Contents

Parallel Session A, Wednesday 26th August 14.45-16.15 ................................................................. 4
  Session A1 – Data Quality. .................................................................................................................. 4
  Session A2 – Data Science. ................................................................................................................. 9
  Session A3 – Joined-Up Information Governance ......................................................................... 15
  Session A4 – Naturalistic Cohort Studies ....................................................................................... 19
  Session A5 – Naturalistic Cohort Studies ....................................................................................... 24
  Session A6 – Linked Observation For Health Policy ...................................................................... 29

Parallel Session B, Wednesday 26th August 17.00-18.30 ............................................................... 34
  Session B1 – Statistics & Analytics. ................................................................................................. 34
  Session B2 – Data Quality. ............................................................................................................... 38
  Session B3 – E-Infrastructure. ......................................................................................................... 42
  Session B4 – Geographical Studies & Methods .......................................................................... 46
  Session B5 – Naturalistic Cohort Studies. ...................................................................................... 50
  Session B6 – Public Engagement in Research. ............................................................................... 55

Parallel Session C, Thursday 27th August 09.00-11.05 ................................................................. 61
  Session C1 – Data Linkage Methods. ............................................................................................... 61
  Session C2 – Data Quality: Primary Care. ...................................................................................... 67
  Session C3 – Joined-up Information Governance ........................................................................ 73
  Session C4 – Linked Observation for Health Policy ..................................................................... 81
  Session C5 – Pharmacoepidemiology. ............................................................................................ 88
  Session C6 – Research Cohort Extension. ..................................................................................... 95

Parallel Session E, Thursday 27th August 14.15-16.15 ................................................................. 102
  Session E2 – Kidney Disease. ........................................................................................................ 102
  Session E3 – Trial’s Methodology ................................................................................................. 108
  Session E4 – Research Cohorts. .................................................................................................... 112
  Session E5 – Statistics & Analytics. ............................................................................................... 118
  Session E6 – Natural Experiments. .............................................................................................. 123

Parallel Session F, Friday 28th August 10.15-12.15 ................................................................. 128
  Session F1 – Apps and Interventional Informatics .................................................................... 128
  Session F2 – E-Infrastructure ....................................................................................................... 135
  Session F3 – Geographical Studies & Methods. .......................................................................... 141
  Session F4 – Linked Observation for Health Policy ................................................................. 145
  Session F5 – Natural Experiments. ............................................................................................. 151
  Session F6 – Joined-Up Information Governance ..................................................................... 158

Parallel Session G, Friday 28th August 13.55-15.15 ................................................................. 161
  Session G1 – Data Science. ............................................................................................................ 161
  Session G2 – Joined-Up Information Governance ................................................................. 165
  Session G3 – Linked Observation for Health Policy ............................................................... 168
  Session G4 – Natural Experiments. ............................................................................................ 172

The Farr Institute International Conference 2015: Data Intensive Health Research and Care
26-28 August 2015, St Andrews
Session G5 – Naturalistic Cohort Studies.................................................................176

Poster Abstracts ........................................................................................................180
Apps and Interventional Informatics........................................................................180
Statistics and Analytics .............................................................................................182
Data linkage methods.................................................................................................189
Data quality................................................................................................................190
Data Science ...............................................................................................................191
E-Infrastructure ............................................................................................................196
Geographical Studies and Methods ..........................................................................200
Joined-Up Information Governance ..........................................................................203
Patient and Public Involvement ...............................................................................209
Linked Observation for Health Policy ......................................................................211
Natural Experiments .................................................................................................218
Naturalistic Cohort Studies .........................................................................................224
Research Cohort Extension.........................................................................................232
Trial’s Methodology and Innovation ...........................................................................238
First Nations organizations and communities in Ontario, Canada have had limited access to First Nations-specific health services utilization information for policy development and analysis. In large part, this is because the data have lacked inclusive and reliable identifiers of First Nations identity.

One initial method used to highlight First Nations populations in administrative data has been to identify geographic areas that represent the 133 First Nation communities in Ontario. However, these “on-reserve” populations represent fewer than half of the over 175,000 registered First Nations people in Ontario and may include non-First Nations individuals. In order to extend beyond this cursory approach to identification, the Chiefs of Ontario (COO) entered into a Data Governance Agreement with the Institute for Clinical Evaluative Sciences (ICES), which enabled the encoded identification of registered First Nations people at an individual level. This central linkage of the federal Indian Registry System (IRS) with the Ontario Registered Persons Database (RPDB) allows for the potential unlocking of the many population-based provincial health and social databases held by ICES for use by First Nations.

This presentation will highlight details of the probabilistic linkage of the IRS to the RPDB and the ensuing validation work to establish a foundation for the technical use of the data. To date, 176,266 First Nation individuals (93%) from the IRS have been successfully linked to the RPDB. Those who we could not successfully link were more likely to be female, living off-reserve, and married at the time of registration in the IRS. Overall, approximately half of those linked were living off-reserve at registration. The linkage with the RPDB allows for the cross-validation of several demographic data elements found in the IRS, including date of birth and confirmed deaths. We will compare the profiles of the First Nations populations identified via the geography-based method and the registry-based method. The resulting recommendations inform the broader international discourse around methods of Indigenous identification in administrative data.

These are the first steps to unlocking First Nations-specific administrative health data as a tool in policy development and in telling the stories of their people and communities. The benefits to First Nations ownership, control, access and possession of data include the ability to highlight disparities in the use of a wide range of health services, to highlight communities with particularly high and low rates of chronic conditions, and evaluate the effects of policy changes at community and regional population levels.
DATA LINKAGE TO IMPROVE IDENTIFICATION OF MATERNAL AND NEONATAL OUTCOMES IN ELECTRONIC HEALTH RECORDS

Authors: Katie Harron¹, Ania Zylbersztejn², Hannah Knight³, Ruth Gilbert², David Cromwell¹, Jan van der Meulen¹
¹London School of Hygiene and Tropical Medicine, ²UCL, ³RCOG

Abstract 1643:

Background: Conditions and events can be coded in numerous ways within electronic health records relating to the same individual. Where there is inconsistency within records, there can be uncertainty about the most appropriate criteria for defining a cohort or coding a risk-factor. One example is stillbirth, which can be coded in multiple ways in baby records. We evaluated whether linkage to information on maternal records improved ascertainment of stillbirths in Hospital Episode Statistics (HES).

Methods: All records relating to birth episodes for babies and delivery episodes for mothers admitted to NHS hospitals in England between April 2010 and March 2012 were extracted from HES (n=1,361,039). Baby and maternal records were linked together using clinical information recorded for each birth/delivery episode. We used three categories of codes for identifying stillbirths in baby records only: discharge method, birth status and ICD-10 diagnosis. We then determined whether additional stillbirths could be identified in the linked baby-maternal records using birth status and ICD-10 diagnosis recorded in the maternal record. Linkage to subsequent infant readmissions was used to identify live births wrongly classified as stillbirths. Consistency of codes within and between records was evaluated. Rates of stillbirths as identified in baby records only or in the linked baby-maternal records were compared with national statistics.

Results: Agreement between codes identifying stillbirths was high within baby records (kappa=78.84%) and maternal records (kappa=80%). Infant readmissions were observed in 3/4033=0.1% of baby records coded as stillbirths using discharge method, 25/3769=0.7% (baby) and 27/4374=0.6% (maternal) using birth status, and 52/5119=1.0% of maternal records using ICD-10 codes, suggesting errors in coding.

The national rate of stillbirths occurring in hospitals in England between 2010-2012 was 0.53% (Office for National Statistics). Based on the presence of any relevant code, stillbirths were under-reported in baby records (0.43%). When information on the maternal record was linked to the baby record, the stillbirth rate based on any code rose to 0.49%. There was close agreement between stillbirth codes in baby and maternal records (kappa=92%).

Discussion: The number of stillbirths identified in the linked baby-maternal hospital records for England was approximately 8% lower than in national statistics. On inspection, it appeared that a contributing factor to this under-ascertainment may have been multiple births sharing records within HES. Ultimately, combining information from more than one source using data linkage can be used to help validate data quality and increase ascertainment of important conditions in electronic healthcare data.
VALIDITY OF SELF-REPORTED PARENTAL HIP FRACTURE: A POPULATION-BASED PARENT-OFFSPRING RECORD LINKAGE STUDY

Authors: Lisa M. Lix¹, William D. Leslie², Shuman Yang³, Suzanne N. Morin³, Sumit R. Majumdar⁴, Leslie L. Roos²
¹University of Manitoba, ²University of Manitoba, Winnipeg, Manitoba, Canada, ³McGill University, Montreal, Quebec, Canada, ⁴University of Alberta, Edmonton, Alberta, Canada

Presenting Author: Lisa M. Lix

Corresponding author email: Lisa.Lix@umanitoba.ca

Introduction: Multiple risk factors have been tested for their association with osteoporosis-related fracture, including parental hip fracture. This information is usually collected via self-report from offspring, a method potentially prone to recall bias. Our purpose was to test the validity of self-reported parental hip fracture using linked administrative health databases.

Method: The study was conducted with population-based databases from Manitoba, Canada (population 1.2 million; public healthcare system), including hospital discharge abstracts, population registry files, and a unique province-wide bone mineral density (BMD) registry containing test results and clinical risk factors for fracture. The study cohort included individuals 40+ years with self-reported parental hip fracture information in the BMD registry between January 2006 and March 2013; BMD testing selection criteria emphasize screening for women ages 65+ and targeted testing in men and high-risk younger women. The population registry (1970 – 2013) was used to identify parent-offspring clusters and dates of health insurance coverage; offspring whose parents did not have continuous health insurance coverage in the observation period were excluded. Diagnoses recorded with the International Classification of Diseases (ICD) in hospital discharge abstracts (1970 – 2013) were used to ascertain parental hip fracture (ICD-8 N820; ICD-9-CM code 820; ICD-10-CA S72.0–S72.2; all persons with hip fracture are hospitalized). Sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and kappa (κ; > 0.60 indicates substantial agreement) were estimated.

Results: We identified 7,359 offspring in the BMD registry with parental information; 12.2% reported a parental hip fracture. Mean offspring age was 52.5 years (SD 5.7) and 85.7% were female. Self-report sensitivity was 0.69 (95% confidence interval [CI] 0.67, 0.72), specificity was 0.96 (95% CI 0.96, 0.97), PPV was 0.76 (95% CI 0.73, 0.79), NPV was 0.95 (95% CI 0.95, 0.96), and κ was 0.68 (95% CI 0.66, 0.71). There was 92.9% overall agreement between self-report and administrative data. Sex-stratified analyses produced sensitivity for males of 0.57 (95% CI 0.49, 0.64) and for females of 0.72 (95% CI 0.69, 0.75); for males κ was 0.62 (95% CI 0.54, 0.69) and for females it was 0.70 (0.67, 0.72). Sensitivity was lowest for 40-44 years (0.65; κ = 0.56) and highest for 55-59 years (0.72; κ = 0.69).

Conclusions: Offspring self-reported parental hip fracture exhibited very good to substantial agreement with diagnoses in administrative health data. The unique record linkage capabilities in Manitoba presents a number of opportunities for including parental health history in offspring disease risk prediction.
CREATING DISEASE E-REGISTRIES: DATA COMPARISON BETWEEN THE ACUTE MYOCARDIAL INFARCTION REGISTRY AND THE NATIONAL ELECTRONIC HEALTH RECORD IN ESTONIA.

Authors:
Taavi Tillmann¹, Marje Johanson², Tanel Pärnamaa³, Sven Laur³, Timo Petmanson², Jaak Vilo³
¹UCL Epidemiology and Public Health; STACC Estonia, ²Software Technology and Applications Competence Center (STACC), Estonia, ³Institute of Computer Science, University of Tartu, Estonia; STACC, ⁴Software Technology and Applications Competence Center, Estonia

Presenting Author: Taavi Tillmann
Corresponding author email: t.tillmann@ucl.ac.uk

Background: Disease-specific registries are important, yet their cost is often prohibitive. We assessed the feasibility of reproducing a registry-like dataset using Electronic Health Records (EHR) in Estonia, a small country at the forefront of EHR innovation and adaptation. We compared our results to the official Acute Myocardial Infarction Registry, initially in unlinked format.

Method: The complete national EHR dataset (1 million patients) was used to analyse data from 3.5 million discharge summaries and 4.3 million GP consultation notes in a linked and pseudonymized format spanning January 2012 – January 2014. From this we sampled 2012 data for case identification. Linkage was facilitated by a national ID-number system. We adapted available text-mining algorithms into the Estonian language to predict and validate STEMI/NTEMI status, smoking status and drugs prescribed at discharge (other risk factors will be analysed by August 2015).

Results: During 2012, we identified 2965 AMI events in 2804 patients using inpatient discharge summaries. This is 3.0% more events and 1.6% more persons than that reported by the official registry. Mean sample age (70.9 and 70.4 years, respectively) and gender (58% and 60% male) were comparable, and the age profiles of both cohorts matched closely. In addition, we also identified 5% additional events and 6% additional persons from GP records. The distribution of smoking (20%, 24%) and NSTEMI (53%, 51%) was comparable between EHR and registry data. The proportion of patients whose discharge medications did not include at least 2 recommended drugs was lower in registry data (18%) than in EHR data (27%, especially with antplatelets), suggesting that registry data may include an optimistic bias. The unadjusted 30-day mortality rate is currently known to be at least 10%. Subsequent linkage to the mortality database will clarify this. Mortality appears higher in NTEMI (11.3%) than STEMI (5.3%) cases.

Conclusion: EHR data in Estonia appears to match surprisingly closely to official AMI registry data in terms of absolute numbers and the demographic profile. Extracting data on risk factors and treatment can be challenging, but innovative text-mining tools can facilitate this sufficiently. Future work in the development of e-Registries could lead to: a) cost savings in the processes of existing registries b) cost-effective development of novel “virtual” registries and c) easier international comparisons of health system outcomes. Overall, e-Registries may identify areas of clinical care that would benefit from improvement, as well as providing baseline epidemiologic data for disease surveillance, and optimum healthcare resource management.
Abstract 1655

KNOWN UNKNOWNS AND UNKNOWN UNKNOWNS: CAN DATA LINKAGE BE USED TO EVALUATE UNCERTAINTY IN ROUTINE HEALTH DATA SOURCES: RESULTS FROM THE BORN IN BRADFORD COHORT

Authors: Emily Petherick¹, S.J. Prady², D Mason³, B Kelly³, K.e. Pickett², J Wright³

¹Bradford Institute for Health Research, ²Department of Health Sciences, University of York, ³Born in Bradford, Bradford Institute for Health Research

Presenting Author: Emily Petherick

Corresponding author email: emily.petherick@bthft.nhs.uk

Background: Linkage of routine health data to data collected for research purposes has the capacity to enhance expand epidemiological analyses across a wide range of clinical areas. For most clinical data collections validation of available information is usually only possible in small subsets due to the costs and data protection concerns. Linkage of research and clinical data sets provides a useful framework to evaluate completeness and accuracy of routine clinical data. This will be explored using two exemplar data items; ethnicity and pregnancy.

Methods: 12450 pregnant women were recruited to the Born in Bradford study, typically at 26-28 weeks gestation. At recruitment women completed a detailed questionnaire which captured self-reported ethnicity in addition to a wide range of socio-demographic information. This research data was deterministically linked via their NHS number to primary care records provided by SystmOne, along with the boundary dates of the known pregnancy from maternity records. The primary care records were searched for the presence of Read codes indicating pregnancy within the known pregnancy period, and ethnicity. The completeness of recording of pregnancy and ethnicity, and sensitivity and specificity of the previously published Read code lists to detect them were calculated.

Results: Complete data were available to permit data linkage between primary care records, maternity records and recruitment questionnaire data for 11980 pregnancies in 10978 women. 98.2% had a Read code in their records indicative of pregnancy, with sensitivity and specificity in excess of 97%. Ethnicity was only recorded for 58.3% of women were not missing at random, with rates higher for women who self-reported as Other, White British and lowest in mothers of Black, Mixed Ethnicity, Pakistani and Bangladeshi ethnicity. Where Read codes were available agreement with self-reported ethnicity was good, with sensitivity and specificity in excess of 95%.

Conclusions: Routine primary care data may be used to accurately identify index, and therefore future, pregnancies. Ethnicity data may not always be recorded in the records of the maternal samples but where it is available is likely to be accurate. Data linkage between routine health and research data is a novel method by which bias in missing data of routine data sources can be identified. Data are likelier to be more complete for known clinical events than for unknown socio-demographic characteristics.
Abstract 1732
DATASHIELD: TAKING THE ANALYSIS TO THE DATA (NOT THE DATA TO THE ANALYSIS) IN THE FARR INSTITUTE

Authors:
Paul Burton¹, A W Boyd², J T Minion², R C Wilson², A J Turner², M J Murtagh², R P Cornish², A J V Davies²
¹University of Bristol, ²School of Social and Community Medicine, University of Bristol

Presenting Author: Paul Burton
Corresponding author email: Paul.burton@bristol.ac.uk

Research in modern biomedicine and social science commonly requires large sample sizes and a capacity to integrate high quality data from different scientific domains. Some of the greatest challenges derive from data governance. Physical pooling or sharing of information from individuals in a central database to be queried by researchers raises important ethical-legal questions regarding data security and disclosure of identity and issues relating to intellectual property and the rights and responsibilities of varied stakeholders. With development funding contributed by both the Welsh and Scottish Farr Institutes, DataSHIELD provides a novel technological solution that circumvents some of the most basic challenges in facilitating the access of researchers and other healthcare professionals to individual-level data. DataSHIELD comes in two fundamental flavours: horizontal (HDS) and vertical (VDS).

Development of the HDS - for secured meta-analysis of ‘horizontally partitioned’ data where different sources hold the same data but on different individuals – is further advanced than VDS. Data sets are analysed simultaneously but in parallel; the analyses being linked by non-disclosive summary statistics and commands transmitted between a central analysis computer (AC), and each data computer (DC). Technical implementation employs a specially modified R statistical environment linked to an Opal database deployed behind the computer firewall at each data source. VDS is primarily being developed under the Welsh and Scottish Farr Institutes working with ALSPAC data. VDS addresses the secure analysis of linked data: different data-sources hold different data on the same individuals. In contrast to HDS, VDS requires sequential rounds of encryption and decryption to maintain data security. It is therefore more demanding technically and presents a range of interesting methodological challenges. However, in relation to the management of sensitive data under Farr, it is now recognised that VDS may be complemented by HDS to provide a flexible, cost-effective, open-source solution to secure data access to sensitive data-bases that have already been created via record linkage. This is useful if a linked data set – e.g. created using the SAIL infrastructure – needs to be accessible for third-party analysis but the individual-level data must remain strictly confidential and some level of disclosure control is desired over the analyses that may be undertaken. This presentation will describe the evolution to-date of the Vertical DataSHIELD and the potential application of Horizontal DataSHIELD to Farr Institute data. We hope to promote discussion across the Farr Institute of potential uses for DataSHIELD and priorities for its ongoing development.
Introduction: In many scientific domains, there is increasing emphasis, and in some cases, an insistence, on ensuring that research can be replicated, and that research results are reproducible when the original research is replicated. These concerns have been motivated by the realisation that a significant proportion of published research results cannot be reproduced, either because the methods used were inadequately documented, or have been kept secret, or because an error was made in the formulation of the methods or in their execution. There is no reason to believe that health research using linked data is immune to such problems - indeed, the size and complexity of linked data set analyses make errors far more likely than in many other types of research. Fortunately, there exists a range of techniques, tools and technologies which can be employed during the data management and analysis phase of linked data health research projects and studies in order to minimize the probability of researchers making mistakes, and to maximize the probability of other researchers finding mistakes, should they still occur.

Methods: Review of methods used in cognate disciplines to ensure or promote reproducibility of research findings, and description of some secure data sharing platforms now available to population health and health services researchers which facilitate replication of analyses while still maintaining privacy and confidentiality. The tools and techniques used to promote reproducible research include: the use of publicly accessible programming code repositories (such as GitHub); literate programming approaches in which analysis computer code is interleaved with explanatory text and results, for publication in conjunction with the formal scientific paper; and software testing methods to check that computer code written by health researchers is doing what the researchers hope it is doing.

Results: A motivating example will be presented: the application of reproducible research methods enabled detection and elucidation of serious calculation errors in a published, widely-cited and policy-relevant research into bicycle helmet efficacy, and they ensured reproducibility when the erroneous research was replicated, without errors, to correct the scientific record. The way in which these methods can be used in a typical linked data analysis will be briefly demonstrated.

Conclusions/implications: Health researchers using linked data should routinely be using readily available reproducible research techniques and technologies in order to ensure their studies are replicable and their results are reproducible and thus likely to be free of errors in their execution.
Clinical coding systems enable clinicians and healthcare professionals to record clinical concepts in Electronic Health Records (EHRs) for direct care purposes using standardized numerical or textual codes. Use of clinical data within these EHRs for research purposes often focuses on analytical concepts (e.g. “cancer diagnosis”) that are more abstract than can be referenced with a single clinical code. To represent these abstract analytical concepts, clinical codes can be grouped into sets (often referred to as “clusters” or “code lists”), such that any reference to a code in the set within an EHR can be considered to be a reference to the abstract analytical concept.

Abstract analytical concepts are fundamental to the lifecycle of research using EHRs. They determine: (1) which data items are required for each patient from a custodian of EHRs to undertake a particular analysis, (2) how EHRs are processed once received from custodians to answer a specific research question, and (3) how results are interpreted and compared with other analyses. Indeed, the importance of these concepts to the lifecycle of research has been recognised, with many academic journals enabling the concepts defined (or reused) for a particular analysis using EHRs to be published as supplementary material for an article, and online repositories emerging that enable concepts to be published for reuse by the research community.

To ensure the provenance, repeatability and validity of the research, such concepts need to be systematically integrated into the lifecycle of research. However, these concepts are often published using a variety of ad-hoc and informal representations. Such representations present challenges for systematic integration into the research lifecycle, with inconsistent syntax and ambiguous semantics. Additionally, there is an absence of software tools to support the various stages of the research lifecycle that can consume and act upon such representations.

Accordingly, we have developed an XML-encoded language with which sets of clinical codes can be represented consistently and unambiguously. This language has been used to represent abstract analytical concepts defined as part of the Quality Outcomes Framework (QOF) in primary care, and to represent a set of clinician-defined concepts relating to specific musculoskeletal conditions. Additionally, we have developed associated software tools that enable these concept representations to: (1) drive requests for data items from custodians of EPRs, (2) be seamlessly integrated into scripts used for processing EHR data, and (3) be published, compared and reused within research communities.
Abstract 1743

COLLABORATIVE RESEARCH ACROSS CENTRES: WHY EXTRACTION TRANSFORM & LOAD (ETL) SHOULD NOT BE A FORGOTTEN DARK ART?

Authors: Kevin Ross, Ian Ford, Sharon Kean, Colin McCowan
University of Glasgow

Presenting Author: Kevin Ross

Corresponding author email: colin.mccowan@glasgow.ac.uk

Background: Increasingly clinical research centres are looking to do collaborative research work using their routinely collected clinical data sources. One key mechanism for this work is the ability for centres to prepare their data to a standardised format which then allows centralised tools to work with and analyse this data. This process generally known as Extraction, Transform & Load (ETL) requires an in depth knowledge of the original data source, technical programming skills to produce customised coding to move it to a shared platform and documentation of the processes to allow for data quality and provenance checks.

Methods: Two recent projects requiring collaboration across different centres, Electronic Health Records for Clinical Research (EHR4CR) and PROMISE-RA have required the extraction of clinical datasets from the same source files within the Glasgow Safe Haven with subsequent transformation and load to a project specific platform and to an i2b2[1] clinical data warehouse. We undertook ETL processes for hospital episode data, prescribing information, specified laboratory results, specialist clinic data and demographic information.

Results: Routinely collected data on over 548,000 patients was used within the EHR4CR project and more detailed granular information on 450 patients with rheumatoid arthritis was used for PROMISE-RA. As EHR4CR is pan-European, a project requirement was to present clinical data in a number of pre-agreed international coding formats eg. LOINC for laboratory data, ICD10 for diagnoses, ATC for drugs, Pathlex for pathology and SNOMED for patient demographics. Mapping between coding formats has been a problematic and time consuming task as there are rarely one-to-one matches or obvious links between coding systems. Laboratory test and drug data were particularly challenging as Scottish clinical data is coded using UK specific coding systems (Read, BNF) which have no direct international equivalents.

Discussion: Collaborative work across different data centres is an expanding area of work which is likely to become more important in the future. The ability of individual centres to participate in such research is predicated in good understanding and metadata of their original clinical data but also having technical skills and knowledge in how to perform ETL processes maintaining data quality and provenance. This is a key step for clinical collaborations but the effort and resource required is often underestimated and the process requires greater transparency.

References

[1] i2b2 (Informatics for Integrating Biology and the Bedside) is an NIH-funded National Center for Biomedical Computing based at Partners HealthCare.
Abstract 1585

**ApiNATOMY: A PHYSIOLOGY KNOWLEDGE MANAGEMENT TOOLKIT TO ORGANIZE MOLECULAR AND CLINICAL DATA**

**Authors:** Bernard de Bono, S Islam, S Alexander, P Grenon
*Farr Institute, London*

**Presenting Author:** Bernard de Bono

**Corresponding author email:** bernard.de.bono@gmail.com

We report on the development and application of an open-source graphical 'GoogleMaps'-style software toolkit, ApiNATOMY. In particular, this report provides an account of the application of ApiNATOMY in support of physiology-related data analytics, modelling and simulation within pharmaceutical research cycle.

ApiNATOMY manages physiology knowledge to semantically bridge data from multiple structural scales. The core of the ApiNATOMY toolkit is a knowledgebase (KB) that takes into account anatomical locations and their physiological connections, as represented through the application of standard reference ontologies. A graphical user interface (GUI) module is linked to this KB to display, navigate, query and annotate a map of this physiology knowledge in the form of a circuit board.

This toolkit is being applied, in the context of three projects funded by the Innovative Medicines Initiative (IMI). The combined breadth of these three project applications cover a substantial proportion of knowledge management in the pharmaceutical research cycle, namely (i) Drug Discovery [OpenPHACTS project], (ii) Drug Development [DDMoRe project], and (iii) Clinical Trials [AETIONOMY project]. In this report, we outline the role ApiNATOMY plays in these three key areas and indicate the benefits of the future application of ApiNATOMY to semantically bridge large datasets across these three disciplines.

In effect, ApiNATOMY provides the means to consistently and coherently represent body compartments, and applies this representation to depict routes of communication between compartments. In a pharmaceutical context, this approach allows for:

- the mapping of the results of different types of assays, diagnostic tests and clinical examinations onto the same circuit board, and
- the inferring of the physiological relationship, over the circuit board, between the above resources as a means to propose mechanistic interpretation to statistical correlation between results, as well as to simplify the generation of mathematical models that study this relationship between results.

In an IMI project context, ApiNATOMY integrates knowledge about the anatomical locations:

1. of expression and other assay data about enzymes, transporters and receptors relevant to drug discovery [OpenPHACTS];
2. of assays and variables from mathematical models relevant to the simulation of drug absorption, distribution, metabolism and excretion processes, as well as downstream drug effects [DDMoRe];
3. that, by consensus, clinical experts impute as relevant to symptoms and signs elicited from patient history and examination [AETIONOMY].

We demonstrate the application of ApiNATOMY to the above three areas, and discuss the implications of this method to cross-disciplinary data analysis over the entire span of the pharmaceutical research cycle.
Abstract 1763

TEAM SCIENCE – THE ASTHMA-E-LAB

Authors: Philip Couch, Christopher Munro, Andrew Broadbent, Stephen Lloyd, Paul Stephenson, Angela Simpson, John Ainsworth, Adnan Custovic, Ian Buchan
University of Manchester

Presenting Author: Philip Couch

Corresponding author email: philip.couch@manchester.ac.uk

Here we present the informatics approach that underpins the Asthma-e-Lab, an application of the HeRC e-Lab platform, bringing together Clinicians, Health Informaticians, Statisticians and Data Managers to establish an improved understanding of early life asthma. There is growing evidence to suggest that asthma is an umbrella diagnosis that includes multiple diseases with different underlying mechanisms. It is thought to be unlikely that these different diseases will respond in the same way to therapeutic treatments. The aim of the STELAR project is to perform a statistical analysis across data gathered as part of multiple UK based birth cohort studies to identify clusters of characteristics (or phenotypes) of asthma that may in turn relate to specific endotypes. Identification of these endotypes may then support an improved stratified approach to the treatment of individuals with the disease.

The e-Lab platform can be used to create collaboration spaces that can be accessed in a straightforward manner using a standard web browser. These spaces can be tailored for specific teams, making available tools that include those to manage documents, wikis, blogs and provide social networking capability. E-Labs offer a standard set of tools that can be incorporated into these spaces, but also implement an architecture that allows new software components to be easily incorporated.

In support of the STELAR project, e-Lab modules have been developed to allow the integration of data sets gathered across the five birth cohort studies. These data sets can be imported into the e-Lab and associated with a data description. The descriptions are developed by the data managers of the cohorts and provide detailed information about the imported data. This includes information about how data are represented, the semantics and additional context required for correct interpretation. These data sets can then be integrated in a manner that allows research questions to be asked across multiple data sets, created by different communities. The e-Lab is unlike many other data management systems in that it is designed to operate on any data. Data are stored in a domain independent way such that the software can be easily extended and applied in different contexts. The e-Lab builds on other standards and we present the application of Minimum Information Models to support the capture of data context and Research Objects to aggregate and share the contents of a collaboration space.
SESSION A3 – JOINED-UP INFORMATION GOVERNANCE.
CHAIR: PROF GRAEME LAURIE, UNIVERSITY OF EDINBURGH

Abstract 1529 HOW CAN FARR LEAD THE WAY WITH RESPONSIBLE DATA LINKAGE FOR HEALTH RESEARCH?

Authors: Graeme Laurie1, Nayha Sethi1, Christine Dobbs2, James Cunningham3, Nathan Lea4, Kerina Jones2
1University of Edinburgh, 2University of Swansea, 3Manchester University, 4UCL

Presenting Author: Graeme Laurie

Corresponding author email: graeme.laurie@ed.ac.uk

This paper provides an account from the IG Working Group within the Farr Institute of essential elements of good governance of data linkage for health-related research. Drawing upon lessons learned from previous projects, and reflecting on regulatory failures such as care.data in NHS England, the discussion will centre around those governance mechanisms which have been deployed in practice and will offer an assessment of these as elements essential to achieving good governance. The argument will be made that a crucial feature in any step-increase of research capability must be the adoption of adaptive governance models. These must recognise a range of approaches to delivering safe and effective data linkage while remaining responsive to public and research user needs and expectations as these shift and change with time and experience. The targets are multiple and constantly moving. There is not, nor should we seek, a single magic bullet in delivering good governance in health research. Proportionality and mutually of recognition of "what works" are crucial elements of any multi-site initiative to promote effective sharing of data for health research.

Abstract 1580 HOW DO ENGLISH GPS EXPLAIN THEIR INFORMATION SHARING BEHAVIOUR? A QUALITATIVE STUDY FRAMED BY PRINCIPLES FROM THE LAW, ETHICS AND PROFESSIONAL GUIDELINES.

Authors: Jon Fistein1, Justin Keen2, Jeremy Wyatt3
1Leeds University/Medical Research Council, 2Leeds University, 3Leeds

Presenting Author: Jon Fistein

Corresponding author email: j.l.fistein@leeds.ac.uk

Patients expect their General Medical Practitioners (GPs) to use information about them to provide safe, efficient care. There are several circumstances when GPs must share patient information beyond their Practice with colleagues in other healthcare organisations to support care, and other circumstances where they may be asked to share patient information for other purposes for example, with governmental agencies, for health protection, for public health and for medical research. This current research aims to investigate the rationale that GPs use for either sharing or not sharing information across a range of scenarios.

GPs have a duty of (common law) confidentiality not to disclose patient information without lawful justification. The data are usually stored on the GPs’ practice systems and are regarded as ‘personal’ data under the Data Protection Act 1998, meaning that the GP is generally considered to be the ‘Data Controller’ with obligations only to process the data ‘fairly and lawfully’. The GP must also handle the data ethically and professionally, for example as described in the General Medical Council’s
guidance.

The research aims to evaluate GP knowledge, understanding and beliefs about the existing legal, ethical and professional frameworks and to assess whether these frameworks are effective in achieving particular policy outcomes.

GPs are being asked to describe their current information sharing actions, providing real examples from their experience or that of colleagues, and then to provide their explanation for these behaviours. A two-stage interview technique is being used. In the first stage, the GP is invited to share narratives about their professional experiences of information sharing. In the second stage, a topic guide is used to structure directed questions about particular concepts in more depth and to cover other areas that have not emerged spontaneously from the interviewee’s narrative. The interview material is then transcribed, coded and thematically analysed. This is then compared with the framework concepts identified in sources of English law, bioethics literature and professional guidelines.

This preliminary report presents the early findings from the study, outlining the underlying legal, ethical and professional frameworks, and presenting the analysis of the first round of GP interviews.

Abstract 1676

ARE REPRODUCIBILITY, EFFICIENCY AND ACCURACY OF SCIENTIFIC METHODS AT ODDS WITH PROTECTING PERSONAL DATA WHEN LINKING DATA FOR RESEARCH

Authors: Emily Jefferson
University of Dundee

Presenting Author: Emily Jefferson

Corresponding author email: e.r.jefferson@dundee.ac.uk

Linking routine and research datasets offers unique opportunities for cost-effective research at large-scale. However, data sharing and linkage raises significant privacy and confidentiality concerns, especially when considering unconsented data. There are many different safeguards which have been adopted when linking data to minimise privacy concerns such as:

- Anonymisation of data and provision to researchers using project specific pseudo-identifiers.
- Provision of only the minimum amount of data to answer a specific scientific question.
- Separation of the roles of indexer and linker so that no one organisation can observe both linked descriptive data and identifiable data together.
- Analysis of data by researchers only within safe haven environments, where only aggregate level data can be exported.
- Linking of data from different organisations afresh for each research project rather than the creation of permanently linked large data warehouses.

These safeguards satisfy many confidentiality concerns, however, the question is if we adopt these safeguards how do they impact the requirements of research?

- If analysis in safe haven environments is enforced, what happens if safe haven environments are not fit for purpose (i.e. lacking either the hardware or software required for the research project)? This is especially a problem when Big Data is considered.
- If a Data Controller organisation has to produce bespoke extracts each time for each research project, how will they ensure that resource is available in a timely fashion to produce the extract? How will they ensure that the extract will be reproducible in 5 years’ time, ensure consistency of extract independent of the individual producing the extract and an audit trial? Who will ensure that there are no bugs in the extract creation?
- Data obtained directly from Data Controllers can be of very poor quality; and thus, metadata and data standards are crucial to data analysis. Who is best placed to clean, transform and add metadata to research datasets: the Data Controller organisation, a safe haven/linked data provider or each separate research group? If the cleaning/transformation steps cannot
be reproduced/shared, then how can the scientific methods be followed? How is credit apportioned?

- How can non-hypothesis driven research be carried out if only the minimum amount of data is provided to answer a specific research question?

This talk looks into solutions for addressing the requirements of research whilst adhering to the safeguards which minimise privacy concerns and the possibility of proportionate use of safeguards dependant on the sensitivity of the linked data.

Abstract 1633

BARRIERS AND ENABLERS TO CROSS-ORGANIZATIONAL AND CROSS-SECTORAL DATA SHARING AND LINKAGE ACROSS THE UK: A CONSULTATION WORKSHOP.

Authors:
Christine Dobbs1, J.A. Cunningham2, G Laurie3, N Lea4, N Sethi3
1Farr Institute CIPHER, Swansea University, 2Farr Institute HeRC, University of Manchester, 3Farr Institute Scotland, University of Edinburgh, 4Farr Institute London, University College of London

Presenting Author: Christine Dobbs

Corresponding author email: c.dobbs@swansea.ac.uk

Problem: It is widely recognised that there is a complex array of legislation and regulations around the use of data. Furthermore, guidelines are not always clear, and interpretations vary. Models of data management and access also differ across organisations.

Objectives: In November 2014 members of the Farr Institute Innovative Governance Workstream ran a full-day consultation workshop to bring together key stakeholders from the research community. The four main objectives were to (1) note past experiences, (2) understand future expectations for cross-organizational and cross-sectoral linkage, (3) identify lessons to be learned around good IG delivery and (4) prioritise the emergent enablers and barriers, in order to highlight which issues are most pressing.

Method: Nominal Group Technique (NGT: e.g. Centers for Disease Control and Prevention, 2006; Jones & Hunter, 1995) was the chosen methodological approach, because it is particularly inclusive in terms of encouraging and facilitating participation from all stakeholders and in facilitating group consensus. By consensus attendees identified ten barriers and ten enablers to data-sharing, and in conclusion ranked these lists individually.

Participants: In total 12 participants (‘delegates’, data-linkage researchers) and five Information Governance experts (‘active observers’) attended. These included academic affiliates from all of the Farr nodes, the NHS, the MRC and other data-linkage professionals.

Results: Ten barriers and enablers were identified during the course of the workshop, and in the paper we provide more in-depth findings. Here are the top three listed in order of importance:

Barriers
1. Risk aversion by data custodians in the form of ‘if in doubt, say no’.
2. Lack of understanding of the benefits and value of the data
3. Data sharing hindered because of misunderstandings or knowledge gaps in legislative requirements

Enablers
1. Civic engagement with the wider population must be continued and developed further.
2. Stakeholder map. Possibly a RACI (Responsible, Accountable, Consulted, Informed) matrix.
3. Metadata: To inform researchers’ decision-making processes; data set quality, integrity, provenance and usability.

Discussion: Participation was extremely high and delegates were very supportive of the event (rated as 4.3 out of 5) also sharing with us what they thought the next steps might be. There was a clear set of
messages that emerged from the workshop to maintain a strategy that emphasises civic engagement, and key points to handle the identified barriers. In conclusion we wish to enter into dialogue with the audience to discuss ramifications and to gather their thoughts on potential ways forward.

Abstract 1746

SETTING UP THE INFORMATION GOVERNANCE FOR A UK SECURE eRESEARCH PLATFORM (UKSeRP) – “IT’S A PEOPLE THING, NOT JUST A SET OF CLEVER IT BOXES”

Authors: Martin Heaven¹, Simon Thompson², Ronan Lyons³
¹FARR Institute, Swansea University, ²SAIL at Swansea University, ³FARR Institute Swansea University

Presenting Author: Martin Heaven

Corresponding author email: m.l.heaven@swansea.ac.uk

SAIL Databank/FARR Wales has completely re-engineered and expanded its core infrastructure to create the UK Secure Research Platform. This initiative enables FARR Wales to provide a platform as a service (PaaS) to internal programmes of work as well as externally run national programmes. This work builds on the years of experience gained by successfully running the SAIL Databank and represents the next evolution of our facilities.

However, if it is to set up and run a UKSeRP, any research organisation/collaboration will be required to implement a complete and robust Information Governance model to demonstrate the safety of the data in practice. This needs to satisfy legal requirements and to give confidence to prospective data providers that their data will be handled appropriately whilst in the SeRP environment.

Within the SAIL infrastructure, there are certain formal separations and firewalls in place between the technical infrastructure team and the information analysts who carry out the research on SAIL Databank. We have been able to utilise this internal structure to explore the Information Governance set up from the perspective of a research organisation /collaboration wishing to instigate a UKSeRP: Much of this is about how to inform and manage people.

The presentation will report on work done to develop the essential requirements of a generic UK SeRP Information Governance System specification, by considering both what has worked in the past for SAIL databank and the requirements of ISO 27001.

We will propose a succinct list of essential requirements to help when faced with navigating the considerable rulebooks, guidelines, policies and procedures related to Information Governance.

Abstract 1657

INCREASING ACCESS TO HEALTH ADMINISTRATIVE DATA WITH ICES DATA & ANALYTIC SERVICES

Authors: R Saskin, L Ishiguro, N Gunraj, E Yates, JC Victor
Institute for Clinical Evaluative Sciences

Presenting Author: R Saskin

Corresponding author email: reflik.saskin@ices.on.ca

The Institute for Clinical Evaluative Sciences (ICES) is a not-for-profit organization that conducts research to evaluate healthcare delivery and outcomes in Ontario, Canada. Established in 1992, ICES houses a vast and secure array of linkable, de-identified health-related data on more than 13 million Ontarians, including health services data, healthcare provider data, registries and population-based health surveys.

In 2014 as part of Ontario’s commitment to the Strategy for Patient-Oriented Research (SPOR)—a multi-collaborative initiative between government ministries and funding agencies—ICES launched the Data...
& Analytic Services (DAS) platform. With a primary objective of increasing access to health administrative data for external researchers, ICES DAS provides: access to highly de-identified, risk-reduced datasets created from ICES’ data holdings; analytic support; and complete data analysis and report writing services. ICES DAS is also able to import researcher collected data for linkage to the ICES data holdings. Services are available to publicly-funded researchers, healthcare providers and administrators, policy-makers and students.

All requests submitted for consideration are adjudicated by the ICES DAS team for feasibility and eligibility. Researchers are required to obtain research ethics board approval prior to project initiation.

As part of providing access to data, researchers who choose to perform their own analyses access a secure virtual desktop infrastructure via an encrypted internet connection and use a range of available statistical software. The analytic environment is fully compliant with requirements set out by the Information and Privacy Commissioner of Ontario. Alternatively, projects seeking analytic services are led by ICES scientists with clinical and subject matter expertise and carried out by analysts and epidemiologists who are highly skilled in the use of ICES data and statistical methods. In addition, re-identification risk assessment is conducted on any research results that clients request for the preparation of subsequent reports or manuscripts.

As of abstract submission, ICES DAS has received over 160 requests from across Canada of which 140 have been deemed feasible and eligible. Research topics have been varied and include assessments of healthcare provision by sector, chronic disease prevalence and treatment, and statistical methods. Of the feasible and eligible projects, 20 requested the importation and linkage of external data, 92 involved access to data only and 48 requested analytic services.

ICES DAS has demonstrated preliminary evidence of a promising approach to increase access to one of the most comprehensive linked health administrative data repositories globally and thus increase research capacity for the scientific community in Canada.

**SESSION A4 – NATURALISTIC COHORT STUDIES.**

**CHAIR: DR HESTER WARD, NHS NATIONAL SERVICES SCOTLAND**

**Abstract 1326**  
**RISK FACTORS FOR RECURRENT INJURIOUS FALLS THAT REQUIRE HOSPITALISATION FOR THOSE WITH AND WITHOUT DEMENTIA: A POPULATION BASED COHORT STUDY**

Authors:  
Lynn Meuleners¹, Mk Bulsara², K Chow¹, Jq Ng³  
¹Curtin University, ²Notre Dame University, ³University of Western Australia

Presenting Author: Lynn Meuleners

Corresponding author email: l.meuleners@curtin.edu.au

**Objective** Identify risk factors for recurrent injurious falls that require hospitalisation for people with dementia compared to those without dementia at the population level

**Design** A retrospective whole-population cohort study.

**Setting** De-identified data was obtained from Western Australian Hospital Morbidity Data Collection and the Western Australian Mortality Database using the Western Australian Data Linkage System from 2001 to 2013.

**Participants** There were 32,519 individuals with an index hospital admission for dementia and 33,103 individuals without a diagnosis of dementia who were aged 50+ years.

**Main outcome measures** Recurrent injurious falls that resulted in a hospitalisation.

**Results** There were 23,903 injuries due to a fall that resulted in a hospitalisation for 65,622 participants with (n=32,519) and without dementia (n=33,103) during the study period. Each previous fall increased the risk of a recurrent fall for those with dementia by 2.76 times (95% confidence interval [CI] 2.57–2.79) compared to those without dementia. Men were 8% more likely to have recurrent falls and those with
at least one comorbid health condition were at a significantly increased risk for recurrent injuries due to fall (adjusted hazard ratio 1.08, 95% CI 1.04–1.12). Not being married (single, widowed or divorced) significantly increased the risk of recurrent injuries due to a fall compared to those who were married (adjusted hazard ratio 1.08, 95% CI 1.05–1.12). Increasing age was associated with increased risk of falls. Compared to those aged 50 to 59 years, the risk of falls was 2.48 times higher in those aged 75 to 79 years, 3.62 times higher in those aged 80 to 84 years, and 6 times higher in those aged 85+.

**Conclusion** The risk of recurrent injurious falls that require hospitalisation is a problem for those with dementia. Increasing age, males, not being married, and the presence of at least one comorbid health condition are particularly major risks for recurrent injurious fall requiring hospitalisation.

---

**Abstract 1335**

**MATCHED POPULATION-BASED STUDY EXAMINING THE RISK OF TYPE 2 DIABETES IN PEOPLE WITH AND WITHOUT DIAGNOSED HEPATITIS C INFECTION**

**Authors:** Christian Schnier¹, Sarah Wild², Zain Kurdi², Chris Povey³, David Goldberg⁴, Sharon Hutchinson⁵

¹University of Edinburgh/ UK Biobank, ²Edinburgh University, ³Information Division Scotland, ⁴Health Protection Scotland, ⁵Glasgow Caledonian University

**Presenting Author:** Christian Schnier

**Corresponding author email:** christian.schnier@ed.ac.uk

In Scotland in 2012, an estimated 37,000 people were living with chronic infection with Hepatitis C virus (HCV). People infected with chronic HCV have an increased risk of liver disease, especially when affected by co-morbidities. Meta-analyses have found HCV infection to be associated with an increased risk of type 2 diabetes mellitus (T2DM). Here, we examine this association within a large population-based study, according to chronic and resolved infection.

**Methods:** An HCV-positive cohort was identified using data held at Health Protection Scotland. The HCV-positive cohort consisted of 21,929 people diagnosed with HCV between 1985 and 2011. 15,827 people in the cohort had chronic HCV and 3927 resolved HCV (2175 with unknown status). An HCV-negative cohort was then identified using data from the Community Health Index (CHI). For each person in the HCV-positive cohort, 3 people (in total 65,074 people) were selected from CHI matched on neighbourhood, sex and year of birth. A diagnosis date was attributed to each matched group, equal to the date of HCV diagnosis of the HCV-positive member. Both cohorts were then linked to a database holding information on over 300,000 people with diagnosed diabetes in Scotland. We followed up each cohort member through 3 time periods: up to one year before (pre-HCV) / within one year of (peri-HCV) / more than one year post (post-HCV) the date of HCV-diagnosis. Data were analysed in multivariate regression models including a random group effect.

**Results:** T2DM had been diagnosed in 2.9% of the HCV positive cohort and 2.7% of the HCV negative cohort. A higher proportion of T2DM was diagnosed in the peri-HCV period (i.e. around the time of HCV-diagnosis) for the HCV positive cohort (22%) compared to the HCV negative cohort (10%). In both the pre-HCV and post-HCV periods, only those in the HCV positive cohort living in less deprived areas (13% of the cohort) were found to have a significant excess risk of T2DM compared to those in the HCV negative cohort (adjusted odds ratio (95% CI) in the pre-HCV period: 4.02 (2.32–6.96); for females and 2.33 (1.42–3.83) for males; adjusted hazard ratio in the post-HCV period: 1.53 (1.14–2.04)). These findings were similarly observed for both with chronic HCV and those with resolved HCV.

**Conclusions:** In the largest study of T2DM among chronic HCV-infected individuals to date, there was no evidence to indicate that infection conveyed an appreciable excess risk of T2DM at the population level.
Abstract 1457

CAUSE-SPECIFIC MORTALITY OF MULTIPLE SCLEROSIS PATIENTS FROM BRITISH COLUMBIA, CANADA

Authors: Elaine Kingwell\textsuperscript{1}, Charity Evans\textsuperscript{2}, Feng Zhu\textsuperscript{3}, Tom Duggan\textsuperscript{3}, Joel Oger\textsuperscript{3}, Helen Tremlett\textsuperscript{3}

\textsuperscript{1}University of British Columbia, \textsuperscript{2}College of Pharmacy and Nutrition, University of Saskatchewan, \textsuperscript{3}Faculty of Medicine (Neurology), University of British Columbia

Presenting Author: Elaine Kingwell

Corresponding author email: elainejk@mail.ubc.ca

Background: Lifespan is 6-10 years shorter in people with multiple sclerosis (MS) than in the general population, but the reasons are not well understood. We used linked clinical and population-based administrative health databases to compare mortality risk by specific cause in a Canadian cohort of MS patients with that in the general population.

Methods: MS patients residing in British Columbia (BC), Canada, registered with a BC MS clinic between 01/January/1980 and 31/December/2004 were selected from the BCMS clinical database. Emigration and mortality information were obtained by linkage to BC Health Registry data (from the BC Ministry of Health) and multiple-cause death data (from the BC Vital Statistics Agency). Follow-up was from the most recent of 18th birthday, first BCMS clinic visit or 01/January/1986, to the earlier of death, emigration from BC or 31/December/2009. Cause-specific standardized mortality ratios (SMRs) were calculated by comparison to equivalent cause mortality rates in the BC population (all deaths at ≥18 years of age; 1986-2009) stratified by age group, sex and calendar year. Causes were grouped using International Classification of Disease codes.

Results: Of 6,656 MS patients with 84,433 patient-years of follow-up, 1,091 had died by the study end. MS was the underlying cause in 50.8%, and an underlying or contributing (‘any mention’) cause in 79.5% of deaths. Other major underlying causes included cancer (14.5%); cardiovascular (13.5%); respiratory (including infections) (6.6%); suicides (2.5%) and accidents (2.0%). Relative to the BC population, the risk of death was increased for respiratory illnesses (SMR 2.99; 95% CI 2.36-3.73); suicide (SMR 2.64; 95% CI 1.74-3.84) and cardiovascular disease (SMR 1.44; 95% CI 1.21-1.69), while rates of death due to cancer (SMR 1.00; 95% CI 0.85-1.17) and accidents (SMR 0.92; 95% CI 0.58-1.39) were similar to those in the general population.

Conclusions: The most frequently recorded underlying cause of death in our MS cohort was MS itself, accounting for more than half of all deaths. The risk of death due to respiratory illness or suicide was 2-3 times greater in the MS cohort than in the general population, both of which may be directly attributable to MS. The risk of a cardiovascular-related death was also increased. Mortality risk by ‘any mention’ of causes will be assessed with reference to multiple-cause death rates in the general BC population to gain a better understanding of the causes that contribute to death among people with MS.
Abstract 1531  
Quantification of risk factors for postherpetic neuralgia in a cohort of herpes zoster patients

Authors: Harriet Forbes, K Bhaskaran, S L Thomas, L Smeeth, T Clayton, SM Langan
London School of Hygiene and Tropical Medicine

Presenting Author: Sinead Langan

Corresponding author email: harriet.forbes@lshtm.ac.uk

Objective: Herpes zoster commonly causes disabling neuropathic pain called postherpetic neuralgia. We aimed to investigate risk factors for postherpetic neuralgia.

Methods: Using primary care data from the Clinical Practice Research Datalink, we fitted multivariable logistic regression models to investigate potential risk factors for postherpetic neuralgia (defined as pain ≥90 days after zoster, based on diagnostic and/or prescription codes), including demographic characteristics, co-morbidities, and characteristics of the acute zoster episode. We also assessed whether their effects were modified by antiviral use.

Results: Of 119,413 zoster patients, 6956 (5.8%) developed postherpetic neuralgia. Postherpetic neuralgia risk rose steeply with age, most sharply between 50-79 years (adjusted odds ratio for a 10-year increase, 1.71, 99% confidence interval 1.63-1.78). Postherpetic neuralgia risk was higher in women (6.3% vs 5.1% in men; OR=1.19, 1.10-1.28); and those with severely immunosuppressive conditions, including leukaemia (14.4%; 2.07, 1.08-3.96) and lymphoma (12.1%; 2.45, 1.53-3.93); autoimmune conditions, including rheumatoid arthritis (9.1%; 1.21, 0.99-1.46); and other comorbidities including asthma and diabetes. Ex-smokers also had an increased risk (7.1%: OR vs never-smokers =1.14, 1.05-1.24) as did underweight individuals (8.7%; OR vs healthy weight=1.25, 1.01-1.54). Antiviral use was not associated with postherpetic neuralgia (OR=1.04, 0.97-1.11). However the increased risk associated with severe immunosuppression appeared less pronounced in patients given antivirals.

Interpretation: Postherpetic neuralgia risk increased with age and among those with severe immunosuppression, who remain contraindicated for zoster vaccination.

Abstract 1562  
UTILISING SCOTTISH HEALTH INFORMATICS TO ASSESS RISK FACTORS OF PROGRESSION TO CRITICAL CARE AND DEATH AMONG INDIVIDUALS WITH ACUTE PANCREATITIS (AP)

Authors: Peter T Donnan1, Usha Gungabissoon2, Philip Johnston3, Hester Ward3, Andrew M Lawton2, Damian J Mole4
1University of Dundee, 2Healthcare Information Factory, Observational Data Analytics Group, Worldwide Epidemiology, GSK, UK, 3edRIS, NHS National Services Scotland, UK, 4MRC Centre for Inflammation Research, Queen's Medical Research Institute, University of Edinburgh, UK

Presenting Author: Peter T Donnan

Corresponding author email: p.t.donnan@dundee.ac.uk

Background: Acute pancreatitis (AP) is a common and devastating inflammatory disease of the pancreas usually triggered by gallstones or alcohol misuse. Systemic complications can develop in AP that frequently requires support in critical care (CC). The evolution of a harmonized informatics infrastructure in Scotland has accelerated the development of a national unified health record. Our aim was to use this infrastructure to elucidate the epidemiology of AP in Scotland, with a specific focus on deterministic and prognostic factors for severe AP (sAP), that is, critical care admission or death.

Methods: A retrospective observational cohort design included individuals with at least one episode of AP (ICD10 K85) occurring in four Scottish Health Boards from 1st April 2009 to 31st March 2012. Data from...
sixteen sources, including general practice health records, community prescribing, A&E attendances, hospital in-patient, critical care, and mortality registries were linked by a unique patient identifier in a national safe haven. Logistic regression and gamma models were used to define independent predictive factors for critical care admission and mortality.

Findings: 3340 patients were coded as AP. 2053 (61.5%) met the definition for true AP (tAP). A disproportionate number of tAP admissions (35.3%) were from the most socially deprived quintile. 368 patients (17.9% of tAP) were admitted to CC at a median of one day after hospital admission. 101 patients (27.4% of tAP) had more than one CC admission during a single continuous inpatient stay. Median (1st, 3rd quartile) total CC stay was 107 (45, 221) hours. Factors which predicted CC admission in tAP were: history of angina or hypertension, hypocalcaemia at presentation, prior emergency admission with AP, age 20-29 years if concurrent type 2 diabetes mellitus, and age 40-49 years. Overall mortality in hospital in tAP was 102/2053 patients (5.0% of tAP). 22 patients died without being admitted to CC. Mortality in patients with tAP requiring CC was 80/368 (21.7%), of whom 54/368 (14.7%) died in CC and 26/368 (7.1%) died in hospital after discharge from CC. Factors independently predicting death in tAP patients admitted to CC were cardiovascular stenting and thyroid or parathyroid surgery within the previous year.

Interpretation: National record-linkage analysis from a wide variety of routinely collected health data sources is feasible and a powerful resource for public benefit. Factors that predict sAP are distinct from those that prognosticate death in hospital from AP. Mortality in patients with AP who require CC admission remains high.

Abstract 1785

LESSONS LEARNT: EVALUATION OF SCOTTISH HEALTHCARE DATA TO ELUCIDATE THE EPIDEMIOLOGY OF SEVERE ACUTE PANCREATITIS

Authors: Hester Ward¹, Usha Gungabissoon², Damian Mole³, Philip Johnston¹, Lynda Cochrane⁴, Grant Wyper¹, Leanne Hopkins¹, Christos Skouras⁵, Andrew M Lawton⁶, Peter T Donnan⁷

¹eDRIS, NHS National Services Scotland, UK, ²Worldwide Epidemiology, Quantitative Sciences, GSK, UK, ³MRC Centre for Inflammation Research, Queen’s Medical Research Institute and Clinical Surgery, University of Edinburgh, UK, ⁴Clinical Statistics Consultants, UK, ⁵Clinical Surgery, University of Edinburgh, UK, ⁶Healthcare Information Factory, Observational Data Analytics Group, Worldwide Epidemiology, GSK, UK, ⁷Dundee Epidemiology and Biostatistics Unit (DEBU), Division of Population Health Sciences, Medical Research Institute, University of Dundee, UK

Presenting Author: Hester Ward

Corresponding author email: hester.ward@nhs.net

Background: Acute pancreatitis (AP) is a common inflammatory disease of the pancreas usually triggered by gallstones or alcohol misuse. Approximately 20% of patients with AP develop a severe form of the disease, often with multiple organ failure. This group of patients require management in critical care with associated costs of approximately £100 million per annum in the UK. GlaxoSmithKline (GSK) and the University of Edinburgh have a Discovery Partnership with Academia collaboration with the aim of developing a potential therapy to minimise the risk of organ failure amongst individuals with AP. The reasons why a proportion of patients develop severe AP are poorly understood. The harmonised informatics infrastructure in Scotland has enabled the study of risk factors for the development of severe disease.

Aim of study: The aim of the present study was to provide a better understanding of the epidemiology of AP, to identify subsets of the AP population most likely to benefit from future medicines and inform early drug development activities.

Methods: This study was commissioned by GSK through the Farr@Scotland Institute. A retrospective analysis of linked observational healthcare data amongst individuals with ≥1 episode of AP occurring in
four Scottish Health Boards from April 2009 to March 2012 was performed. Data from sixteen sources, including primary care, community prescribing, accident and emergency, hospital in-patient, critical care, and mortality registries were linked by a unique patient identifier and analysed in the national safe haven. Additionally, a detailed case note review was required to verify the diagnosis of AP.

Challenges: This was the first major project to test the Farr@Scotland capabilities and several challenges had to be overcome.

- Development of protocol and statistical analysis plan, to meet needs of all stakeholders
- Non-standardised recording of data, in particular laboratory tests, across regions.
- Initial feasibility assessment to determine if data were of required quality
- There was no existing process for streamlining multiple simultaneous approvals (ethics, Privacy Advisory Committee etc.)
- Delays in obtaining certain data

Successes

- Project delivered within agreed timelines, successful study management Establishment of a national approvals system for Scotland
- Collaborative team approach with shared goals
- Application of data to gain greater disease understanding [etiology and progression of AP]

This study demonstrated the utility of linked patient data within the Farr@Scotland environment. It is hoped further applications of these data will continue to improve trust amongst data providers and data quality.

SESSION A5 - NATURALISTIC COHORT STUDIES.
CHAIR: DR JOHN MORIARTY, QUEEN’S UNIVERSITY BELFAST

Abstract 1727 EDUCATION AS A PREDICTOR OF MENTAL HEALTH AFTER BEREAVEMENT: A POPULATION BASED RECORD LINKAGE STUDY

Authors: Aideen Maguire1, Mark McCann2, John Moriarty1, Dermot O’Reilly1
1Queen’s University, Belfast. 2MRC Unit Glasgow

Presenting Author: John Moriarty
Corresponding author email: a.maguire@qub.ac.uk

Introduction: Bereavement affects almost everyone in their lifetime and is a psychologically stressful event that may pose a risk for longer term mental health problems. There are many factors that can influence how people cope with loss, including cognitive resources, stress, and the circumstances surrounding death. This study assesses the mediating effect of education on likelihood of poor mental health after bereavement based on various bereavement exposures.

Methods: This is a secondary data-linkage study utilising antidepressant drug uptake as an indicator of poor mental health. Data comprise Census information from the Northern Ireland Longitudinal Study (NILS), linked to mortality data (Northern Ireland Mortality Study) and antidepressant medication uptake data (Enhanced Prescribing Database), providing a picture of demographic and household factors, bereavement experiences and prescription drug information for a representative 28% sample of the NI population. A nested case control approach was taken. Multi-level logistic regression models were constructed to determine the likelihood of poor mental health given bereavement exposure by level of educational attainment, adjusting for factors known to be associated with mental health.

Results: Individuals who are bereaved have poorer mental health than those who are not bereaved. Over a quarter of those bereaved by suicide have subsequent poor mental health, as measured by
receipt of antidepressant medication. Education protects against poor mental health post bereavement, especially in cases of sudden death. Those bereaved by a sudden death who have a third level education are 73% less likely to have poor mental health compared to those with no qualifications (OR=0.27, 95%CI 0.09, 0.75). Education does not protect against poor mental health in those bereaved by suicide (OR=1.42, 95% CI 0.29, 7.00).

**Conclusion:** The relationship between education and mental health may be more complex than first assumed. Education is not merely an indicator of socio-economic status but more likely improves mental health outcomes by developing cognitive skills, increasing resilience and providing the personal and psychological skills necessary to prevent poor mental health reactions to acutely traumatic events. Being bereaved by suicide is a special case that cannot be rationalised in the same way that other bereavements can.

### Abstract 1630

**PROGRAMMED FROM BIRTH? EFFECTS OF MATERNAL SMOKING ON OFFSPRING REPRODUCTIVE HEALTH**

**Authors:** Sam Tweed¹, Sohinee Bhattacharya², Paul Fowler³  
¹University of Aberdeen, ²Dugald Baird Centre for Research on Women’s Health, University of Aberdeen, ³Chair in Translational Medical Sciences, Institute of Medical Sciences, Division of Applied Medicine, University of Aberdeen

Presenting Author: Sam Tweed

Corresponding author email: s.tweed.10@aberdeen.ac.uk

**Background:** Exposure to cigarette smoke in utero due to maternal smoking has been associated with reduced fertility in later life. Existing epidemiological studies are inconsistent, but generally suggest a small decrease in fertility of exposed women.

**Aims:** To compare the fertility-related outcomes of women exposed to cigarette smoke with women not exposed to cigarette smoke in utero.

**Methods:** The Aberdeen Maternity and Neonatal Databank (AMND) is comprised of routinely collected data on all deliveries in Aberdeen, Scotland since 1950. This retrospective cohort study included all women born on or before 31st December 1972 recorded in the AMND. Intergenerational record linkage was performed to identify if these women had themselves had any pregnancies recorded in the AMND over their reproductive lives. Women with one or more pregnancies were further compared according to exposure to cigarette smoke in utero on age at first delivery, total number of pregnancies and outcome of each pregnancy. Generalised estimating equations were used to generate adjusted odds ratios (OR) and 95% confidence intervals (CI) whilst accounting for clustering of data in sibling groups.

**Results:** In the cohort of 37,107 women, 12,321 had data recorded on mother’s smoking status during pregnancy. All results were adjusted for mother’s age at delivery and social class, as well as offspring year of birth, gestation and weight at delivery. Women whose mother smoked during pregnancy were more likely to have a pregnancy than those whose mother did not smoke; adjusted OR 1.25 (95% CI: 1.13-1.38). Women exposed to cigarette smoke in utero were also likely to have a delivery earlier (adjusted OR for age at first delivery ≤19 years 1.48 (95% CI: 1.28-1.71)), have more pregnancies in total (adjusted OR ≥3 pregnancies 1.18 (95% CI: 1.04-1.34)) and more live births overall than women who were not exposed (adjusted OR ≥3 live births 1.24 (95% CI: 1.06-1.44)). Women exposed to cigarette smoke in utero were significantly more likely to have a miscarriage than those not exposed; adjusted OR 1.25 (95% CI: 1.10-1.42).

**Conclusion:** In contrast to existing literature, this study suggests exposure to cigarette smoke in utero increases fertility in women. However, a lack of exposure data and residual effects of social class and education may present potential confounding factors. An increased risk of miscarriage in women exposed to cigarette smoke in utero is a novel finding with biological plausibility, demanding further scientific exploration.
Abstract 1695

IS POOR EDUCATIONAL ATTAINMENT A RISK FACTOR FOR ADVERSE HEALTH IN ADOLESCENCE?

Authors: Joanne Demmler¹, A Bandyopadhyay¹, M Rahman¹, A Evans², S Paranjothy², M A Healy³, R A Lyons¹, S Brophy¹
¹Swansea University, ²Cardiff University, ³Swansea University

Presenting Author: Joanne Demmler

Corresponding author email: j.demmler@swansea.ac.uk

Objectives: This study examines if education attainment in primary school can be used as a predictor of adolescent health.

Methods: The Wales Electronic Cohort for Children (WECC) was linked to the Department for Children, Education, Lifelong Learning and Skills (DCELLS) dataset to obtain key stage 1 and key stage 2 results for 172,421 children for the years 2005 to 2013. This cohort was then stratified and compared in terms of: (i) those that achieved key stage 1 and key stage 2 (achieving), compared to those who achieved key stage 1 but not key stage 2 (declining in education) and (ii) those that did not achieve key stage 1 but did achieve key stage 2 (improving) and those who did not achieve key stage 1 or key stage 2 (not achieving). Subsequently these children were then linked to primary and secondary care admissions for injuries between the ages of 12 and 18. The hazard risks were then calculated using Stata 12.

Results: The children who achieved key stage 1 but not key stage 2 (i.e., declining in education) were at higher risk of injury in adolescence (that resulted in a hospital admission) compared to those who achieved both stages (Hazard Ratio, 1.28 [95% CI: 1.13 to 1.44] – boys, 1.27 [95% CI: 1.06 to 1.50] girls. There was no significant difference between children who did not achieve key stage 1 (but then went on to achieve key stage 2 – improving in education, and those who continued to not achieve – not achieving). Declining students exhibited more injuries related to cuts, burns and substance abuse and were more likely to self-harm, whilst improving students exhibited more sports injuries, falls and bumps, which were more likely to be unintentional. Not achieving key stage 1 was furthermore associated with more motor vehicle traffic crash injuries and more assaults.

Conclusion: Children at higher risk of intentional injuries (e.g., substance abuse, self-harm & assaults) in adolescence can be identified in their primary school years as those who initially were doing well in education but then decline in achievement. Thus, poor adolescent health/ risk behaviours in adolescent can be predicted from education achievement in primary school.
Abstract 1609

SCOTTISH LEARNING DISABILITIES OBSERVATORY: RAISING THE VISIBILITY OF PEOPLE WITH LEARNING DISABILITIES IN SCOTTISH DATASETS

Authors:

S-A Cooper¹, C McCowan¹, A Henderson¹, L Hughes-McCormack¹, M Jacobs¹, L Allan², J Pell¹

¹Institute of Health and Wellbeing, University of Glasgow, ²Directorate of Population Health Improvement, Care, Support & Rights Division, Scottish Government

Presenting Author: S-A Cooper

Corresponding author email: Sally-Ann.Cooper@glasgow.ac.uk

Background: People with learning disabilities are thought to experience premature death, and substantial health inequalities compared with the general population. England’s recent confidential inquiry into deaths of people with learning disabilities has shown that many are potentially preventable. Scotland’s rich, routinely collected health statistics do not include an identifier for learning disabilities, but linkage to other created data sources could provide a major route forwards towards understanding their poor health, its determinants, potential solutions, and trends over time. The Scottish Government has funded a new Scottish Learning Disabilities Observatory from March 2015, to start to take this work forward.

Methods: We identified and compared datasets recording people with learning disabilities, and scope for record linkage. Work streams include trends over time in detection of learning disabilities and autism in children (Scotland’s pupil census data); trends over time in terminations of trisomy pregnancies (clinical genetics services data); management of long term conditions in primary care (Greater Glasgow and Clyde primary care data extraction); multi-morbidity (data from the Primary Care Clinical Informatics Unit at the University of Aberdeen); household, employment, health and personal characteristics (national census data); and dental health data. Work in progress includes work with SPIRE for a national bespoke data extraction of primary care data, and with Analytic Services Division and Scottish Consortium for Learning Disabilities using data on people known to Local Authorities as having learning disabilities.

Results: Scotland’s pupil census shows consistent rates of learning disabilities in recent years, but markedly increasing identification of autism. Matching to SMR2 reveals learning disabilities to be associated with low apgar, and low birth weight.

Primary care management of long term conditions is substantially poorer across all long-term conditions for adults with learning disabilities (n=5,800) compared with the general population. Initial data extraction revealed some challenges, but was considerably more efficient than methods used in 2010 when individual consenting was undertaken to answer the same question on a subset of the population (n=700).

The extent of multi-morbidity in 20 year olds with learning disabilities is similar to that of 50 year olds in the general population (Aberdeen data).

Data will be available for presentation.

Conclusions: This work is at an early stage, and shows both considerable potential, and challenges to address. People with learning disabilities have been a hidden population in routine statistics for too long, and this new focus on their needs is a positive development.
INCIDENCE AND DETERMINANTS OF MENTAL HEALTH DISORDERS IN CHILDREN OF PARENTS WITH MULTIPLE SCLEROSIS

Authors: Neda Razaz1, Helen Tremlett1, Thomas Boyce2, Martin Guhn1, Ruth Ann Marrie3, K.s. Joseph1
1University of British Columbia, 2University of California, 3University of Manitoba

Presenting Author: Neda Razaz

Corresponding author email: neda.razaz@gmail.com

Purpose: Early-life stressors can have harmful effects on the social and emotional functioning of children. Although parental multiple sclerosis (MS) may put children at increased risk for mental health disorders such as anxiety and depression, the incidence and determinants of such disorders have not been determined.

Methods: We carried out a retrospective matched cohort study in British Columbia, Canada, using linked population-based education and health databases. Children who were born between 1993 and 2006, with an MS parent, were compared with age-matched children whose parents did not have MS. Parents with MS were identified using a validated algorithm as those with ≥3 health claims related to MS. The outcome of interest was the diagnosis of depression or anxiety disorders in children identified through hospital, physician and prescription drug claims. Children were followed from the index date until the first diagnosis of a mental health disorder, emigration from British Columbia or 31 December 2011. Cox regression was used to estimate adjusted hazard ratios (aHR) and 95% confidence intervals (CI). Further the Sobel test was used to assess the potential mediation of the relationship between parental MS and mental health disorders in children by parental mental health morbidity.

Results: The study included 1,028 children of MS parents, 4,010 children of unaffected parents, and 25,464 child-years of follow up. Mental health disorders were more common among MS parents vs. unaffected parents (50.4% vs. 33.1%, P<0.001) and among MS affected mothers vs. MS affected fathers (54.6% vs. 40.2%, P=0.01). The incidence of mental health disorders in children was 8.3 and 6.3 per 1000 child-years among children of parents with and without MS, respectively. Sex of the MS affected parent modified the relationship between parental MS and mental health disorders in children (P=0.04). Compared with children of unaffected mothers, children of mothers affected by MS had higher rates of mental health disorders (HR 1.65, 95%CI: 1.14-2.39), whereas children of fathers affected by MS did not (HR 0.51, 95%CI: 0.15-1.74). The Sobel test showed that maternal mental health disorders significantly mediated the association between maternal MS and mental health disorders in children (P <0.001).

Conclusion: Maternal MS is associated with a higher rate of mental health disorders in children and this association is mediated by maternal mental health morbidity. Prevention efforts aimed at mitigating the impact of parental MS on mental health disorders in children should address the psychological needs of MS parents and their children.
SESSION A6 – LINKED OBSERVATION FOR HEALTH POLICY.
CHAIR: MICHAEL WALLER, THE UNIVERSITY OF QUEENSLAND

Abstract 1506
STANDARDISED MORTALITY RATIOS FOR SOCIALLY EXCLUDED POPULATIONS: MAKING THE INVISIBLE VISIBLE THROUGH DATA LINKAGE

Authors: Andrew Hayward¹, R Aldridge², A Story³
¹Farr London, UCL Institute of Health Informatics, ²Farr London, UCL Institute of Health Informatics, ³Farr London, UCL Institute of Health Informatics

Presenting Author: Andrew Hayward

Corresponding author email: A.hayward@ucl.ac.uk

Background: Homeless people, prisoners, drug users and sex workers are key exemplars of overlapping groups suffering from extreme exclusion. Geographical indices of social inequality are frequently used to identify social gradients in health but these extremely excluded populations are not routinely identifiable in health data statistics, effectively making their needs invisible to policy makers.

Methods: We reviewed the international literature to identify studies from high income countries that had used linkage of datasets of these populations to mortality datasets to infer standardised mortality ratios compared to the general population. ICD 10 disease chapters are classified according to summary estimates across population and disease groups within each chapter to identify disease categories with evidence of increased mortality.

Results: Most information was available for homeless populations, but where available results in the other, overlapping populations were similar. For socially excluded men standardised mortality ratios of: between 1.5 and 2 were seen for neoplasms, endocrine system, nervous system, and circulatory system disease; between 2 and 5 for infectious and parasitic disease, diseases of blood forming organs, respiratory system and gastrointestinal system disease; and between 5 and 10 for mental and behavioural disorders and deaths due to external causes, and injury and poisoning. For socially excluded women standardised mortality ratios of between 1 and 1.5 were seen for neoplasms; between 1.5 and 2 for infectious and parasitic disease and circulatory disease; between 2 and 5 for endocrine system, mental and behavioural, nervous system respiratory system, gastrointestinal system diseases and external causes; and between 10 and 20 for injury and poisoning. All cause standardised mortality ratio was 2.76 (95% CI 2.70-2.82) in socially excluded men and 3.45 (95% CI 3.25-3.65) in socially excluded women.

Conclusion: Linkage of datasets from socially excluded groups to health datasets uncovers extreme health inequalities and is essential information to advocate for measures to prevent exclusion and provide for the health needs of socially excluded groups. Challenges and progress on establishing linkages between UK homeless data sets and mortality and hospital episode statistics within an NIHR funded grant will be presented.
Abstract 1481  

**Hip fracture risk in chronic kidney disease**

**Authors:** Lynn Robertson¹, Angharad Marks¹, Corri Black¹, Sharon Gordon¹, Rosemary Hollick², Gordon Prescott¹  
¹University of Aberdeen, ²NHS Grampian

**Presenting Author:** Angharad Marks

**Corresponding author**  
email: l.robertson@abdn.ac.uk

**Introduction:** Chronic kidney disease (CKD) has many complications. For those on renal replacement therapy (RRT) there is an increase in the risk of fracture. The evidence regarding those with less advanced CKD is not so clear. We aimed to understand the association between CKD and hip fracture-related hospital admissions and mortality using data linkage methods.

**Research questions**

1. Are there increased admissions for hip fractures in patients with CKD?
2. What proportion of hospital admissions of people with hip fractures could be attributed to the presence of CKD?
3. Does having CKD increase the risk of death in patients with hip fractures?

**Methods:** This is a longitudinal cohort study using the established Grampian Renal Observatory (datasets), including the Grampian Laboratory Outcomes Mortality and Morbidity Study (GLOMMS-II). Data-linkage to hospital episode statistics and National Records of Scotland allowed ascertainment of comorbidities, hip fracture and mortality rates. In addition to the descriptive analysis included here, Poisson regression will be used to provide IRRs for the different levels of renal function, unadjusted and adjusted for baseline characteristics. Population attributable risk will be investigated, as will mortality rates in those who suffer fractures.

**Results:** Of 70,906 individuals, there were 19,775 with normal renal function (47% males, mean age 52 years) and 19,635 with CKD stage 3-5 (35% male, mean age 75 years). There were 1,745 individuals who had at least one hip fracture during follow-up. People with CKD have higher rates of hip fracture than those with normal renal function (IRR 6.90 [95%CI 5.83, 8.15] in univariate analysis. Adjusted analysis will seek to address potential confounding by age and gender.

**Implications of research:** This is the first in-depth, population level study of the risk of renal bone disease in a Scottish population, and is therefore directly applicable to the Scottish healthcare setting. The results of this study will quantify the workload faced by the NHS as a result of CKD related bone disease. We will identify and characterise the patients at greatest risk.

This project showcases Scotland’s ability to link healthcare data for long term follow up of patient cohorts; it will allow the research potential of linked data for future studies of appropriate therapies and recommendations for clinical practice to be explored.
Abstract 1547  EXPLORATORY DATA LINKAGE STUDY OF PALLIATIVE CARE REGISTER USE IN PRIMARY CARE

Authors: Ting Wang, Sarah Rees, Daniel S Thayer, Mathilda Castagnet, Susie Lucas
SAIL Databank, College of Medicine, Swansea University

Corresponding author email: t.wang@swansea.ac.uk

Background: The Palliative Care Register (PCR) is a Quality and Outcomes Framework (QOF) indicator requiring GP practices to maintain a list of patients expected to die within 12 months. There is evidence suggesting that compared to patients with cancer, patients with other conditions are less likely to be on a PCR and are added at later stages of disease. General practitioners generally do not have access to detailed outcome data for those they place on the PCR, a gap that can be filled using linked data.

The study aimed to describe the quality of available data as well as assess aspects of care provided.

Objectives: Provide a report to GP practices showing how palliative care at their practice compares to their health board and Wales, with a particular focus on:

- Rate of registration on PCR by disease group and cause of death
- Duration on PCR prior to death
- Place of death (hospital or other), comparing those on a PCR with those not on a PCR

Methods: Data from ONS mortality, primary Care and the Welsh Demographics Service were linked at patient level.

PCR Read codes specified in QOF guidelines were used to create a list of patients added to a PCR between 2008 and 2012. Data for deaths deemed as predictable were extracted from ONS Mortality Records. Primary care diagnoses identified patients on the PCR who had cancer, congestive heart failure, lung disease or dementia.

Death data and primary care data were linked to derive survival time from PCR registration to death.

Results: Of patients added to a PCR between 2008 and 2012, 72% had a code for cancer in their GP history, 15% for lung disease, 9% for congestive heart failure and 6% for dementia.

Of predictable deaths between 2008 and 2012, 18% of patients who died from cancer were on a PCR compared with 3% of those with other conditions.

An association was found between higher PCR use at a practice and a lower rate of death in hospital (t=2.446; p=0.01; SE – 0.068) which persisted even after controlling for age and deprivation.

Conclusions: These results suggest an under-identification in the PCR of patients with diseases other than cancer and a greater duration on the PCR among cancer patients. The results may further suggest that use of the PCR leads to fewer deaths in hospital, a desirable outcome, though more evidence is needed to establish causality.
Abstract 1738

PREVALENCE AND TREATMENT OF ACTIVE ASTHMA IN SCOTLAND USING THE PRESCRIBING INFORMATION SYSTEM

Authors: Markus Steiner¹, G Devereux¹, S Turner¹, J McLay², G Wyper³

¹University of Aberdeen, ²University of Aberdeen, ³NHS National Services Scotland

Presenting Author: Markus Steiner

Corresponding author email: m.steiner@abdn.ac.uk

Background: Many studies of the clinical epidemiology of asthma that have used routinely collected drug prescription or dispensing data have been limited to samples assumed to be representative of the national population from which they are drawn. Our aim was to describe asthma prevalence and treatment in children and young adults using the Prescribing Information System (PIS), a national prescribing and dispensing database for Scotland.

Methods: For more than 95% of the dispensed prescriptions in primary care from December 2009 a valid patient identifier is available and the database includes some socio-demographical characteristics (age-group, sex, SIMD) of the patients. Data were also linked to hospital admission and A&E data for the identification of exacerbations. The analysis was limited to patients aged up to 50 years to reduce contamination by COPD.

Data: We identified from the first extract until 2012 358,804 patients with 2,809,563 dispensed prescriptions for inhaled therapies used for asthma; equating to a prevalence of 11.4% of the 3,139,356 people aged 0-44 registered with a GP in Scotland. However, 95,207 patients had only one or two dispensed prescriptions for short-acting beta2-agonists (SABA) and no other inhaled therapies in the two years; we consider these patients to be unlikely to have active asthma (table). Additionally, 1,041 cases on inhaled therapy had hospital admission(s) with a diagnosis of COPD (ICD10: J40-J44) and are excluded from further analysis. 6,056 (2.3%) of people collecting inhaled therapy (>2 SABA) had at least one hospital admission with a primary diagnosis of asthma.

Conclusion: This current and whole population database indicates that the prevalence of asthma is approximately 10% in young adults and 15% in children living in Scotland but prevalence of active asthma is approximately 8% in adults and 10% in children. The dataset has been recently updated for another year and will allow us to present updated figures and allow some refinement of our results presented so far.
THE CANADIAN CHRONIC DISEASE SURVEILLANCE SYSTEM (CCDSS): METHODS FOR NATIONAL DISEASE SURVEILLANCE WITH LINKED ADMINISTRATIVE HEALTH DATABASES

Authors:
Lisa M. Lix\footnote{University of Manitoba}, Kim Reimer\footnote{British Columbia Ministry of Health, Vancouver, British Columbia, Canada}

Presenting Author: Lisa M. Lix
Corresponding author email: lisa.lix@umanitoba.ca

Introduction: The Public Health Agency of Canada (PHAC) established the Canadian Chronic Disease Surveillance System (CCDSS) in 2009. The CCDSS uses population-based linked health administrative databases in all provinces/territories (P/Ts) to report the burden and use of health services and health outcomes for multiple chronic diseases. The CCDSS is undergoing rapid expansion, leading to new methodological challenges. Our purpose is to describe the current structure and future directions of this successful national and P/T collaboration.

Methods: The CCDSS is founded on deterministic linkage of three administrative databases in each Canadian P/T via a unique lifetime identifier: health insurance registration files, physician billing claims, and hospital discharge abstracts. Disease case definitions are developed by expert Working Groups after a comprehensive literature review, diagnostic validation studies, and feasibility studies. Analytic code developed by PHAC is distributed to all P/Ts to ascertain disease cases. The extracted summary data are approved by Scientific Committee and Technical Committee members in each P/T and then submitted to PHAC for further analysis and reporting.

Results: National surveillance is ongoing for diabetes, hypertension, selected mental illnesses, and chronic respiratory diseases. Pilot studies have been conducted for heart disease and musculoskeletal conditions as a precursor to national surveillance. CCDSS working groups and committees are instrumental to success because they collaborate with partners in disease surveillance, including Aboriginal groups, universities, and health associations, and contribute expertise about administrative health data quality and linkage. The advantages of the distributed analytic approach are that small changes in methodology can be easily made and technical expertise is not required in each P/T. Future expansion plans include neurological conditions, heart failure, stroke, injuries, and multimorbidity (i.e., co-occurrence of two or more chronic conditions). Methodological challenges include lack of completeness of disease case capture in physician billing claims, poor measurement validity of diagnosis codes for some chronic conditions, and the balance between comprehensively reporting all cases and following disclosure guidelines to ensure data confidentiality.

Conclusions: Other jurisdictions can learn from the successful implementation of the CCDSS model. A multi-region disease surveillance system to monitor a broad range of health conditions on a routine basis can contribute to improvement public health policy, programs, and practice.
SESSION B1 – STATISTICS & ANALYTICS.
CHAIR: DR HILIARY PINNOCK, UNIVERSITY OF EDINBURGH

Abstract 1620 DEVELOPMENT AND VALIDATION OF PROGNOSTIC MODELS FOR ATEROThROMBOTIC EVENTS AND HOSPITALISED BLEEDING IN STABLE MYOCARDIAL INFARCTION SURVIVORS

Authors: Laura Pasea1, Sheng-Chia Chung1, Mar Pujades Rodriguez1, Em Jennings2, Cathy Emmas3, Harry Hemingway1
1Farr Institute for Health Informatics Research, University College London, London, UK, 2Global Payer and Real World Evidence, AstraZeneca R&D, 3Real World Evidence, AstraZeneca UK

Presenting Author: Laura Pasea
Corresponding author: l.pasea@ucl.ac.uk

Background: The risk of recurrent cardiovascular (CV) events and bleeding remains high in stable myocardial infarction (MI) survivors in the US, England, Sweden and France [Rapsomaniki 2014]. However there are no prognostic models to support risk stratification of this population. The study aim was to develop and validate prognostic models for atherothrombotic events and bleeding in MI survivors in England.

Methods: We examined linked electronic health records (CALIBER 2000-2010) of patients who had survived 1 year after their last MI. Weibull parametric survival regression models including demographics, clinical history, clinical biomarkers and medications were used. C-indexes were used to estimate model discrimination. To assess model calibration the validation cohort was divided into risk groups according to cut points at the 16th, 50th and 84th percentiles of the development cohort linear predictor. The two smaller groups at the top and bottom 16th percentiles correspond to low and high risk and the two larger central groups correspond to medium low and medium high risk. The unequal grouping retains more information than equal grouping. [Cox 1957] For each model the predicted events were compared with the observed events in the validation cohort by risk group, graphically.

Results: 12,694 patients were in the development cohort from 159 general practices in South England and 5,613 in the validation cohort from 61 practices in the North. The 5-year Kaplan-Meier event estimates were 22.4% [95% CI: 21.4%-23.3%] for CV death/stroke/MI and 7.2% [6.6%-7.8%] for hospitalised bleeding. In validation the c-indexes of the models were 0.75 [0.74-0.77] and 0.66 [0.63-0.70] and the hazard ratios for the high: low risk contrast were 11.5 [8.4-15.5] and 6.0 [3.6-10.2], respectively. The models were well calibrated. Although the models have largely different prognostic factors, the validation cohort were found to have moderately concordant cardiovascular and bleeding predicted risk groups.

Conclusion: Prognostic models derived from population based electronic health records allow us to risk stratify patients and may aid balancing benefits and harms for care and enable effective long-term clinical management after acute MI.
IDENTIFYING INCIDENT CASES OF DEMENTIA IN A LARGE BIORESOURCE USING A HEURISTIC PHENOTYPE ALGORITHM IN COMBINATION WITH GENETIC INSTRUMENTS

Authors: Wilfred Bonney1, Colin N. A. Palmer1, Peter Connelly2, Emily Jefferson1, Alex S. F. Doney1
1University of Dundee, 2NHS Tayside

Presenting Author: Wilfred Bonney
Corresponding author email: w.bonney@dundee.ac.uk

Background and Objective: Dementia has catastrophic implications for affected individuals, their family and wider society. Combining information about genes with detailed clinical information in large populations provides a powerful way of investigating why people vary in susceptibility to dementia and how they will respond to treatments. However, patients with dementia are challenging to study because their diagnoses are not easily observable in the electronic medical records (EMRs) and so are greatly underrepresented in large population-based bioresources. The objective of this study was not only to develop a heuristic phenotype algorithm to identify patients who have developed dementia since being recruited into the large Genetics of Diabetes Audit and Research in Tayside Study (GoDARTS) bioresource, but also to use established genetic instruments for validation of the phenotype.

Methods: A heuristic phenotype selection algorithm was applied iteratively on five selected datasets (i.e. General Registrar Office - Death Certification (GRO), Prescribing, and Scottish Morbidity Records: SMR00, SMR01, and SMR04) from the GoDARTS bioresource. Available genomic data was linked to EMR for the GoDARTS’ participants using the HIC Research Data Management Platform (RDMP). Cox proportional hazards’ models were constructed to estimate the risk of dementia association with the apolipoprotein E epsilon 4 (ApoE4) single nucleotide polymorphism (SNP) and a weighted Genetic Risk Score (wGRS), calculated using 20 known susceptibility SNPs associated with Alzheimer’s disease.

Results: The algorithm identified 855 definite cases of dementia from the GoDARTS population of 18,190. Amongst the 855 dementia cases identified, 446 individuals were found to be genotyped for both the ApoE4 and the wGRS SNPs. In a Cox regression model involving age and gender as covariates, the study demonstrated that the ApoE4 (Hazard Ratio (HR) = 1.98; 95% Confidence Interval (CI): 1.59–2.46; p-value = 1.01 x 10^-9) and the wGRS (HR=1.63; 95% CI: 1.14–2.32; p-value= 7.04 x 10^-3) were independently associated with dementia, with non-statistically significant gender differences. The combined model of the wGRS and the ApoE4 also indicated a significant association with dementia (HR=2.13; 95% CI: 1.60–2.84; p-value=2.31 x 10^-7), with non-statistically significant gender differences. The overall incidence rate of dementia in the GoDARTS bioresource was found to be 7.28 per 1000 person-years.

Conclusion: The study demonstrates that it is possible to identify patients with dementia from large population-based bioresources using genetic instruments for validation. Such an approach can generate a dementia specific sub-cohort for the international research community to support translational research.
Abstract 1683

PROGNOSTIC VALUE OF TRENDS IN BLOOD PRESSURE: A JOINT MODELLING STUDY OF 85,428 MEASUREMENTS IN 5,246 PATIENTS USING ELECTRONIC HEALTH RECORDS

Authors: Michael J Crowther1, Anoop D Shah2, Keith R Abrams1, Harry Hemingway2
1University of Leicester, 2Farr Institute of Health Informatics Research, University College London

Presenting Author: Michael J Crowther

Corresponding author email: mjc76@le.ac.uk

**Background:** Electronic health records frequently contain repeated clinical measurements on patients, yet most prognostic models consider only a single static value of a biomarker. In cardiovascular disease little is known about how to model both trends over time, and also their association with clinical endpoints – thus opening up the possibility of real-time risk prediction and subsequent treatment modification. We aimed to determine among type 2 diabetic patients with stable angina, the extent to which current values of systolic blood pressure (SBP) and trends over time, add information beyond a single baseline value in the prediction of death, myocardial infarction (MI) or stroke.

**Patients:** 5,246 patients diagnosed with stable angina and type 2 diabetes mellitus (T2DM) in England between 1997 and 2010, with a total of 85,428 BP measurements

**Setting:** Population based electronic health records in England linking primary care, hospitalisation registry, heart attack and death registries (the CALIBER programme).

**Methods:** We used joint longitudinal-survival analysis techniques to model repeatedly measured systolic blood pressure and the time to composite end point (all-cause mortality/non-fatal myocardial infarction/non-fatal stroke).

**Results:** We found 1,748 events during a median follow-up of 3.33 years (max: 12.22 years). The median number of SBP measures per patient was 12 (range: 1, 336). An average decreasing trend in SBP was found over time (reduction of -0.965 mmHg (95% CI: -1.088, -0.844) per year for a male aged 68 with a body-mass index (BMI) of 30 kg/m² and HbA1c of 58 mmol/mol, no previous heart failure, no use of loop diuretics or BP lowering medication). The current value of systolic blood pressure (association = -0.012 [95% CI: -0.017,-0.006]) and rate of change (association = -0.122 [95% CI: -0.166,-0.077]) of SBP over time are strongly associated with the risk of death, MI or stroke, independent of age, sex, BMI, heart failure, loop diuretic use, SBP lowering medication, and HbA1c level.

**Conclusions:** Lower values of SBP and a declining trend in SBP may play an important role in identifying patients at increased risk. Joint modelling can provide an effective tool for investigating the association between a repeatedly measured biomarker and a time to event, providing a dynamic risk prediction framework.
Abstract 1745

**A RISK CALCULATOR FOR MULTI-DRUG RESISTANT TUBERCULOSIS IN LONDON**

**Authors:** Oliver Dukes1, A Hayward1, A Story2, C Anderson3, H Maguire3, M Pavlou4
1Farr Institute of Health Informatics Research, University College London, 2University College London Hospitals NHS Foundation Trust, 3Public Health England Health Protection Directorate: Field Epidemiology Service (Victoria Office, 4University College London Department of Statistical Science

**Presenting Author:** Oliver Dukes

**Corresponding author email:** o.dukes@ucl.ac.uk

**Background:** Rapid nucleic acid amplification technologies provide a fast and effective way to diagnose multidrug-resistant tuberculosis (MDR-TB). However, there is concern about the number of false positive results produced by over-testing in low risk groups. Treatment for MDR-TB is time-consuming and expensive, with severe side effects for patients. There is therefore a need for a reliable and readily accessible risk calculator that can be used by clinicians in order to effectively target rapid testing.

**Methods:** Using Enhanced TB Surveillance data collected through the London TB Register from 2009-13, a multivariable risk prediction model was developed based on a priori risk factors of MDR-TB in London. Pre-specified interactions were fitted and restricted cubic splines were used to flexibly model continuous predictor variables. Bootstrapping was used to internally validate the model and decision-analytic techniques enabled exploration of how the calculator may be used effectively in a clinical setting.

**Results:** The dataset included 7,292 individuals, 173 (1.83%) of whom tested positive for MDR-TB. Using a multivariable penalised logistic regression model, the strongest risk factors associated with MDR-TB included previous diagnosis of TB, the burden of MDR-TB in a patient’s country of birth, length of time in the UK if foreign born and age. The effect of age was non-linear, with odds of MDR-TB highest in adolescents and young adults. The calibration slope was 0.90 after internal validation, with a small decrease in the c-statistic from 0.72 to 0.70. Non-parametric calibration curves indicated good performance for the majority of patients, but with overestimated risk for higher risk individuals. The model has been developed into a working calculator using Microsoft Excel.

**Conclusions:** The estimates from the model are the basis for a calculator which predicts individual risk. We plan to integrate this calculator within the London TB Register so that it becomes embedded in clinical practice. Future work could include external validation of the calculator using TB data from other European cities. There is also a need for health economic work to help identify a range of cut-off risks at which testing is worthwhile.

Abstract 1777

**LEARNING TO CARE: USING MACHINE LEARNING TO IMPROVE PREDICTION OF COPD ADMISSIONS**

**Authors:** Hilary Pinnock1, Felix Agakov2, Peter Orchard2, Anna Agakova2, Mary Paterson1, Lucy McCloughan1, Chris Burton3, Brian McKinstry1
1Usher Institute of Population Health Sciences and Informatics, University of Edinburgh, Edinburgh, 2Pharmatics Ltd, Edinburgh Bioquarter, Edinburgh, 3Centre of Academic Primary Care, University of Aberdeen, Aberdeen

**Presenting Author:** Hilary Pinnock

**Corresponding author email:** hilary.pinnock@ed.ac.uk

**Background:** Telehealth aims to detect imminent exacerbations in order to facilitate prompt action to prevent admissions. However, recent randomised trials have failed to demonstrate reductions in admissions when telehealth has been applied and resulted in large numbers of alerts which did not
require clinical intervention (false positives). Pilot work suggested that current algorithms, based on simple additive methods were be poor predictors of exacerbations. We aimed to develop a more advanced algorithm to estimate risk of COPD-related hospital admissions with better specificity/sensitivity.

**Methods:** We linked an existing telemonitoring dataset of 133 COPD patients monitored on average for 430 days, with their baseline data from the randomised controlled trial (including demographic and assessments of illness severity) and data extracted from their electronic health record. Using this enhanced dataset, we developed a probabilistic machine learning algorithm to predict next-day admissions due to COPD. The algorithm used 44 features extracted from time series of telemonitoring measurements. The majority of patients in our dataset had 0-2 admissions due to COPD per annum, so the algorithm was designed to address classification of imbalanced data. We considered the complete case and imputed scenarios. The quality of predictions was evaluated by 10-fold nested cross-validation. Test folds included only previously unseen patients, and were used neither for feature selection nor for tuning parameters of the predictive algorithm.

**Results:** We compared our machine learning algorithm with two standard symptom-counting algorithms. In one of them, next-day admissions were predicted when the number of symptoms observed on a given day exceeded a pre-defined threshold. In the other one, admissions were predicted when two days of elevated symptom-based scores were preceded by two days of normal scores. For both the complete-case and the imputed scenarios, our algorithm demonstrated significant improvements in the prediction of future admissions over these symptom-counting methods. In the imputed scenario, the two standard symptom-counting algorithms resulted in the test AUC of approximately 0.50, with the slightly different confidence intervals of [CI95% 0.49, 0.50] and [0.46, 0.53]. Our algorithm resulted in a test AUC=0.73 [CI95% 0.67, 0.79].

**Conclusion:** Our machine learning algorithm has significantly improved the ability of telemonitoring to predict COPD admissions. This offers the potential to improve the effectiveness both of telehealth and COPD self-management. On-going research is observing the impact of including meteorological and viral load data in the algorithm to see if this further improves predictivity.

**Funding:** MRC Confidence in Concept
the individual in their home as part of an interview assessing for housing repairs.

**Results:** We report that address data from a patient-provided routine data source are reliable when compared to interview data for a cohort of 80,000 mainly older individuals undergoing interview assessment for home improvements.

**Conclusion:** The convenience sample of older adults, who would like to remain in their own home, provides a conservative estimate. Further work is needed to assess younger, more mobile population groups. The evaluation of patient-provided routine data against addresses collected during interviews is a useful first investigation into the accuracy of a patient-population routine data source.

---

**Abstract 1678**

**ESTABLISHING THE ‘MISSING LINKAGE’ WITH IMAGING IN SCOTTISH HEALTH DATA: A PILOT FOR INTRACEREBRAL HAEMORRHAGE**

**Authors:** David Bailey¹, Rustam Al-Shahi Salman², Aidan Hutchison³, Hester Ward⁴, Steve Pavis⁵, David Rodriguez Gonzalez⁶, Gordon Sommerville⁶, John Macdonald⁶, Alan Fleming⁶, Lorna Ramsay⁶, Janet Murray⁶, Jackie Caldwell⁶


**Presenting Author:** David Bailey

**Corresponding author email:** david.bailey5@nhs.net

**Background:** The Electronic Data Research & Innovation Service (eDRIS) provided by NHS National Services Scotland (NSS) allows access to linked datasets in a secure environment. The existing infrastructure facilitates linkages between many data sources, but not one of Scotland’s most recently established national datasets, the national Picture Archiving and Communications System (PACS), containing imaging from every Scottish NHS Board.

Imaging data are stored in Digital Imaging and Communications in Medicine (DICOM) format, which is very different from the format of other national datasets currently used by NHS NSS. The MRC has funded a pilot project to develop the systems and processes necessary to link imaging data from Scotland’s national PACS to healthcare data for patients with intracerebral haemorrhage (ICH) and so provide a de-identified research dataset.

**Aim:** The project aims to acquire the permissions, develop the technologies, and create a ‘missing linkage’ from NHS Scotland healthcare datasets to imaging on the national PACS, thus enabling analysis in the public interest. It will:

1. test an approach to acquiring and anonymising imaging studies associated with a linked dataset;
2. create a dataset to enable audit, quality improvement and research projects on ICH;
3. determine the feasibility of future, similar linkages for other diseases;
4. if successful, establish a modus operandi for integrating imaging into linkages that are currently undertaken for administrative purposes in Scotland.

**Method:** Patients suffering from ICH will be identified using the Scottish Stroke Care Audit (SSCA) and their relevant CT and MR imaging will be extracted from the national PACS. These images will be anonymised using DICOM Confidential, an open source, policy driven, DICOM anonymisation tool developed by collaborators at Informatics and the Brain Research Imaging Centre (BRIC) at the University of Edinburgh. Following successful anonymisation the de-identified images will be transferred to the national safe haven where they can be viewed and linked to data from the SSCA, records of hospital admissions and discharges, death records, and community prescribing data. This linked
dataset will then be used to fulfil the project’s clinical aims.

**Next Steps:** The presentation will give an update on our progress with this pilot project. We will detail the steps remaining and discuss our experience thus far and its bearing on future Scotland-wide linkages to PACS imaging datasets.

**Abstract 1577**

**HOW WELL DOES NOTIFICATION DATA CORRELATE WITH LABORATORY DATA FOR NOTIFIABLE RESPIRATORY INFECTIONS?**

**Authors:** Faye Janice Lim¹, C C Blyth², N de Klerk³, P Fathima³, H C Moore³

¹Wesfarmers Centre of Vaccines and Infectious Diseases, Telethon Kids Institute, The University of Western Australia, ²Wesfarmers Centre of Vaccine and Infectious Diseases, Telethon Kids Institute, The University of Western Australia; School of Paediatrics and Child Health, The University of Western Australia; Department of General Paediatrics, Princess Margaret Hospital for Children; PathWest Laboratory Medicine WA, Princess Margaret Hospital for Children, ³Wesfarmers Centre of Vaccine and Infectious Diseases, Telethon Kids Institute, The University of Western Australia

**Presenting Author:** Faye Janice Lim

**Corresponding author email:** Janice.Lim@telethonkids.org.au

**Introduction:** Vaccine preventable diseases such as influenza, pertussis and invasive pneumococcal disease, substantially contribute to morbidity and mortality. As such, they are notifiable diseases in Western Australia. These mandatory notifications are recorded on the Western Australian Notifiable Infectious Diseases Database (WANIDD), which is managed by the Western Australia Department of Health. PathWest Laboratory Medicine (PathWest) is the only state-wide public laboratory system in Western Australia undertaking testing on the majority of diagnostic samples in children.

By linking PathWest and WANIDD data, we sought to quantify how well WANIDD captured laboratory-confirmed detections of respiratory pathogens, initially focusing on influenza viruses.

**Method:** Data were selected from a larger population-based data linkage program of work with a total population cohort of approximately 470,000 Western Australian births from 1996 to 2012 to investigate the pathogen-specific burden of respiratory infections.

We selected all PathWest records of influenza-positive detections from respiratory specimens and WANIDD notifications of influenza in 2001-2012. Records with a specimen collection date within 48 hours of each other were considered to relate to the same episode of infection and were linked together.

We then calculated the number of records identified in PathWest only, WANIDD only or in both datasets and also compared the incidence rates for influenza, using person-time-at-risk as the denominator.

**Results:** We identified 5208 influenza-positive records from PathWest and 5169 influenza notifications from WANIDD. After linking both datasets, there were a total of 5991 influenza cases. Of these, 822 (13.7%) were identified in PathWest but not in WANIDD and 783 cases (13.1%) were identified in WANIDD but not PathWest.

Using either PathWest data alone or WANIDD data alone yielded an incidence of 1.5 per 1000 child-years for influenza. Combining both datasets increased the incidence by 20% (95% CI=1.1, 1.2) compared to using either dataset alone, resulting in an incidence of 1.7 per 1000 child-years. This increase in incidence was consistent across all age groups and locations.

**Discussion:** WANIDD is a valuable resource for capturing notifications from private and public laboratories. Including additional data from PathWest resulted in hundreds of additional cases that would have been missed if we relied on WANIDD alone. This highlights the benefits of combining information from multiple datasets through data linkage to accurately estimate disease burden at the population level. Future work includes validating WANIDD’s capture of pertussis and invasive pneumococcal disease as well as identifying ways of improving reporting of notifiable diseases.
VALIDITY OF CANCER DATA IN UK PRIMARY CARE DATABASES: COMPARISON OF OBSERVED AND EXPECTED CANCER INCIDENCE RATES

Authors: M Ranopa¹, I Douglas¹, T van Staa², L Smeeth¹, K Bhaskaran¹
¹Department of Non-Communicable Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, United Kingdom, ²Institute of Population Health, University of Manchester, Manchester, United Kingdom

Abstract 1451

Background: UK primary care data are frequently used in research studies with cancer outcomes, but the validity of cancer diagnosis information in such databases is unclear. Recent studies have shown a disparity in cancer incidence rates (IRs) from UK primary care databases compared to national rates based on collated cancer registrations. Whether this disparity is still observed when linking cancer registry data to primary care data is unknown.

Objectives: For the four most common cancers among men and women in the UK, we aimed to compare incidence rates calculated using Clinical Practice Research Datalink (CPRD) primary care data, with published national incidence rates based on cancer registrations. Secondly we aimed to assess the impact of incorporating cancer registry data linked to the CPRD when estimating incidence rates.

Methods: A random sample of 2 million patients was selected from the CPRD. Cases of breast, colorectal, lung, and prostate cancer were identified over the period 2000-2010 using Read codes. We then identified corresponding cancer diagnoses in the NCDR using ICD-10 codes. Two sets of age- and sex-specific IRs were estimated: (i) CPRD rates; (ii) updated CPRD rates incorporating linkage to the National Cancer Data Repository (NCDR). IRs were compared with national estimates published by the UK Office for National Statistics (ONS).

Results: Overall CPRD cancer IRs (per 100k Person-Years) were lower compared to ONS reported rates: Breast (CPRD, 140.2 vs ONS, 145.2); Colorectum (Male: CPRD, 52.9 vs ONS, 66.0; Female: CPRD, 41.2 vs ONS, 52.2); Lung (Male: CPRD, 52.6 vs ONS, 73.8; Female: CPRD, 39.7 vs ONS, 50.8); Prostate (CPRD, 108.1 vs ONS, 117.8). Larger disparities were observed among the elderly, and during earlier years of the study period. Compared to ONS rates, updated IRs incorporating linked NCDR data (including all cases from both data sources) were similar for colorectal and lung cancer, but higher for breast and prostate cancer: Breast (159.3); Colorectum (Male: 67.7; Female: 55.1); Lung (Male: 71.6; Female: 53.3); Prostate (135.6).

Conclusions: Consistent with previous studies, CPRD cancer IRs were lower compared to ONS IRs and the disparity varied by age and cancer type. This was no longer the case when linkage to the NCDR was incorporated, but for breast and prostate cancer, incidence rates were then higher than expected, implying that a proportion of these cancer cases in CPRD are either false-positive, or not registered nationally.
UK SECURE RESEARCH PLATFORM – SAIL-ING TO A NEW WORLD

Authors: Simon Thompson
Swansea University

Presenting Author: Simon Thompson

Corresponding author email: simon@chi.swan.ac.uk

SAIL Databank/FARR Wales has completely re-engineered and expanded its core infrastructure to create the UK Secure Reasearch Platform. This initiative enabled FARR Wales to provide a platform as a service (PaaS) to internal programmes of work as well as externally run national programmes. This work builds on the years of experience gained by successfully running the SAIL Databank and represents the next evolution of our facilities.

The UKSeRP will provide much better return on investment for research bodies funding research programmes as the speed of providing secure, robust infrastructure is greatly reduced as is the cost and risk associated with a project. For the research group the complexities of doing the research is reduced substantially as they can focus on the research questions and have a wide choice of diverse technologies developed, maintained, provisioned and securely hosted by specialist IT staff, removing the need for project level expertise in these areas.

The UKSeRP will allow a much wider set of governance models than that chosen by the SAIL Databank and forms part of an initiative to enable wider NHS use of anonymised linked data for planning and operational intelligence.

This presentation will also look at how this has been achieved and what impact extra professional accreditation has had on this development. Incorporated into UKSeRP is a NRDA (national data research appliance) which provided completely devolved account management, data managements, data documentation, etc allowing for the research programme wo self-manage their environment.

The UKSeRP is extensible for example the Dementia Platform UK (DPUK http://www.dementiasplatform.uk ) will be using an instance of UKSeRP to deliver on the informatics work streams which incorporate a challenging set of use cases.

The UKSeRP through the CLIMB project (http://www.climb.ac.uk) has been extended to provide very high CPU/memory virtual servers for research groups working in the field of microbial bioinformatics. FARR Wales as well as being a delivery partner in the project is investigating how these two world start to interact for mutual benefit.

UKSeRP/NRDA DATA QUALITY & DATA CATALOGUE

Authors: Simon Thompson
Swansea University

Presenting Author: Simon Thompson

Corresponding author email: simon@chi.swan.ac.uk

The SAIL Databank has been successfully operating for many year and has experienced a massive growth in the number of projects and datasets held and curated. The SAIL databank remains one of
the most technology advanced data linkage centre in the world, however the way we provide datasets to research projects has not changed much from the platforms inception.

As part of the work undertaken by the FARR Wales programme SAIL have been able to develop an electronic data catalogue system to provide much more understanding to end users and help reduce the complexities of working with a set of datasets.

As part of the NRDA development the system captures the high level description and knowledge of a dataset and this documentation stage is mandatory for loading a dataset using the NRDA into SAIL/UKseRP. The NRDA also as part of providing dataset management allows for the collection of variable level documentation as well as automatically computing descriptive metrics during the data loading process. The system also lets you define simplistic rules about variable validation which will be viable as a marker of quality.

The dataset documentation is then available through a web based data catalogue which is delivered through the NRDA product, embedded into SAIL/UKseRP. When a dataset is transferred from one NRDA to anther the dataset documentation is also transferred so that when the received data is loaded into the end system it is fully documented.

There have been many challenges implementing this into a brownfield site such as SAIL as it has many dataset and many versions of dataset, as well as hundreds of project specific cuts of datasets and project specific additional data files. As such SAIL/UKSeRP has an auto documentation feature which will respond to database changes for core dataset holdings as well as project specific areas, a key develop area has been that a cut of a dataset needs to inherit the documentation of the variables which have been provisioned but need to recomputed the descriptive metrics as the cut could be for a subset of the held dataset.

This initiative is a significant improvement to both to SAIL, NRDA and the planned future instances of UKSeRP. The development road map will see many more features, knowledge and insight being made available to researchers.

Abstract 1645

HIC RDMP: A PROCESS-DRIVEN RESEARCH DATA MANAGEMENT PLATFORM

Authors: E Jefferson, T Nind, G McAllister, W Bonney, D Scobbie

University of Dundee

Presenting Author: T Nind

Corresponding author email: e.r.jefferson@dundee.ac.uk

Background & Objectives: The Health Informatics Centre (HIC) at the University of Dundee provides a service to securely host pseudonymised, cohort linked extracts of clinical datasets to researchers so as to enable them to answer key research questions.

A major challenge for research using routinely collected clinical datasets is that they are often not ‘research-ready’ or fit-for-purpose. Researchers continue to struggle in acquiring and reusing large longitudinal datasets for meaningful translational research. The objective of the HIC Research Data Management Platform (RDMP) is to (a) accumulate, manage and control shared access to research data; (b) transform and maintain transformation state information about research data; (c) analyse and investigate data in related sets using open and bespoke tools; and (d) publish results to a secure safe haven.

Methods: This talk will describe the design and implementation of the HIC RDMP. The RDMP consists of a set of data structures and processes, sharing a core Data Catalogue, to manage electronic health records, genomic data, and imaging data throughout their lifecycle from identification and acquisition to safe disposal or archival and retention in secured safe havens. The architecture is centred on a core Data Catalogue with five internal processes: Data Load Process, Catalogue Management Process, Data Quality Process, Data Summary Process, and Data Extraction Process.
The tools and processes that encompass the HIC RDMP not only fulfill the research data management requirements of researchers, but also support seamless collaboration of data cleaning, data summarization and data quality assessment activities by different research groups.

**Results:** An 8 month internal evaluation was performed to assess the performance of the HIC RDMP and indicated that the implementation of the HIC RDMP has reduced the delivery time of data cohorts from six months down to two weeks and supported over 185 research projects.

**Conclusions:** The HIC RDMP (a) provides the research community with a system for data management and curation; (b) removes cost of data quality assessment and improvement from individual research projects and contain this in the management of the data source; (c) provides a system for continuous improvement in data quality in large scale administrative datasets; and (d) emphasizes expert-lead process rather than technology to address the issue of research data management.

**Abstract 1632**

**TOWARDS A NATIONAL RESEARCH DATASET OF ROUTINELY-COLLECTED CLINICAL IMAGING**

**Authors:** Gordon McAllister  
*University of Dundee*

**Presenting Author:** Gordon McAllister

**Corresponding author email:** g.i.mcallister@dundee.ac.uk

**Background:** Farr Scotland is creating a national research dataset from imaging data contained in the national clinical PACS (Picture Archive and Communication System). This contains all clinical radiological imaging captured across Scotland since 2006, and therein a wealth of untapped phenotypic data. It is believed this will be the largest resource of its kind in the world.

The project leverages the Health Informatics Centre (HIC) Research Data Management Platform (RDMP), extending it with the capability to manage large-scale image data.

**Aims**

1. Enable novel research by providing a linkable, research-ready radiological imaging dataset and image analysis tools in a Safe Haven environment.
2. Provide workflows and infrastructure for ‘big data’ analysis of imaging through co-location with an HPC resource.
3. Develop a general imaging data extension to the HIC RDMP to be used as the basis for managing other imaging datasets.

**Methods:** Creating a large-scale repository of unconsented clinical imaging data involves significant challenges beyond those encountered with text-based electronic records. There are technical challenges related to the sheer volume of data, but there are also significant challenges around anonymisation, governance and security. With 700TB of data, increasing at approximately 20TB/month, any solution requires not only adequate storage for the dataset and project extractions but network, infrastructure and management software with sufficient throughput to keep pace with the influx of new imaging. This talk will discuss these challenges and the solutions that have been devised and implemented.

There are further challenges in extracting sufficiently rich metadata for useful cohort identification, and the potential for creating new metadata through automatic analysis of the pixel data and DICOM tags. This talk will examine these and other researcher-oriented challenges including data extraction and analysis tool provision in the Safe Haven environment.

**Results:** Solutions for the historic and ongoing transfer processes have been developed, with 120TB of historic data currently saved. The logical software architecture for the system at the Advanced Computing Facility has been created and software components have been implemented, including a robust caching mechanism and a DICOM file-specific load module for the RDMP Data Load Engine.

**Conclusions:** This project will unlock a whole new class of data for researchers to use in large-scale
linkage studies, supplying both the data and tools with which to perform novel research that has not been possible before now, and could have wide-ranging benefits for public health both in Scotland and beyond.

Abstract 1634  GenomicsDb – BUILDING A CONSENSUS GENOME DATABASE AS PART OF THE HIC RESEARCH DATA MANAGEMENT PLATFORM

Authors: Phil Appleby
University of Dundee

Presenting Author: Phil Appleby

Corresponding author email: p.appleby@dundee.ac.uk

Background: The Genetics of Diabetes Audit and Research in Tayside (GoDARTS) genomic data set currently has consented DNA SNP data for approximately 9500 diabetics and 8500 controls from across Tayside, Scotland. Experiments were performed using several different SNP-array assay platforms which differ in coverage. The data has typically been processed on an ad-hoc basis by researchers and has been held in a collection of flat genotype files per assay platform often with multiple copies in modified forms in existence. This has made maintenance and results reproducibility challenging. Phenotype data from routinely collected ehealth records for the GoDARTS Cohort is held within the Health Informatics Centre (HIC) Research Data Management Platform (RDMP). There has, however, been no easy method to link phenotypes to genotypes and querying the data has been a manual exercise for researchers, often taking several days to find the required information.

Aims:
- To introduce standard work flows for processing Genomic data, with each action in the process logged and reproducible, removing the requirement for multiple copies.
- To create a secure, searchable data store for rapid data retrieval, linking the genotype and phenotype data within the RDMP platform.
- To combine results from multiple assay platforms.
- To produce a consensus cross-assay platform genome for each genotyped individual.

Methods: Standard work flow pipelines for called-variants were created using configurable shell, python, perl and R scripts. The pipelines were configured for SNP QC, Individual (sample) QC, pre-phasing, imputation and database load processing.

The GenomicsDb created consists of SNP and sample indexes held in a MongoDB NoSQL database with the bulk of the data being retained in compressed, indexed Variation Call Format (VCF) files.

Results: Data from pipelines and the web-based database search tool are being used by research groups with the time to query and retrieve data now a matter of minutes rather than days as before. Combining data from multiple assay platforms has enabled a “best guess” consensus genome view to be built for any individual.

Conclusions: The standardisation of work flows has enabled processing steps to be retraced for reproducibility. Securing the data in a shared database now protects a definitive version of the data from unauthorised and unaudited modification.

The Genomics data work flow and database as a component of the RDMP will allow the easier exploration and analysis of phenotype data in conjunction with genotype data and should facilitate novel pattern discovery in the data.
SESSION B1 – GEOGRAPHICAL STUDIES & METHODS.
CHAIR: DR RICHARD FRY, SWANSEA UNIVERSITY

Abstract 1379

EXPOSURE TO AMBIENT AIR POLLUTION AND FETAL GROWTH IN NORTH EAST SCOTLAND - A WHOLE POPULATION STUDY

Authors: T Clements¹, C Dibbens¹, S Bhattacharya², S Turner²
¹University of Edinburgh, ²University of Aberdeen

Presenting Author: T Clements

Corresponding author email: Tom.Clemens@ed.ac.uk

Background: Maternal exposure to environmental insults during pregnancy has been associated with a number of adverse pregnancy outcomes. Exposures to increased concentrations of air pollutants such as nitrogen dioxide (NO₂) and suspended particulate matter less than 10 microns in size (PM₁₀) have been linked to reduced birth weight and prematurity. Studies in the Netherlands, Spain, France and Australia have linked maternal exposures to NO₂ and PM₁₀ to reduced fetal head size as measured on antenatal ultrasound scan. Our hypothesis was that exposure to air pollutants will be associated with reduced fetal size and growth for the population of Grampian born between 1985 and 2011.

Methods: A pollution climate mapping model which incorporates information such as point and area emissions totals and meteorological patterns was used to estimate an annual averages for NO₂ and PM₁₀ in each 1x1 km grid square across Scotland. This was combined with time series data from a fixed site pollution monitor to estimate individual 2 month exposure windows for each mother based on their residential postcode. Maternal and fetal data held on the Aberdeen Maternity and Neonatal Databank were linked to ambient air exposures by postcode and date.

Results: First trimester fetal size was linked to maternal exposures in 24,582 pregnancies and second trimester size to maternal exposures in 54,772 pregnancies. There were negative relationships between maternal NO₂ exposure and first trimester fetal size (crown rump length) and second trimester fetal measurements including head circumference and femur length; typically measurements were reduced by 1mm when the individuals with the highest exposures quartile were compared to those in the lowest quartile. Similarly, higher maternal exposures to PM₁₀ were associated with shorter first and second trimester fetal measurements. Change in fetal measurement will be linked to maternal ambient air exposures.

Conclusions: In this whole-population linkage study, the largest ever undertaken, we have replicated findings made elsewhere. Our next step is to undertake novel analyses which test the hypothesis that increased maternal ambient air exposures will be associated with reduced fetal size between the first and second trimesters.

Abstract 1515

SOCIAL DETERMINANTS OF VACCINE UPTAKE IN THE UK: USE OF ELECTRONIC HEALTH RECORDS FOR ASCERTAINMENT OF INEQUITY PARAMETERS

Authors: Anu Jain, Albert Jan van Hoek, Sara L Thomas
(1) Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London (2) Immunisation, Hepatitis and Blood Safety Department, Public Health England, London

Presenting Author: Anu Jain

Corresponding author email: anu.jain@lshtm.ac.uk
**Background:** Health inequities are a fundamental issue for public health. One important aspect of inequity is disparity in the uptake of vaccines to prevent childhood or adult infections. However, the extent to which national vaccination policies benefit individuals from all backgrounds is often poorly understood because relevant data on the social determinants of vaccine uptake are lacking. The increasing availability of linked electronic health records (EHR) creates exciting new possibilities for exploring vaccine-related health inequities. There has to date been little research to assess how well routinely collected data capture relevant social factors. Therefore, the potential for using existing and new data linkages to better characterise these factors needs to be determined.

**Methods:** In this presentation, we will report the results of a detailed investigation of social factors that are potentially available in selected EHR data from England. A range of datasets are being examined, including those from general practice, hospital inpatient settings, and child health records. Existing algorithms and other methodologies devised by data providers and by EHR researchers to enable identification of relevant factors will also be summarised. The data that are available in each data source will be compared with structural and social determinants of health inequities highlighted in common conceptual frameworks (eco-social models), and important gaps in availability will be discussed. Suggestions will be made for possible additional linkages that would enable more extensive examination of social inequities using EHR, but also potential barriers to effecting these linkages.

**Conclusion:** This work will help in utilising routinely collected data to measure health inequities, for example to enable better targeting of specific population groups for vaccination and to monitor vaccine-preventable disease burden by socioeconomic position. The methods identified in this investigation can be extended to other health outcomes in addition to vaccine uptake. The findings will also inform discussions between EHR researchers and data providers about possible future linkages.

**Funding:** This research is funded by the National Institute for Health Research Health Protection Research Unit (NIHR HPRU) in Immunisation at the London School of Hygiene and Tropical Medicine in partnership with Public Health England (PHE).

---

**Abstract 1530 PILOT STUDY LINKING PRIMARY CARE TO CENSUS AND HEALTH DATA IN SCOTLAND: FEASIBILITY AND VALUE FOR UNDERSTANDING ETHNIC VARIATIONS IN HEALTH**

**Authors:**
Anne Douglas¹, Genevieve Cezard¹, Colin R Simpson¹, Colin Fischbacher², Markus FC Steiner³, Aziz Sheikh¹, Hester Ward², Raj Bhopal¹

¹University of Edinburgh, ²NHS National Services Scotland, ³University of Aberdeen

**Presenting Author:** Anne Douglas

**Corresponding author email:** anne.douglas@ed.ac.uk

**Background:** Ethnic health inequality is striking. Quantifying this inequality is important. The Scottish Health and Ethnicity Linkage Study (SHELS) linked NHS hospital discharges and mortality to the 2001 Scottish Census. However, crucial risk factors are not recorded in either dataset. This project explored the feasibility and value of linking, at individual level, primary care risk factor data to Census and health records, and tested their use in risk models examining ethnic variations in cardiovascular disease (CVD) and mortality.

**Methods:** Ten of 17 general practices in Glasgow and Edinburgh with significant ethnic minority populations consented to provide anonymised data that were then linked to Census records using encrypted identifiers. Further linkage to CVD hospitalisations and all-cause mortality datasets created a retrospective cohort study.

Linkage rates of primary care to Census records were calculated and reasons for non-linkage explored. Completeness and validity of primary care risk factors were examined. Risk ratios (RR) for hospitalisation or death were calculated using Poisson regression with robust variance, adjusting for age, socio-
economic status (SES, from the Census) and risk factors (from primary care).

Findings: The process of gaining consent and extracting a complete, valid dataset was challenging and time consuming. 52,975 (50.5%) individuals from the 108,608 primary care records extracted from general practices were linked to 2001 Census data. Main reasons for non-linkage were having a first primary care record after 2001 (n=12,558) or no entry in the encrypted identifier key table (n=29,174). Risk factor data completeness was similar across ethnic groups. To explore the additional value of primary care data our analysis focused on two exemplar risk factors: smoking and diabetes. 48,325 (91%) of 52,975 records had a valid smoking status and 2900 (5·5%) people had a primary care record of diabetes.

Adding smoking and diabetes into age and SES adjusted risk models gave plausible results. For example, Pakistani women had higher age-adjusted risk of a first CVD event, but lower risk of all-cause mortality compared to White Scottish women. Adjustment for smoking increased these risks, whereas adjustment for diabetes reduced the risk, reflecting the high diabetes and low smoking prevalence in this group.

Conclusion: We overcame ethical and technical obstacles to extract a valid sample of primary care data and link at individual level to Census and health records. We have shown ‘proof of concept’ and potential value of hitherto unexplored record linkage methods, which could be adopted at national level.

Abstract 1682 GEOGRAPHICAL DIFFERENCES IN THE UPTAKE OF BOWEL AND BREAST CANCER SCREENING IN SCOTLAND

Authors: Gillian Libby1, Wenbai Yang2, Alistair Y Geddes1, A Stewart Fotheringham3, Patsy J Whelehan1, Robert JC Steele1
1Dundee University, 2None, 3Arizona State University

Presenting Author: Gillian Libby
Corresponding author email: g.libby@dundee.ac.uk

Aim: To increase understanding of geographically varying determinants of uptake, primarily for bowel cancer screening but also for breast cancer screening in Scotland.

Background: Bowel and breast cancer are among most prevalent types of cancer in Scotland. Population-based screening programmes were implemented 2007-2009 and 1988-1991 respectively, for residents aged 50-74 years for bowel screening and women 50-70 years for breast screening. Uptake levels differ between programmes, and key performance indicators show declines in uptake from lowest to highest multiple area deprivation quintile within all Scottish NHS health boards. This has drawn attention to strategies for increasing uptake in areas of highest deprivation, though this is confounded by previous research elsewhere suggesting uptake at small area level is independent of area deprivation.

Methods: An exploratory, cross sectional analysis was conducted including all invitees for bowel and breast screening between 01/01/2009-31/12/2011 and 01/04/2008-31/03/2011 respectively. Individual level screening data were aggregated into crude uptake percentages for the 6,505 Scottish Data Zones (a core local-area geography and also the reporting units for the Scottish Index of Multiple Deprivation, SIMD). Area deprivation levels including both the overall SIMD score and component-level scores were investigated as potential determinants of screening uptake. Other possible determinants also at Data Zone level were obtained from the 2011 population census. Key methods of analysis included both conventional linear regression and geographically weighted regression (GWR). Modelling was done at three levels, including all Data Zones within i) all Scotland, ii) each Health Board, iii) each city.

Results: Uptake exhibited considerable variation across Data Zones (7.7-80.3% for bowel screening and 0-100% for breast screening). Using conventional regression techniques, the associations between uptake and area deprivation scores were found to be inconsistent across both NHS Boards and cities.
However GWR modelling drew attention to localised areas where components of overall area deprivation and area-based ethnicity measures both were significantly negatively associated with bowel screening uptake. The GWR modelling also highlighted areas where bowel screening uptake was either lower or higher than would be predicted given values for area deprivation and census measures. Moreover, the discrepancy between observed and expected level of uptake was seen in Data Zones in both high and low SIMD quintiles.

**Discussion:** This analysis using linked data provides a more detailed view than previously of geographical variations in bowel and breast screening uptake. A significant negative association between bowel screening uptake and area deprivation was found only in localised areas.

**Abstract 1699**

**LINKING THE OUTPUTS FROM COMPLEX SPATIAL MODELS TO ROUTINE HEALTH RECORDS**

**Authors:** Richard Fry¹, S E Rodgers², J Morgan³, S Orford⁴, F D Dunstan⁵, D Fone⁶

¹Swansea University, ²Farr, Swansea University, ³School of Mathematics, Cardiff University, ⁴WISERD, Cardiff University, ⁵School of Medicine, Cardiff University

**Presenting Author:** Richard Fry

**Corresponding author email:** r.j.fry@swansea.ac.uk

**Introduction:** Linking spatially modelled environmental data to health data can provide valuable insights into environmental exposures that may affect population health. This paper explores some of the complex GIS methods developed at the Farr Institute, Swansea University and in particular uses the NIHR Funded CHALICE study as an example of linking complex spatial models to routinely collected health data held in the SAIL databank.

**Method:** The SAIL databank is uniquely placed in the UK to link spatial models of exposure at household level via a Residential Anonymised Linking Field (RALF). The CHALICE study developed complex measures of density at the household level, which were then aggregated to LSOA level for analysis. The study measured the spatio-temporal variation in alcohol outlets using a Geographic Information System (GIS) to model the local alcogenic environment over time. We calculated densities at household level, to more accurately define the local environment, creating distance-weighted walking neighbourhoods by defining 800m network based buffers with exposures to alcohol outlets weighted using a Butterworth filter.

**Results:** A localised spatial model of exposures can influence our understanding of health outcomes and provide a richer understanding of exposures when compared to areal-based spatial models. In the CHALICE project we used household level density scores, aggregated to LSOA level, for 24 time periods (quarterly between 2006 q1 and 2011 q4) for 1.4 million homes within Wales. Modelled data were linked to a variety of individual-level and aggregate health data (e.g. Hospital admissions data - within the SAIL databank, Welsh Health Survey) and crime data sourced from the police forces in Wales. The linked data showed that changes in the alcogenic environment had effects on health conditions and violent crime rates during the study period.

**Conclusion:** The CHALICE project demonstrates the power of linking localised spatial models of exposures to health data. The benefits of this approach included; linking health data at the household level, aggregating the density measures to any desired spatial unit (which is necessary where some data sources have been aggregated to maintain anonymity); and reducing the impact of the ecological fallacy. The methods developed here are being employed in other research projects where a local environment exposure may play an important role in population health studies (e.g. Alcogenic environments around schools, fast food density around routes to schools) and have helped us understand some of the issues surrounding the use of complex spatial models in an anonymised databank.
SESSION B5 – NATURALISTIC COHORT STUDIES.
CHAIR: PROF ADAM TIMMIS, BARTS HEART CENTRE

Abstract 1641
AMBULATORY CARE DATA IN A STABLE CHEST PAIN COHORT (N=8762):
PROGNOSTIC MODELING OF 10 YEAR CORONARY MORTALITY

Authors:
Adam Timmis1, Neha Sekhri2, Tim Clayton3, Gene Feder4, Harry Hemingway5, Pablo Perel3
1NIHR Cardiovascular Biomedical Research Unit, BartsHealth, London, 2BartsHealth, London, 3London School of Hygiene and Tropical Medicine, 4Centre for Academic Primary Care, School of Social and Community Medicine, University of Bristol, UK, 5Farr Institute of Health Informatics Research at London, London, UK

Presenting Author: Adam Timmis
Corresponding author email: a.d.timmis@qmul.ac.uk

Objective: Ambulatory care data in the hospital setting are rarely available for clinical research. We recorded clinical data in patients referred for outpatient assessment of chest pain and then followed them up for 10 years. The study objectives were to develop a new prognostic model to inform clinical management and to determine relations between mortality risk and the probability of coronary disease utilising modified Diamond-Forrester and Duke score estimates.

Methods and Results: In a multicenter study, clinical data were recorded electronically in 8762 ambulatory patients attending for cardiac outpatient assessment of previously undiagnosed chest pain. The patients were followed-up in the linked national death registry for a median 10 years (minimum 9 years) during which 233 coronary deaths were recorded. A prognostic model was developed with internal validation showing strong association with 10 year mortality for age, sex and chest pain typicality as well as ethnicity, pulse rate, smoking status, diabetes and ECG findings. Model discrimination was good (Harrel’s c=0.83), patients in the highest risk quartile accounting for 177 coronary deaths during follow-up compared with a total of 56 deaths in the lower risk quartiles. Observed 10 year coronary mortality increased progressively as the predicted risk of coronary disease by the Duke score increased, ranging from 0.2% to 10.6% in patients with predicted risks of coronary disease <10% and >90%, respectively. Only in groups with a ≥30% predicted risk of coronary disease did the observed 10 year coronary mortality exceed 1%.

Conclusion: For the first time in ambulatory patients with suspected angina, a prognostic model is presented based on simple clinical factors available at the initial cardiac assessment. The model discriminated powerfully between patients at high risk and at lower risk of coronary death during the follow-up period. Its potential clinical utility was reflected in the prognostic value it added to probability estimates of coronary disease by the Duke score.
Abstract 1426

LONG TERM HEALTH CARE USE AND COSTS IN PATIENTS WITH STABLE CORONARY ARTERY DISEASE: A POPULATION BASED COHORT USING LINKED ELECTRONIC HEALTH RECORDS (CALIBER)

Authors: S Walker¹, M Asaria¹, S Palmer¹, A Manca¹, H Hemingway², M Sculpher¹
¹Centre for Health Economics, University of York, ²Institute of Health Informatics, UCL

Presenting Author: M Asaria

Corresponding author email: simon.walker@york.ac.uk

Objectives: To examine long term health care utilisation and costs of patients with stable coronary artery disease (SCAD).

Methods: A cohort study of 94,966 patients with SCAD (47.4% of patients had stable angina, 13.5% unstable angina, 6.6% STEMI, 9.8% NSTEMI and 22.6% other coronary heart disease (CHD) as their qualifying diagnosis) in England between 1st January 2001 and 31st March 2010, identified in four prospectively collected, linked electronic health record (EHR) sources (primary care, secondary care, disease registry and death registry) from which health care resource utilisation was quantified. Resource use and costs in the first year and the first year following a non-fatal event (myocardial infarction, ischaemic or haemorrhagic stroke and Cost predictors in the first year with SCAD were estimated. Mean 5-year and lifetime healthcare costs were estimated by cardiovascular disease (CVD) risk level.

Results: 20.5% of patients were hospitalised for CHD in the first year rising to 66% in the year following a non-fatal event. Mean total health care costs in the first year were £3,133 per patient (56.8% CHD-related), increasing to £10,377 per patient in the year following a non-fatal event (66.2% CHD-related). First year predictors of cost included gender (mean cost £654 lower in women); SCAD phenotype (NSTEMI cost £745 more than stable angina); and co-morbidities (heart failure cost an additional £802 per patient). In patients in the lowest risk decile (5 year CVD risk of 3.4%), 5-year costs were £9,335 (44.7% CHD-related) and lifetime health care costs £116,888 (40.1% CHD-related). In patients in the highest risk decile (5 year CVD risk of 43.7%), 5-year costs were £23,391 (53.0% CHD-related) and lifetime health care costs £43,020 (51.9% CHD-related).

Conclusion: Patients with SCAD incur substantial health care utilisation and costs, which varies and may be predicted by 5-year CVD risk profiles. High risk patients have markedly higher costs over the initial 5 years but lower lifetime costs than low risk patients as a result of shorter life expectancy. Improved cardiovascular survivorship and an ageing UK population will require stratified care in anticipation of the burgeoning demand. The methods used here could be readily applied to other chronic diseases to better inform clinical decision making.

Abstract 1421

HEART RATE AND CARDIOVASCULAR DISEASE PROGNOSIS OF PATIENTS WITH STABLE CORONARY ARTERY DISEASE (SCAD): A CALIBER STUDY

Authors: O Archangelidi¹, M Pujades-Rodriguez¹, A Timmis², S Denaxas¹, H Hemingway¹
¹Farr Institute of Health Informatics Research, ²Barts and The London, NHS Trust

Presenting Author: O Archangelidi

Corresponding author email: o.archangelidi@ucl.ac.uk

Background: Studies assessing the association of resting heart rate with coronary heart disease and heart failure in populations with pre-existing stable coronary artery disease present conflicting findings. And whether heart rate is associated with cerebrovascular or peripheral cardiovascular diseases (CVDs) or cardiac rhythm disorders is not known.
Methods: Linked primary care, hospitalisations, myocardial infarction, and cause-specific mortality records (CALIBER dataset) of people aged ≥30 years with history of CVD in 1997-2010 were included. Associations between heart rate and ten fatal and non-fatal CVDs were assessed using stratified Cox proportional hazard models.

Results: Among 51,703 study patients experienced 21,918 events during 134,106 person-years of follow-up. Strong associations were found with heart failure (HR=2.96, 95% CI 2.59-3.39 for >90 vs. <60bpm), atrial fibrillation (HR=2.04, 95% CI 1.88-2.21) and sudden cardiac death (HR=1.30, 95% CI 1.10-1.55). These associations were present regardless of the type of pre-existing coronary event. Moderate associations were found with all-cause stroke which were diminished for those with pre-existing MI or UA. No associations were found with myocardial infarction or peripheral arterial disease. Associations did not differ by sex. The risk for heart failure, atrial fibrillation and cardiovascular mortality steeply increased above 70bpm, while the risk of sudden cardiac death was observed for values above 90bpm.

Conclusions: Resting heart rate in stable coronary disease population showed heterogenous associations with recurrent CVD events that varied with the type of pre-existing coronary artery disease. These findings support conducting trials for secondary prevention among patients in early stages of heart failure development.

Abstract 1484 ETHNICITY AND INCIDENCE OF TWELVE CARDIOVASCULAR DISEASES: AGE AND GENDER SPECIFIC ASSOCIATIONS IN 1 MILLION PEOPLE

Authors: Julie George1, Rohini Mathur2, Mar Pujades Rodriguez1, Eleni Rapsomaniki1, Anoop Dinesh Shah1, Spiros Denaxas3
1Farr Institute of Health Informatics Research (London), UCL, 2London School of Hygiene and Tropical Medicine, 3Farr Institute of Health Informatics Research (London)

Presenting Author: Julie George
Corresponding author email: j.george@ucl.ac.uk

Introduction: Little is known about ethnic differences in initial lifetime presentation of specific cardiovascular diseases (CVDs) across a range of acute and chronic disease presentations in the general population. Understanding ethnic disparities in the incidence of CVDs is an important step in addressing inequalities in health outcomes between ethnic groups.

Methods: Using linked data from English primary care (Clinical Practice Research Database, (CPRD)), secondary care (Hospital Episode Statistics (HES)), acute coronary syndrome clinical registry (Myocardial Ischaemia National Audit Project) and mortality records (Office for National Statistics) from the CALIBER research programme, a cohort of 1,068,318 patients aged ≥30 and free from clinical cardiovascular disease was constructed. The initial lifetime presentation of 12 CVDs including cardiac, cerebrovascular, abdominal and peripheral arterial disease were identified across all data sources, using variable definitions from CALIBER (https://www.caliberresearch.org/). Ethnic group was identified from records in CPRD and HES, using a validated algorithm. CPRD provided information on covariates. The associations between ethnic group and initial lifetime CVD presentations was modelled using disease-specific Cox proportional hazard regression, using both complete case and multiply-imputed data.

Preliminary Results: 90.9% of patients were White, 3.6% South Asian and 2.9% Black, and contributed 5.9 years median follow-up time and 95,224 CVD events to the analysis. The mean age at initial lifetime CVD presentation for White patients was 10 years older (71.2 (95% CI 71.1-71.2)) than for South Asian (61.5 (60.9-62.0)) or Black patients (62.1 (61.4-62.9)). In age-sex adjusted models, South Asian patients were more likely to present with coronary heart diseases (CHDs) compared to White patients (Hazard ratio (HR) for stable angina: 1.68 (1.52-1.85); unstable angina: 1.83 (1.57-2.14); myocardial infarction: 1.67 (1.49-1.87)); Black patients were less likely than White to present with unstable angina (HR: 0.76 (0.59-0.97)) or myocardial infarction (HR: 0.50 (0.40-0.62)). Both South Asian (HR 0.70 (0.57-0.86)) and
Black (HR 0.63 [0.50-0.80]) were less likely than White patients to present with peripheral arterial disease, but more likely to present with ischaemic stroke (South Asian HR 1.30 [1.03-1.63]; Black 1.24 [0.97-1.59]).

Discussion: Ethnic differences in initial lifetime presentation of CVDs were observed in preliminary analyses, particularly in the association of ethnic group with CHD. Additionally, South Asian and Black patients experienced their first CVD approximately 10 years earlier than White patients. Results of additional modelling will be presented, along with reflections on the use of ethnicity from multiple data sources and implications of our findings for primary disease prevention.

Abstract 1698

USING WELSH GP RECORDS TO ASSESS CARDIOVASCULAR RISK

Authors: A Lacey1, N Gollop2, R Lyons1, J Halcox1
1College of Medicine, Swansea University, 2Cardiff University

Presenting Author: A Lacey

Corresponding author email: a.s.lacey@swansea.ac.uk

Purpose: To examine the extent of risk factor (RF) assessment and documentation in primary care (PC) in the 5 years preceding de novo presentation with premature cardiovascular disease (CVD) in Wales, to assess utility of conventional CVD risk scores to identify high-risk individuals and consider the potential impact of new NICE-guidance on CVD risk assessment practice.

Method: We accessed records for 1,255,578 residents <65 years of age stored on the Secure Anonymised Information Linkage (SAIL) database. De novo ICD-10 secondary care diagnoses of CAD were identified for men aged <55 years and women <65 years between 01.01.2006- 01.01.2013. Linked primary care records (PCR) datasets were accessed to determine documentation of conventional RFs and CVD risk scores (Framingham [FRS] and QRISK2) during the 5-years prior to index event. We also calculated FRS/QRISK2 scores, post-hoc where possible, and selected the corresponding Framingham laboratory-based (FRS) and non-laboratory-based (FRSnl) scores that could be calculated within the same timeframe for comparison.

Results: De novo premature CVD was diagnosed in 8157 individuals (3544 men and 4633 women). We found that less than 6% of the cohort had a QRISK2 or FRS risk score documented in PCR. We were able to retrospectively calculate QRISK2 and FRS scores in 26.2% of men and 34.9% of women. For those where we found the maximum QRISK2 scores per person scored 45%/39%/16% of men and 48%/38%/14% of women as low risk (< 10%), intermediate (10-20%) and high risk (>20%) respectively. FRS scored 18%/42%/40% of men and 35%/38%/26% women as low/intermediate/high risk respectively. FRSnl scored 13%/42%/45% of men as low/intermediate/high risk and 23%/40%/37% women as low risk/intermediate/high risk.

Conclusions: Significant gaps exist in the extent of RF assessment and documentation for primary prevention of premature CVD in Wales. Of those developing premature CVD, a formal risk score was documented in fewer than 1 in 10 records within 5y prior to initial diagnosis. Notably, just under half of these young individuals who go on to develop coronary disease within five years would have been considered at low risk by QRISK2, whereas only approximately 1 in 8 of these men and 1 in 4 women would have been considered low risk by the simpler FRSnl algorithm. NICE guidelines suggest using a QRISK2 score to assess patients cardiovascular risk, however in a younger cohort we suggest using a simple FRS non-lab score.
PATIENT FACTORS ASSOCIATED WITH SSRI DOSE FOR DEPRESSION TREATMENT IN GENERAL PRACTICE: A PRIMARY CARE CROSS SECTIONAL STUDY.

Authors: Chris F Johnson1, Nadine J Dougall2, Brian Williams2, Stephen A MacGillivray3, Alasdair I Buchanan4, Richard D Hassett1

1Pharmacy and Prescribing Support Unit, NHS Greater Glasgow and Clyde, 2Nursing Midwifery and Allied Health Professionals Research, University of Stirling, 3Social Dimensions of Health Institute, University of Dundee, 4NHS Greater Glasgow and Clyde, Glasgow City Community Health Partnership

Presenting Author: Nadine J Dougall

Corresponding author email: c.johnson2@nhs.net

Background: Antidepressant prescribing continues to rise. Increased long-term prescribing and higher doses are contributing to current growth; however, patient factors associated with the use of higher doses remain unknown. This study's aim was to investigate patient factors associated with selective serotonin re-uptake inhibitor (SSRI) prescribed daily dose for depression treatment in general practice.

Methods: A stratified sample of low to high prescribing practices were selected. Routine individual patient-level data were extracted one practice at a time: September 2009 to January 2011. Patients included were ≥18 years, and prescribed an SSRI for depression. Logistic regression analysis was undertaken to assess individual predictor variables on SSRI daily dose by standard therapeutic dose versus higher dose, as SSRIs demonstrate flat dose response curves for depression treatment. Predictor variables included: age, gender, deprivation, co-morbidity, smoking status, being prescribed the same SSRI for ≥2 years, and patients' general practice. For a subgroup of patients a second sub-group analysis included long-term benzodiazepine and/or z-hypnotic (B&Z) as a predictor variable.

Results: Inter-practice SSRI prescribing varied significantly; practice point prevalence ranged from 2.5% (94/3697) to 11.9% (359/3007) of the practice population ≥18 years old; median 7.3% (250/3421) (χ2 = 2277.2, df = 10, p < 0.001). Overall point prevalence was 6.3% (3518/52575), with 5.8% (3066/52575) prescribed SSRIs for depression of whom 84.7% (2596/3066) had data for regression analysis. Higher SSRI doses were significantly associated with, in descending order of magnitude, individual practice attended, being prescribed the same SSRI for ≥2 years (Odds Ratio (OR) 1.80, 95% CI 1.49 to 2.17, p < 0.001) and living in a more deprived area (OR 1.55, 95% CI 1.11 to 2.16, p = 0.009). Higher SSRI doses in the B&Z subgroup were significantly associated with individual practice attended, being prescribed a long-term B&Z (OR 2.05 95% CI 1.47 to 2.86, p < 0.001) and being prescribed the same SSRI for ≥2 years (OR 1.94, 95% CI 1.53 to 2.47, p < 0.001).

Conclusion: Higher SSRI doses for depression were associated with practice attended and being prescribed the same antidepressant for ≥2 years. As long-term antidepressant use increases, the use of higher doses may further contribute to prescribing growth.
Abstract 1781

CONSENTING ADULTS? WHY DO SOME YOUNG WOMEN REFUSE CONSENT TO DATA LINKAGE?

Authors: Anna Graves¹, Jenny Powers¹, Caroline Jackson², Janni Leung³, Deirdre McLaughlin³

¹Research Centre for Gender, Health and Ageing, University of Newcastle, ²Usher Institute of Population Health Sciences and Informatics, University of Edinburgh, ³School of Public Health, The University of Queensland, Herston

Presenting Author: Caroline Jackson

Corresponding author email: anna.graves@newcastle.edu.au

Objectives: The Australian Longitudinal Study of Women’s Health (ALSWH) has been investigating the health of more than 40,000 women since 1996. In 2012, recruitment of a new cohort of young women commenced. This presentation focuses on obtaining consent to data linkage and differences among those who did and did not consent.

Approach: Since 1996, the ALSWH has provided information drawn from three birth cohorts of women (1973-78, 1946-51 and 1921-26) to develop both policy and practice in Australia. In particular, data from the ALSWH were used by policymakers when developing the Australian Women’s Health Policy in 2010. To ensure that young women continue to be represented in future policy and practice, it was decided to recruit a new cohort of women born between 1989 and 1995. Various recruitment strategies were considered, resulting in the development and implementation of an online survey which was widely advertised through social media.

Results: Women born 1989 to 1995 were asked to complete the online survey which contained questions relating to their health and well-being. Initially, the women were required to provide consent to data linkage at the beginning of the survey, however no data were available on these non-consenters. The consent wording was then moved to the end of the survey. After the change, 17,700 women commenced the online survey; 6,400 women (~36%) did not consent to the linking of their data to health-related administrative datasets. Comparisons were done between consenters and non-consenters on a number of variables previously associated with consent or non-consent in large epidemiological surveys. Age, state of residence, income and a number of health conditions differed between consenters and non-consenters.

Conclusions: Linkage to administrative datasets is increasingly recognized as being able to deliver substantial benefits with low risks and manageable costs. Preventing attrition from non-consenters should be a focus of research teams to maximise the information derived from epidemiological surveys.
PUBLIC VIEWS ON IMPLEMENTING AN ELECTRONIC INTERFACE ENABLING ‘DYNAMIC CONSENT’ AND FEEDBACK RELATED TO THE USE OF PERSONAL HEALTH DATA FOR RESEARCH

Authors: Lamiece Hassan1, Caroline Sanders2, Karen Spencer2, Tjeerd Van Staa2, Hawys Williams2, Will G Dixon2

1Institute of Population Health, 2The University of Manchester

Presenting Author: Lamiece Hassan

Corresponding author email: lamiece.hassan@manchester.ac.uk

Advances in computing and data science are creating unprecedented opportunities to accumulate and share data about our individual health, biology and care. Health research data initiatives have the potential to create new knowledge and improve healthcare. Nonetheless, in the wake of concerns raised by the media and privacy groups, conventional approaches towards data governance may no longer be sufficient to respect and protect individual privacy.

Part of the solution may be to consider an innovative model of electronic 'dynamic consent', enabling greater personal control over the re-use of individual health data (Williams et al, 2015). Importantly, patients are not restricted to giving universal consent (or opt-out) for use of data; they can modify consent status and preferences over time. Furthermore, electronic systems provide opportunities for informing patients about the recipients of their data and the results of research to which their data have contributed.

We held a public workshop to consult on methods of implementing an electronic interface for enabling dynamic consent and feedback from research using health data for research purposes. Thirty members of the public were recruited via established public involvement groups and social media. Key concepts relating to dynamic consent were explained, followed by discussions in smaller groups led by researcher-facilitators. Two exemplar scenarios incorporating dynamic consent were introduced: one in a hospital outpatient setting considering permissions for sharing electronic patient records and the other in primary care addressing consent for pragmatic trials. Additionally, we presented a working prototype of the electronic interface for attendees to test themselves on a tablet. Facilitators probed views on proposed procedures for obtaining dynamic consent using the electronic interface, including information content and format, consent options and wider contextual factors (e.g. internet access). Flip chart notes and evaluation forms were synthesised following the workshop.

Overall, people were supportive of the concept of dynamic consent and particularly welcomed being able to receive feedback about research, making them feel valued. Given that people vary in the amount of detail they want and which components of the system they would use, attendees favoured layered information. Feedback indicated some users wanted additional support for using electronic systems and/or alternative non-electronic options to enable dynamic consent. Privacy (e.g. private booths) and reassurance regarding data security were also important considerations. Findings are being used to inform the design of a future research project that will implement and test procedures for obtaining dynamic consent in Salford.
Abstract 1617

CO-DEVELOPMENT OF A SMARTPHONE APPLICATION TO ALLOW RESEARCH PARTICIPANTS TO VIEW THEIR RESEARCH DATA

Authors: Laura Lamming¹, Dan Mason¹, David Robison², Ann Barratt¹, John Wright¹, Mohammed, A Mohammed¹

¹Bradford Institute for Health Research, ²University of Bradford

Presenting Author: Laura Lamming

Corresponding author email: Laura.Lamming@bthft.nhs.uk

Background: Electronic access to personal data records has become routine for many online activities. Research participants may benefit from having access to their personal research data. As part of the development of a managed smartphone mCohort to enhance public engagement activities in the Born in Bradford (BiB) birth cohort, we will develop a new model for participant access to personal research data, and explore attitudes to having personal research data made available. This project will engage directly with participants, building and developing an application (app) collaboratively with volunteers from the BiB cohort.

Methods: A small group (n=4 to 8) of participants have been recruited. Focus groups are being conducted. The key requirement of the groups will be to allow participants to specify what they want from their data and allow development to progress towards a shared purpose for the use of personal research data. Participants will interact with and discuss the prototype, commenting on usability, content, structure, style and context. App development will be iterative, based on feedback from focus groups. Brief content analyses will be conducted.

Results: Initial results indicate an interest in process data (e.g. which studies have participants contributed to) and personalised health status (e.g. detection of health issues, comparison to ‘healthy’ norms). Focusing on personal adult and child data was preferred, although cohort-wide data was also of interest. Participants were positive about the ability to view research data, but may access it only as new information was provided. Internet access and confidence with IT may be barriers to use. Further findings will be reported.

Discussion: Initial focus groups revealed an interest in personal research data. However, there was a preference for comparative information to facilitate understanding of the health status of their child. This would require an unfeasible level of personalisation to accommodate individual differences and avoid undue concern.

Process data was of interest, with comments suggesting that longitudinal studies may struggle with attrition due to infrequent contact. This supports the need for novel methods, (e.g., the mCohort), to further promote and maintain contact within a large research study.

Conclusion: Initial findings suggest that development of an app to support public engagement within a birth cohort is promising. Providing data may be insufficient to maintain participant engagement, but the degree of personalisation suggested by current participants may not be feasible. Longitudinal studies should consider mobile platforms to maintain and facilitate participant engagement over the long term.
Abstract 1397

USING HEALTH INFORMATICS TO ENHANCE PUBLIC ENGAGEMENT IN DIABETES RESEARCH

Authors: John Kerr
Scottish Diabetes Research Network

Presenting Author: John Kerr

Corresponding author email: jkerr@dundee.ac.uk

Background: Enhancing public engagement in clinical research whilst actively encouraging participation remains a great challenge. To address this challenge, the Scottish Diabetes Research Network (SDRN) in collaboration with the SCI-DC Development Team based in the Clinical Technology Centre at Ninewells Hospital developed the SDRN Research Register.

Public Engagement and Recruitment: The Research Register is designed to give patients with diabetes an opportunity to record their interest in participating in diabetes research. The SDRN provide a number of opportunities for patient engagement and registration including:

- approaching patients at regular diabetes clinics and diabetes education sessions;
- presentations at patient conferences and meetings;
- a diabetes news blog on the SDRN website including an option to join the research register;
- arranging for General Practice to invite their patients to register by letter and
- general publicity via SDRN Facebook and twitter pages.

Engagement Statistics

- 70% positive uptake by patients on personal approach.
- 50% positive uptake by letter invitation.
- 10,300 patients on the Research Register.

Linking with Health Informatics

The significance and utility of the Research Register as a tool for recruitment to clinical research lies solely in its integration with SCI-Diabetes; a fully integrated shared electronic patient record used to support the treatment of NHS Scotland patients with Diabetes.

Once a patient has chosen to join the research register, their registration status is recorded in SCI-Diabetes.

The SDRN is then able to use a custom-built search tool in SCI-Diabetes for the application of eligibility criteria against registered patient data.

The results of each Research Register search provides a list of registered patients who match the study criteria entered. These patients now represent a sample of potentially eligible participants for the research study in question. Each patient can then be contacted by the study team before being provided with information and given the opportunity to take part.

Image

Enhancing Public Engagement through Specificity: By harnessing the power of health informatics through SCI-Diabetes, the Research Register can guarantee that patients will only ever be contacted about research, which they may be eligible for. This single advantage not only provides patients with a greater degree of confidence in granting their permission to be contacted but also enhances interest and awareness in suitable research regardless of their choice whether to participate.
mHEALTH APPLICATIONS FOR DIABETES – USER PREFERENCE AND IMPLICATIONS FOR APP DEVELOPMENT

Authors: N. T. Conway¹, I Campbell², P Forbes², S. G. Cunningham¹, D. J. Wake¹
¹University of Dundee, NHS Tayside, ²University of Dundee

Presenting Author: N. T. Conway

Corresponding author email: n.z.conway@dundee.ac.uk

Background: Increasing diabetes prevalence is driving a demand for more sustainable yet person-centred service. As the worldwide smartphone market continues to grow, the number of diabetes self-care mHealth applications also grows exponentially. mHealth can improve clinical outcomes, but current usage patterns, effectiveness and valued features are unclear. This study sought to assess levels of engagement with mHealth technologies within a subset of the Scottish diabetes population; to identify specific demographic sub-groups of interest; and draw comparisons between desirable and currently available features of diabetes mHealth applications.

Methods: A snapshot analysis of the diabetes mHealth app marketplace was undertaken in July 2014. Available features were used to construct a questionnaire. A random sample of 400 patients (stratified by diabetes type and age) was obtained from the Scottish Diabetes Research Network (n=200) and users of patient health record (MyDiabetesMyWay, n=200). Demographic variables (age group, gender and diabetes type) were cross-tabulated with preference for mHealth technologies and loglinear analysis was used to identify significant interactions. Desirable features of a diabetes mHealth app were compared with currently available diabetes apps.

Results: Available app features include: data storage/graphical presentation; integration with other apps; exercise tracking; health/diet tracking; reminders/alarms; and education. 59% (234/400) people responded to the questionnaire; 62% (144/233) owned a smartphone. Most smartphone users expressed a preference towards mHealth apps (101/142 (71%)) (especially younger age groups), although mobile phone app use for diabetes self-management was low (12/163 (7%)). Older women with T2D were significantly less likely to favour diabetes mHealth apps. Respondents favoured a wide variety of potential features, contrasting with current availability: patient education – favoured by 45% (98/220) users but available in 14% (10/74) apps; personal health record - favoured by 40% (89/220) users but unavailable on apps reviewed.

Discussion/conclusion: mHealth has the potential to empower patients; improve outcomes; and provide service in a sustainable way. This study demonstrates that mHealth acceptance is high, but current engagement is low and functionality does not match potential user preferences. Engagement and functionality could perhaps be improved by including relevant stakeholders in future development, driven by clinical and user need.
SUPPORTING RESEARCHERS TO INVOLVE CONSUMERS AND COMMUNITY MEMBERS (PATIENTS AND PUBLIC) IN THEIR RESEARCH

Authors: Hayley Haines, A.E McKenzie
University of Western Australia School of Population Health and Telethon Kids Institute

Presenting Author: Hayley Haines

Corresponding author email: hayley.haines@uwa.edu.au

Background: The University of Western Australia School of Population Health (the School) and Telethon Kids Institute (the Institute) established a Consumer and Community Participation Program (the Program) in 1998 in response to community concerns about linked data research. Both the School and the Institute undertake extensive research projects and programs using linked data research. The multi-faceted Program supports researchers undertaking linked data research to work with consumers and community members to make decisions about research priorities, policy and practice. In order to support the active involvement consumers and the community in research, the Participation Program has developed a range of initiatives to increase the capacity of researchers.

Training: Training workshops have been developed for both researchers on implementing consumer and community participation and the community on basic research information.

Methods: The Program has worked with researchers to develop methods that are particularly relevant for involving consumers and community members in linked-data research. These tried and tested methods include:
- Community steering and reference groups
- Research buddies
- Community conversations
- Community expo’s

Resources: Researchers who attended training requested resources to provide ongoing guidance on the planning and implementation of consumer and community involvement after training. A Factsheet Series was written in response to this request. A more recent addition, Planning for Consumer and Community Participation in Health and Medical Research, focuses on the practical considerations of planning for involvement. These are both valuable resources to researchers using linked data.

Community links: Many researchers who work with whole of population data do not have direct access to the consumers and community members in their research. To facilitate access to the groups the Program established a database of consumers and community members with an interest in research that can be approached to provide a consumer and community voice in research.

Advisory service: For many researchers who have had limited experience of involving consumers and community members in their research, access to a one-to-one advisory services has proven beneficial.

The presentation will showcase a range of supports and enablers used to embed the Consumer and Community Participation Program at the School and Institute.
PARALLEL SESSION C, THURSDAY 27TH AUGUST 09.00-11.05

SESSION C1 – DATA LINKAGE METHODS.
CHAIR: DR ROB ALDRIDGE, UNIVERSITY COLLEGE LONDON

Abstract 1666
IDENTIFYING THE EFFECTS OF K-ANONYMISATION ON EPIDEMIOLOGICAL DATA: AN EMPIRICAL EVALUATION USING CASE STUDIES FROM ALSPAC

Authors: A Davies¹, A Boyd¹, K Northstone², R Cornish¹, J Macleod³
¹ALSPAC, University of Bristol, ²NIHR CLAHRC West, University of Bristol

Presenting Author: A Davies
Corresponding author email: a.davies@bristol.ac.uk

Background: Data protection legislation permits the use of anonymised, rather than individually identifiable, routine records, without explicit consent. This is advantageous to researchers as seeking consent is resource intensive and can lead to reduced or biased samples. However, ensuring anonymity is challenging; in part due to the inherent risk of disclosure found in datasets containing unique combinations of individual level data. In 2013 the Health and Social Care Information Centre (HSCIC) implemented the Anonymisation Standard for the release and publishing of health and social care data. Drawing on recommendations from the Information Commissioner’s Office the standard recommends an anonymisation technique known as K-anonymisation; a de-identification method which controls for levels of uniqueness by ensuring at least k individuals share the same set of data values. While K-anonymisation methodologies feature widely in the literature, there is little evidence describing the effects the process has on the underlying utility of the data.

Methods: We sought to empirically examine the effects of K-anonymisation through replicating a number of existing research projects from the Avon Longitudinal Study of Parents and Children (ALSPAC). One case study investigated the association between breastfeeding and ‘GCSE’ educational attainment at age 16. We recorded the differences in statistical outputs using identifiable data in comparison with: i. data where only pseudo-identifiers were K-anonymised (weak K-anonymisation), and, ii. data where all values except the outcome variable were K-anonymised (strong K-anonymisation).

Results: For case study 1 we used the original identifiable dataset where logistic regression suggests strong evidence that cases breastfed (reference category never breastfed compared with cases whom were breastfed for <6months and cases whom were breastfed 6+months) were more likely to obtain 5 or more GCSE’s at grade C and above [<6months OR=1.30; 95% CI 1.13, 1.50; P<0.001: 6months+ OR=1.72; 95% CI 1.46, 2.05; P<0.001]. When performing preliminary analysis similar results were shown for the weak K-anonymised data, yet discrepancies became apparent when analysing the strong K-anonymised data, mainly the wider confidence intervals due to a smaller sample size [<6months OR=1.32; 95% CI 1.05, 1.65; P=0.016: 6months+ OR=2.05; 95% CI 1.60, 2.62; P<0.001].

Conclusions: Although discrepancies were found in the strong K-anonymisation the overall conclusions remained the same. Preliminary results suggest that weak K-anonymisation doesn’t appear to impact the data yet there is potential that both strong and weak K-anonymisation can impact the data integrity. Final results, limitations and conclusions of all case studies will be presented.
Abstract 1587

**USING A NOVEL BAYESIAN APPROACH TO PROBABILISTICALLY LINK RECORDS IN THE ECFSPR**

**Authors:** Peter D Hurley¹, Seb Oliver¹, Anil Mehta²

¹Department of Physics and Astronomy, University of Sussex, ²Division of Medical Sciences, University of Dundee

**Presenting Author:** Peter D Hurley

**Corresponding author email:** p.d.hurley@sussex.ac.uk

**Background:** Patient registries are essential for outcome analysis, particularly in rare diseases. Such data are mostly collected as annual snapshots. Records from individual patients must then be amalgamated using a unique identifier to understand the longitudinal progress of the patients across years. Such identifiers are often not available when data are collated across countries and hence a robust method of record linkage is required. Here, we apply astronomical techniques to solve this pressing problem, using an international CF registry as a paradigm.

**Methods:** We adopted a hierarchical Bayesian partition model framework, developed in astronomy, to probabilistically link patient records in the European Cystic Fibrosis Society Patient Registry (ECFSPR) using gender, genotype, diagnostic age, body mass index and height as factors. We applied our method to CF patient data held in Denmark for which pseudonimised identifiers were provided but personalised identifiers withheld.

**Findings:** Our algorithm correctly linked the records belonging to the same patients, (with TPR = 99.5\% FPR = 0.5\% for all probability thresholds and a ROC score of 0.97). We serendipitously discovered data errors in the pseudonimised identifiers, some of which were unknown to the registry administrators in the ECFS. Our algorithm is capable of providing an accurate probability of whether a pair of records belong to the same patient or whether preexisting pseudonimised identifiers are actually correct.

**Interpretation:** Our approach could reassemble legacy data sets where there is no likelihood of linkage information becoming available. We demonstrate that unique identifiers, although essential for clinical purposes within a hospital setting, might not always be accurate across years and provide an automated approach to pseudonimise data enabling reconstruction of large data sets across time.

Abstract 1601

**A OPEN PIPELINE FOR MASKING PATIENT IDENTIFIERS IN ELECTRONIC HEALTH RECORDS**

**Authors:** Richard Jackson, I Kartoglu, C Johnston, A Folarin, R Stewart, R Dobson

Kings' College London

**Presenting Author:** Richard Jackson

**Corresponding author email:** richard.r.jackson@kcl.ac.uk

Informatics innovation around the use of Electronic Health Records (EHRs) for research is a rapidly growing discipline, driven by the potential to reduce the costs of patient care and increase research output.

One of the key challenges is enabling researchers access to rich datasets of clinical records, without compromising the information governance policies of care organisations and ethical guidelines protecting patient confidentiality. Many existing research clinical datasets are unable to provide access to free text, as the potential risk of accidentally or maliciously identifying patients from their notes is considered too great. In areas such as Mental Health, this is highly problematic for EHR researchers, as the most important clinical information is often contained within the free text. Regular expression based document masking algorithms are known to achieve high performance in hiding strong patient identifiers, however, there is currently a lack of open source tools to process binary files.
and text into a common format, and deploy anonymisation techniques to make clinical text suitable for research use.

As part of other information governance controls, we have created a pipeline as a means to mask strong patient identifiers in clinical text, while retaining as much structure as possible of the original document. Cognition is an open source database processing pipeline that converts common electronic document formats, PDFs and images into a XHTML representation, and hides strong patient identifiers with a customisable regular expression based algorithm.

Cognition is designed to scale both vertically and horizontally to support easy deployment and for processing very large datasets. In conjunction with other Information Governance controls, we hope that this tool will assist in unlocking the wealth of data hidden in clinical documents.

https://github.com/KHP-Informatics/Cognition-DNC/blob/master/README.md

Abstract 1703 ACCURACY OF PROBABILISTIC LINKAGE USING THE ENHANCED MATCHING SYSTEM FOR PUBLIC HEALTH AND EPIDEMIOLOGICAL STUDIES.

Authors: Rob Aldridge1, Kunju Shaji2, Ibrahim Abubakar3, Andrew Hayward1
1Farr Institute and University College London, 2The Royal Marsden Hospital, 3University College London

Presenting Author: Rob Aldridge

Corresponding author email: r.aldridge@ucl.ac.uk

Background: The Enhanced Matching System (EMS) is a probabilistic record linkage program developed by the tuberculosis section at Public Health England to match data for individuals across two datasets. This paper outlines how EMS works and investigate its accuracy for linkage across public health datasets.

Methods: EMS is a configurable Microsoft SQL Server database program. To examine the accuracy of EMS, two public health databases were matched deterministically using National Health Service numbers as a gold standard unique identifier. Probabilistic linkage was then performed on the same two datasets without inclusion of NHS number. Sensitivity analyses were carried out to examine the effect of varying matching process parameters.

Results: Deterministic matching using NHS number between two datasets (containing 5931 and 1759 records) identified 1071 matched pairs. EMS probabilistic linkage identified 1068 matched pairs. The sensitivity of probabilistic matching was calculated as 99.5% (95%CI: 98.9, 99.8), specificity 100.0% (95%CI: 99.9, 100.0), positive predictive value 99.8% (95%CI: 99.3, 100.0), and negative predictive value 99.9% (95%CI: 99.8, 100.0). Probabilistic matching was most accurate when including address variables and using the automatically generated threshold for determining matches with manual review.

Conclusion: With the establishment of national electronic datasets across health and social care, EMS enables previously unanswerable research questions to be tackled with confidence in the accuracy of the matching process. In scenarios where a small sample is being matched into a very large database (such as national records of hospital attendance) then the positive predictive value or sensitivity may drop according to the prevalence of matches between databases. Despite this possible limitation, probabilistic linkage has great potential to be used where accurate deterministic linkage isn’t possible, including in low-income settings, and for vulnerable populations, where the absence of unique identifiers has historically hindered the ability to identify individuals across separate datasets.
Abstract 1327

**INCREASING THE SECURITY OF BLOOM FILTER BASED PRIVACY PRESERVING RECORD LINKAGE**

**Authors:**
Rainer Schnell  
*University of Duisburg-Essen, German Record Linkage Center*

**Presenting Author:** Rainer Schnell

**Corresponding author email:** rainer.schnell@uni-due.de

Linking data bases is an increasingly used medial research technique, for example in epidemiology and health service research. Since many medical data bases do not contain personal identification numbers, pseudo-identifiers such as names and birthdays have to be used for linkage. Due to privacy concerns, pseudo-identifiers have to be encrypted. Standard encryptions of error-prone pseudo-identifiers will miss many true matching pairs between different data bases. Hence, special techniques for linking databases with encrypted pseudo-identifiers have to be used. This now popular field of research is called privacy preserving record linkage (PPRL).

Most PPRL techniques invented by computer scientists cannot be used under the organizational constraints of medical data bases. Hence, only very few PPRL have been used for real world applications in medical research. Recently, Bloom filters are increasingly used in medical PPRL with very large datasets.

However, two research groups have reported successful attacks on Bloom filter records. The presentation will describe both approaches (a Constraint Satisfaction Approach with limited success and a nearly complete classical Cryptanalysis). Both approaches have led to new insights into how the encryption of Bloom filters can be improved.

The presentation will demonstrate different strategies of hardening Bloom filters against cryptographic attacks such as combining fields, sampling, use of different hash-functions, random-bits and fake injections. Furthermore, recommendations on modifying identifiers in pre-processing will be given.

Finally, results on the record linkage properties of modified Bloom filters will be presented.

---

Abstract 1687

**EXPERIENCE WITH THE NGLMS: ADVANCED DATA MANAGEMENT AND LINKAGE ANALYSIS TECHNIQUES USE NETWORK GRAPHS AND GRAPH THEORETIC TECHNIQUES**

**Authors:** James Farrow  
*Farrow Norris / SA NT DataLink, Australia*

**Presenting Author:** James Farrow

**Corresponding author email:** james@fn.com.au

The Next Generation Linkage Management System (NGLMS) used by SANT DataLink to manage linked population-spine and third-party data, persists and manipulates pairwise record comparisons and comparison weights along with the underlying records. The NGLMS constructs a long-lived network graph of relationships between records (as opposed to merely flagging which records ‘belong together’). The records are stored as vertices and comparisons as weighted edges joining vertices.

Composing projects as layers of data record relationships, customisation by bringing data sets and their effects on data linkage in and out at will. Sensitivity/specificity characteristic of different strategies may be explored by varying query parameters without having to ‘relink’ data since comparison data is persisted rather than discarded.

Linkage and analysis becomes separated into independent steps:

- entry/cleaning
• comparison calculations
• project composition (data/parameter selection)
• record cluster/group extraction

allowing some to be varied while holding the others constant, e.g. the effect of adding/removing a
data set can be examined without recomputing comparisons; or, the effect of varying weight
thresholds and comparison weights can be used to generated confidence intervals around the effect
of the linkage process on subsequent data analysis steps.

By persisting information as a network/graph:
• data is always 'live' and may be amended, augmented or removed at will while minimising recomputation
• the effect of ‘bad’ data sets and historical decision can be removed at a subsequent time despite intervening linkage activity
• multiple projects may use the same data using different thresholding values and quality parameters without interfering with one another and without necessitating multiple copies of master linkage information
• ethically sensitive information (such as identity changes) may be incorporated into some
projects and not others
• other types of linkage information (such as familial or genomic or economic/social
associations) may be stored and utilised
• a rich taxonomy of link types can be utilised rather than just ‘similarity’, e.g. mother/child and
whether this relationship is biological, legal, social, gestational/surrogate
• aspects of the graph topology may be exploited to understand data characteristics, e.g. the
density of edges in a well-connected cluster or the diameter of a loosely connected cluster

The approach lends itself to advanced graph theoretical techniques which may be too expensive to
run on all records (100's of millions) but which, after partitioning into subclusters, may be
computationally feasible, e.g. a min-plus algebra or tropical semiring may be used to analyse the
distance matrix equivalent to the graph to ascertain similarity paths and clusters.

Abstract 1719

BRINGING IT ALL TOGETHER: HOW TO JOIN AND ANALYZE SENSITIVE DATA FROM
MULTIPLE SOURCES

Authors:
F Endel¹, B Glock²
¹Vienna University of Technology, ²DWH

Presenting Author:
F Endel

Corresponding author
e-mail: florian.endel@student.tuwien.ac.at

Routinely collected claims data and other statistical databases storing pseudonymized and
anonymized personal information from the social security and healthcare system are increasingly
available for analysis in many countries. In Austria, many authorities are responsible for different parts of
these systems, depending on region, service type and organizational background. As a result, data
collections are heterogeneous and each one can only provide a restricted and biased view.
Additionally, joint analyses are challenging, not only due to availability, quality and comparability of
data, but also because of technical, organizational, legal and privacy reasons.

We are reporting about the various approaches we have developed in Austria for sharing and
potentially merging healthcare data from multiple sources.

First, primary care, specialized outpatient care and prescription claims data from 19 different social
security institutions had to be merged into an integrated database. While common pseudonyms exist,
challenges including quality issues, the harmonization of accounting and coding systems as well as the
definition of a reliable data model were key aspects.

**Second**, data from these databases had to be linked with inpatient databases which are under still different responsibilities and do not share any personal identifiers. Therefore special techniques of record linkage, including probabilistic matching had to be developed. The specific procedures have been presented at the SHIP conferences in 2011 and 2013. A common database covering all sectors of the healthcare system for 2006 and 2007 is now available, covering over 95% of the Austrian population. While further years are added, procedures are steadily improved.

The **next** challenge is to link this database with data collections outside the health sector (e.g. unemployment data), which usually do not share common identifiers. Conventional record linkage techniques are often not viable here as sharing personal information is legally difficult. For this purpose, we have developed a stepwise procedure minimizing the risk of data leakage.

Finally, international data exchange and common analysis across countries is an emerging challenge. In the EU FP7 project CEPHOS-LINK, fully anonymized data from six participating European countries is planned to be shared cross-border. In addition to requirements of country specific legislations, several technical issues from harmonized variable definition and trustworthy anonymisation up to secure data transfer have to be mastered for joint analyses. As alternatives to actually transferring anonymised and therefore comparably restricted information, new techniques of distributed analysis and statistical modelling are planned to be deployed.

---

**Abstract 1748**

**THE MASSACHUSETTS PREGNANCY TO EARLY LIFE LONGITUDINAL (PELL) DATA SYSTEM: ITS DEVELOPMENT, CHARACTERISTICS AND CHALLENGES IN THE U.S. AND INTERNATIONAL CONTEXT**

**Authors:** Milton Kotelchuck¹, H Diop², E Declercq³, Pell Leadership Team⁴

¹Massachusetts General Hospital, ²Massachusetts Department of Public Health, ³Boston University School of Public Health, ⁴Boston University School of Public Health; Massachusetts Department of Public Health; U.S. Centers for Disease Control and Prevention

**Presenting Author:** Milton Kotelchuck

**Corresponding author email:** mkotelchuck@mgh.harvard.edu

The MA Pregnancy to Early Life Longitudinal (PELL) data system was created to utilize and enhance a broad range of existing public health databases to examine the impact of pregnancy and birth experiences on subsequent maternal and child health.

The aim of this talk is to introduce the MA PELL data system and its development, emphasizing political data access and data confidentiality issues; technical data linkage issues; and administrative and funding challenges – and to use our PELL experiences to highlight the U.S. context for data linkage systems development efforts, and contrast it with other international efforts.

The MA PELL data system is an innovative, population-based, longitudinal reproductive and early life data system, with multiple linked datasets that can be used for cross-sectional and longitudinal analyses. It is perhaps the most preeminent U.S. reproductive public health life course data system. PELL developed as a public-university partnership between the Massachusetts Department of Public Health (MDPH), Boston University School of Public Health/Massachusetts General Hospital and the Centers for Disease Control and Prevention, the initial funding agency.

Currently, PELL contains information on 1,200,000 births and fetal deaths and over 700,000 mothers, starting from 1998. PELL is a relational data system composed of individual datasets linked together by randomly-generated unique IDs for mothers, deliveries, and infants. The core of PELL consists of the annual probabilistic linkage of Massachusetts birth and fetal death certificates with the birth-related hospital discharge records of both mother and infant. The PELL core datasets are then linked to other statewide public health datasets (including non-birth inpatient hospital and emergency department...
utilization; the MA Birth Defects and Cancer Registries; mortality data; various public health program participation records; and Pregnancy Risk Assessment Monitoring System surveys). Birth and death data are geo-coded.

The Massachusetts Commissioner of Public Health authorized the PELL data system and access to the confidential data. Each analytic study must receive further IRB approvals. Data involving confidential identifiers is linked at MDPH on secure servers and de-identified datasets are extracted for analyses.

PELL is now widely used for epidemiologic research, surveillance monitoring, program evaluation, quality improvement initiatives and policy assessments - addressing late-term prematurity, GDM, Early Intervention, autism, substance use, ART, elective C-sections and racial disparities. [Some examples will be highlighted.]

The presentation will conclude with a discussion of new (PELL) data linkage opportunities, and the evolving data access, confidentiality and administrative challenges facing population-based life course databases in the U.S.

SESSION C2 – DATA QUALITY: PRIMARY CARE.
CHAIR: DR TIM WILLIAMS, CLINICAL PRACTICE RESEARCH DATALINK

Abstract 1749 USING ROUTINE HEALTH RECORDS IN DEMENTIA RESEARCH: COMPARISON OF DEMENTIA DIAGNOSES IN HOSPITAL ADMISSIONS RECORDS WITH INFORMATION FROM PRIMARY CARE IN ENGLAND

Authors: Jane Green, Anna Brown, Oksana Kirichek, Valerie Beral
Cancer Epidemiology Unit, University of Oxford

Presenting Author: Jane Green

Corresponding author email: jane.green@ceu.ox.ac.uk

Background: Large cohort studies with long-term linkage to routinely-collected health records provide valuable opportunities for investigation of lifestyle risk factors associated with dementia, with the potential to overcome both limitations of power in smaller epidemiological studies and bias due to reverse causation, a crucial issue for a disease of insidious onset. It is however unclear how reliable are diagnoses of dementia in such records.

Methods: We compared diagnoses of dementia (ICD-10 F00-F03, G30) available from the central NHS hospital admissions database in England (Hospital Episode Statistics, HES) with those recorded in primary care. Linked HES records were available for 1.2 million women in the Million Women Study cohort, and we wrote to the primary care physician (General Practitioner, GP) for random samples of 33 women with and 1004 without a HES diagnosis of dementia, asking whether their patient had ever been diagnosed with dementia.

Results: Usable information was received from GPs for 72% (240/333) of the women with a HES diagnosis of dementia and for 85% (851/1004) of the women without a HES diagnosis of dementia. Preliminary results suggest that about 85% of diagnoses of dementia in HES were confirmed by the GP, with HES diagnoses of specified dementia types such as Alzheimer’s disease being more likely to be confirmed than HES diagnoses recorded only as unspecified dementia. Among women with no HES record of dementia, GPs confirmed the absence of a diagnosis of dementia in 99% of cases; for only one woman out of 851 did the GP report that she had ever been diagnosed with dementia.

Further analyses will be presented, including a comparison between linked electronic HES and Clinical Practice Research Datalink primary care records.

Conclusions: Hospital admissions records of incident dementia in England appear sufficiently reliable for use in large-scale epidemiological studies. This is true particularly of diagnoses of specified types of
dementia, including Alzheimer’s disease.

**Abstract 1448**

**USING NHS PRIMARY CARE DATA TO IDENTIFY UNDIAGNOSED DEMENTIA**

**Authors:**

E Jammeh¹, C Carroll², S Pearson², J Escudero¹, J Zajicek², E Ifeachor¹

¹Plymouth University School of Computing and Mathematics, ²Plymouth University Peninsula Schools of Medicine and Dentistry

**Presenting Author:** J Zajicek

**Corresponding author email:** emmanuel.jammeh@plymouth.ac.uk

**Background:** Up to 50% of patients with dementia may not receive a formal diagnosis, limiting access to appropriate services. It may be possible to build a picture of “underlying undiagnosed dementia” from a profile of symptoms recorded in routine clinical practice.

**Aim:** To develop a machine learning tool to identify patients who may have underlying dementia but have not yet received formal diagnosis from analysis of routinely collected NHS data.

**Method:** Routinely collected NHS READ-encoded data were obtained from 18 consenting GP surgeries across Devon, UK, totalling 26,483 patient records of those aged >65 years. 850 Patients were identified by READ code as having dementia within the 2 year study period (June 2010 to June 2012). We determined other codes assigned to these patients that may contribute to dementia risk. The dataset was used to train a supervised classifier (Naïve Bayes) to discriminate between patients with underlying dementia and healthy controls using a ten-fold cross-validation approach.

**Results:** The model obtained a sensitivity of 84.42% and a specificity of 86.67% for identifying dementia.

**Conclusion:** Routinely collected NHS data can be used to identify patients who are likely to have undiagnosed dementia. This type of methodology is promising for increasing dementia diagnosis within primary care.

**Abstract 1437**

**IS IT TIME FOR A RETHINK? IMPROVING THE CAPTURE AND USE OF FAMILY HISTORY INFORMATION IN PRIMARY CARE**

**Authors:**

Paul Nathan¹, J. C. Wyatt², Susan Clamp³

¹Hollybrook Medical Centre, University of Leeds and RCGP, ²University of Leeds, ³University of Leeds

**Presenting Author:** Paul Nathan

**Corresponding author email:** paul.nathan@nhs.net

Patients often present to primary care clinicians following significant family events to discuss their family history and personal health risk. It has long been known that appropriate identification of patient at higher risk of genetic disease who are offered early screening helps reduce health inequalities. Despite this family history is not used in any proactive manor. NICE gives conflicting advice about when to offer cascade screening, suggesting this should be undertaken with familial hypercholesterolaemia and not with breast cancer.

A Population based study using Researchone was undertaken to review the use of family history codes in patients with familial hypercholesterolaemia and breast cancer under the age of 40 yrs (males and females). Results demonstrate that family history is not consistently recorded in these high risk groups and screening has often not taken place.

Primary care uses family history codes to record this information. There are difficulties with this system, as free text is often used to record exact diagnosis, and which relative they were but this cannot be used
for decision purposes, to aid risk assessment

Patient memory of medical family history deteriorates with time. Therefore many people do not know if their increased risk of developing genetic disease without being offered appropriate screening.

It is time to ensure there are national cascade screening programs and health service clinical systems have the functionality required to ensure patients are identified and screened.

Abstract 1405

**SCOPE, COMPLETENESS AND CONCORDANCE OF NATIONAL CANCER REGISTRATION DATA LINKED TO THE CLINICAL PRACTICE RESEARCH DATALINK PRIMARY CARE DATA**

**Authors:** Krishnan Bhaskaran¹, R Williams², M Ranopa¹

¹London School of Hygiene and Tropical Medicine, ²Clinical Practice Research Datalink

**Presenting Author:** Krishnan Bhaskaran

**Corresponding author email:** krishnan.bhaskaran@lshtm.ac.uk

**Background:** English cancer registrations are collated in the National Cancer Data Repository (NCDR), and are individually linked to primary care records for patients in the Clinical Practice Research Datalink (CPRD) linkage scheme, opening up new possibilities for cancer-related research. We aimed to establish the completeness of key cancer-related variables in NCDR data linked to CPRD primary care data, and to describe concordance between the two datasets for cancer diagnoses.

**Methods:** Using the NCDR dataset linked to CPRD primary care data, we used ICD codes to identify all diagnoses between 1990-2010 of the 20 commonest UK cancers (breast, lung, prostate, colorectal, malignant melanoma, non-Hodgkin lymphoma, bladder, kidney, brain/central nervous system, pancreas, leukaemia, uterus, oesophagus, ovary, stomach, oral, myeloma, liver, cervix, thyroid). We evaluated the completeness of NCDR variables indicating tumour stage, grade, and treatment. We then used Read codes to identify diagnoses of the same 20 cancers in CPRD primary care data, and evaluated concordance between the two data sources by treating NCDR as the gold standard data source, and calculating the positive predictive value (PPV) and sensitivity of cancer Read codes in CPRD primary care data.

**Results:** We identified 294,883 records in NCDR for the 20 cancer sites of interest; numbers for individual cancers ranged from 2,641 (thyroid) to 61,996 (breast). Data on stage at diagnosis were only available for 6/20 cancer sites, with completeness of 67% for colorectal cancer, 61% for cervical cancers, and <50% for malignant melanoma, uterus, breast and ovarian cancers. For 17/20 cancer sites some tumour grade information was present, but completeness again varied by site from ≥70% (colorectal, uterus, oral cavity) to <20% (pancreas, liver, thyroid). A majority of records for all cancer sites included basic data on treatment modalities used. The majority of cancer diagnoses in CPRD primary care records matched a same-site record in NCDR (PPV=82% overall, and >70% for 17/20 cancer sites). The sensitivity of CPRD for capturing site-specific NCDR-recorded cancers was 71% overall but varied by site, ranging from 35% for kidney cancer to 88% for breast cancer.

**Conclusions:** Linked NCDR data added details on cancer diagnosis and treatment not typically present in primary care records, but data completeness was highly dependent on cancer site. Cancer Read codes in CPRD had high positive predictive value, but sensitivity was low for some cancer sites; this may be partly driven by use of clinical codes in primary care that do not definitively classify site or tumour behaviour.
Objective: To assess the frequency, overlap and diagnostic validity of heart failure recording across three national health record data sources in primary care, secondary care and national death certificates.

Design: Cohort study.

Setting: CALIBER linked electronic health record database, containing primary care data from 244 general practices in England participating in the Clinical Practice Research Datalink (CPRD), linked to Hospital Episode Statistics (hospital admissions) and the Office for National Statistics mortality register (cause specific mortality data).

Participants: 2.13 million patients in England registered with a general practice contributing to CPRD between January 1997 and March 2010.

Main outcome measures: Frequency and overlap of heart failure recorded across the three data sources, all-cause mortality, prevalence of heart failure risk factors, heart failure hospital (re-)admissions, and presence of supportive information within linked records concordant with heart failure diagnosis.

Results: Out of 2,134,615 patients, 79,468 patients had a heart failure diagnosis, 47,228(59.4%) were recorded in CPRD, 54,310(68.3%) in HES and 5008(6.3%) in ONS. Overall, 23,681(29.7%) had heart failure registered both in CPRD and HES. Lowest 5-year mortality was observed in 23,547(29.6%) patients recorded in primary care but never hospitalized for heart failure compared to those who had been hospitalized for heart failure at least once (HR, adjusted for age and sex 0.86, 95%CI 0.84 to 0.88, P<0.0001). 38,425(48.3%) patients with heart failure first diagnosed in secondary care showed highest 5-year mortality compared to patients who were first diagnosed in primary care prior to any hospitalization (HR, adjusted for age and sex 1.66, 95%CI 1.63 to 1.69,P<0.0001). Cardiovascular risk factors and comorbidity were increased in patients who were hospitalized for heart failure compared to primary care patients with no hospitalisation record. 17,182(21.6%) patients died in the first three months following heart failure diagnosis, of whom 71.8% had been hospitalized. In 60,042 out of 79,468(75.5%) patients, additional information was present supporting a heart failure diagnosis, of which loop diuretic(64.8%) and signs and symptoms(39.0%) were most frequently reported.

Conclusions: By linking primary care, secondary care and mortality data sources, we demonstrate that (1) HF diagnoses are incompletely captured by each source on its own; (2) there are large differences in disease severity, health care use, risk factor profile and survival in groups defined by EHR source of diagnosis; and (3) The majority of coded HF diagnoses can be corroborated by supporting evidence, thereby adding some support to the validity of the diagnosis once recorded.
Abstract 1717  EXPLORING NOVEL DIABETES SURVEILLANCE METHODS: A COMPARISON OF DIAGNOSTIC, LABORATORY, AND PHARMACY DEFINITIONS

Authors: Bushra Khokhar, G G Kaplan, H Quan, D Rabi
University of Calgary

Presenting Author: Bushra Khokhar

Corresponding author email: bushra.khokhar@ucalgary.ca

Background: Electronic Medical Record databases are widely used to study outcomes of diabetes and improve the quality and efficiency of diabetes care delivery. The objective of this study was to identify individuals with diabetes in a comprehensive EMR using different case definitions (clinical, pharmacy, laboratory definitions and a combination thereof) to identify and understand the differences in patient populations being captured by each definition.

Aim: The objective of this study was to identify individuals with diabetes in a comprehensive EMR using different case definitions (clinical, pharmacy, laboratory definitions and a combination thereof) to identify and understand the differences in patient populations being captured by each definition.

Methods: The data for this population-based cohort study was obtained from The Health Information Network (THIN). THIN is a longitudinal, primary care medical records database of over nine million patients in United Kingdom. The primary outcome was a diagnosis of diabetes, defined by the presence of a diabetes read code, or an abnormal laboratory result, or a prescription for an oral anti-diabetic or insulin.

Results: This study demonstrated that different case definitions of diabetes identify different sub-populations of patients - this is consistent with several recent studies have shown that both the prevalence of diabetes and the subjects diagnosed with diabetes vary when different diagnostic criteria for diabetes are applied. When observing the cohorts based on any measure of central tendency, it appeared as though each of the cohorts is reasonably comparable to each other. However, the sizes of the sub-populations vary widely, and the distribution of each of the cohorts when grouped by age categories and sex, reveal differences.

Conclusions: Our results suggest that different case definitions identify different sub-populations and using multiple case definitions is likely required to optimally identify the entire diabetes population within THIN.

References

Abstract 1613  INTRODUCING THE PRIDAL MODEL FOR LINKING ROUTINE HEALTH AND IDENTIFIABLE PATIENT REPORTED DATA

Authors: Martin Heaven\(^1\), Mark Kingston\(^2\), Hayley Hutchings\(^3\), Alan Watkins\(^2\), Helen Snooks\(^2\), Martin Heaven\(^1\)
\(^1\)FARR at Swansea University, \(^2\)College of Medicine, Swansea University, \(^3\)College of medicine, Swansea University

Presenting Author: Martin Heaven

Corresponding author email: m.l.heaven@swansea.ac.uk

Background: The rise in availability, quality and use of routine health data has resulted in well-developed methods for anonymised linkage of data from multiple sources. However methods for combining identifiable data (e.g. patient-reported questionnaires) with routine anonymised data are not yet tried-and-tested. Linking these data presents opportunities to improve the feasibility and effectiveness of observational and experimental studies, but emerging data linkage processes must address the appropriate balance between data security and usability (1).
Objectives

- To present an efficient privacy-protecting model for linking routine and identifiable patient-reported research data.

Methods: The Process for Routine and Identifiable Data Linkage (PRIDAL) was devised to efficiently link routine hospital data and patient-reported quality of life and quality of care questionnaire data as part of the PRISMATIC trial (2) – a mixed methods cluster randomised trial of the efficacy of an emergency admission risk tool in primary care.

PRIDAL was conceived by a group of specialists in e-trials, health informatics, information governance and process mapping who reviewed data sources, flows, owners, and security to develop a practical and intuitive process model.

Results: We will present the PRIDAL process model for the first time and use PRISMATIC findings from data matching of 2,400 questionnaire responses to patients in an overall study dataset of over 200,000. (Data matching will complete in April 2015). We will demonstrate that the model achieves high matching rate, and consider the lessons learnt in its application.

Conclusions: The linking of routine health and patient self-reported data presents valuable research opportunities, but clear, replicable models, are needed to support ethical and practical data linkage. We present the PRIDAL model as a potential solution.

Note: PRISMATIC is funded by the NIHR HS&DR programme (project number 09/1801/1054). The views and opinions expressed herein are those of the authors and do not necessarily reflect those of the HS&DR programme, NIHR, NHS, or the Department of Health

Abstract 1487  VALIDATION OF THE RECORDING OF ACUTE EXACERBATIONS OF COPD IN THE CLINICAL PRACTICE RESEARCH DATALINK

Authors: Kieran J Rothnie¹, Hana Mullerova², John R Hurst³, Liam Smeeth¹, Kourtney Davis², Sara Thomas⁴, Jennifer K Quint³

¹Faculty of Epidemiology and Population Health, London School of Hygiene & Tropical Medicine, London
²Respiratory Epidemiology, GlaxoSmithKline R&D, Uxbridge
³UCL Respiratory Medicine, University College London, London
⁴Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London

Presenting Author: Kieran J Rothnie

Corresponding author email: kieran.rothnie@lshtm.ac.uk

Background: Electronic health records (EHR) are an increasingly popular resource in which to conduct research as they can provide tremendous statistical power to answer many clinical questions. However, a limitation of EHR research is the validity of diagnostic codes and other strategies used to define outcomes.

Chronic obstructive pulmonary disease (COPD), a common respiratory condition triggered by smoking, is the fourth leading cause of death globally. People with COPD often experience periods of acute worsening of symptoms, known as exacerbations (AECOPD), and these are a substantial healthcare burden.

The Clinical Practice Research Datalink (CPRD) is a UK database of primary-care health records covering 11 million people, including at least 200000 COPD patients. Previous studies of AECOPD in EHRs have used prescription of antibiotics and steroids to identify AECOPD but it is unclear if this strategy is valid. We aimed to validate the recording of AECOPD in CPRD.

Methods: We randomly selected 1485 patients with a validated diagnosis¹ of COPD from CPRD. Combinations of prescriptions of oral steroids and antibiotics, symptoms, diagnostic codes for lower respiratory tract infections (LRTI) and AECOPD were used to create 16 possible definitions of AECOPD.
Questionnaires were sent to GPs asking for confirmation of their patients’ AECOPD on specific dates and for additional supportive information. Responses were reviewed independently by two respiratory physicians along with free-text for events for which GPs were uncertain. We calculated the positive predictive value (PPV) for each of the definitions. Post-hoc analysis excluded events occurring on annual review dates and restricted to those only with additional information.

**Results:** 998 (72%) questionnaires were returned, containing data for 8362 possible AECOPD. AECOPD codes had high PPV (96.1%, 95% CI 94.5-97.2), as did LRTI codes when combined with prescription of antibiotics and oral steroids (88.4%, 84.4-91.3%). The PPV for AECOPD hospitalisation was lower (48.4%, 40.3-56.5%), but increased in post-hoc analysis (95.4%, 84.2-99.4%). Prescription of antibiotics (60.4%, 59.2-61.6%) or steroids (70.9%, 68.3-73.3%) alone had lower PPV and were not improved in post-hoc analysis.

**Discussion:** Valid identification of AECOPD in CPRD is important as definitions used in clinical studies will affect results. Our study illustrates the importance of validation studies for diseases that are not necessarily straightforward to identify in EHRs. This study is a good example of how validation studies can be conducted and our methods can be applied to other diseases and databases.

---

**SESSION C3 – JOINED-UP INFORMATION GOVERNANCE**

**CHAIR:** PROF KIM MCGRAIL, UBC CENTRE FOR HEALTH SERVICES & POLICY RESEARCH

Abstract 1681  
**EXPANDING THE IDEAS OF PROPORTIONATE GOVERNANCE FOR USE IN BRITISH COLUMBIA**

**Authors:** Kim McGrail¹, Kaitlyn Gutteridge², Nancy Meagher²  
¹University of British Columbia, ²Population Data BC

**Presenting Author:** Kim McGrail

**Corresponding author email:** kim.mcgrail@ubc.ca

**Objective:** In British Columbia today, every request for access to data for research purposes is subject to the same time-consuming and intensive review. This presents challenges for timeliness of reviews and scalability as the current process is equally labour intensive for a simple request for limited data, and a request where there is intent to contact. A new framework is sought to support alignment of the level of risk of a given application to a proportionate level of review. The ultimate goal is to have a transparent process outlining what influences initial adjudication and assignment to a type or level of review.

**Approach:** The proposed approach builds on relevant legislation, OECD privacy principles, and experience including Scotland’s implementation of proportionate risk assessment (safe people, data, environment and public interest). We expand the framework to account for emerging access and privacy challenges, such as a broader user base, commercial interests, more and varied data, and progressive analytic approaches (e.g. hypothesis generating). Six proposed core principles of review are science, approach, data, people, environment, and interest, with a spectrum of risk for each ranging from low to very high. The principles and spectrum of risk create a grid, which represents all the possible scenarios under which access to data may be sought and provides a visual venue for contemplation of requests can be mapped to different levels of review. A significant challenge for the implementation of the framework is identifying consensus on what constitutes ‘low risk’ and ‘high risk’ and mapping risk profiles to a clear review process. Achieving this consensus requires engagement of data stewards, research ethics boards, researchers, privacy experts, legislators and the public.

**Results:** Stakeholder engagement is currently underway; the framework has been met with enthusiasm and significant interest. It is seen as supportive of important dialogue in challenging areas, such as commercial access, hypothesis generating investigation, and genomic data linkage. The framework is intended to accommodate these discussions, as it could clearly be applied to requests from individuals outside traditional academic settings, and for questions or techniques beyond standard hypothesis
testing.

**Conclusion:** The structured framework of principles and risk is a strong foundation for moving towards proportionate governance, and is flexible and adaptable for use in other jurisdictions.

---

**Abstract 1784**  
**INNOVATIVE GOVERNANCE IN PRACTICE - ESTABLISHING THE METADAC: A UK MULTI-AGENCY MULTI-STUDY DATA ACCESS STRUCTURE**

**Authors:** Madeleine Murtagh, Paul Burton, Sue Ring  
School of Social and Community Medicine, University of Bristol

**Presenting Author:** Madeleine Murtagh,  
**Corresponding author email:** madeleine.murtagh@bristol.ac.uk

In a world where rapid technological progress and an ageing population place ever increasing demands on medical and social care, one of our most potent weapons is the translation into societal benefit of scientific resources embodied in longitudinal cohorts studies and biobanks. Although it is easy to appeal to the simple idea of using these resources wisely, the reality is more challenging. Internationally, this is a major strategic issue for most large-scale data/sample generating biomedical and social studies as well as for their funders, who wish to see an optimal use of and return from their investments. The apparently simple administrative task of granting access to data and biosamples is often far from straightforward because of potential - or actual - conflicts of interest between study participants, data generators, data users, funders, ethico-legal experts, specialist lobby groups and policy developers. For publically funded studies these are all legitimate stakeholders, with different opportunities to express their perspectives. Appropriately balancing these views is a complex process involving understanding of the underlying science as well as the relevant ethical, legal, social and technical issues. Effectively managing extensive and complex data and biosamples requires active and innovative governance and technical oversight of access to data and biosamples.

From 1 May 2015 a multi-agency multi-study data access structure, the METADAC (Managing Ethico-social, Technical issues and Administration Data Access Committee), was established to service several of the UK’s major cohort studies (1958BC, 1970BC, Millennium BC and Understanding Society) and to provide a scalable mechanism to incorporate additional cohorts in the future (e.g. ELSA). The METADAC will develop, implement and maintain all administrative and technical activities plus policies needed to realise an access mechanism that is fit for purpose given the complex biomedical/social data and samples in question; its primary aim to optimise the use of the scientific resources under its governance. The METADAC is committed to a streamlined, proportionate and responsive form of governance which maintains full regard to the consents and expectations of the study participants who provided the research materials.

Complex decisions about data access are not simply ethical or technical; they go to the heart of the politics of scientific research. In this paper we discuss the METADAC in this context, outlining the technical, ethical and human dimensions of governing data access in practice.
Abstract 1595 PUBLIC BENEFIT AND PRIVACY PANEL

Authors: Janet Murray1, Patricia Ruddy1, Muriel Douglas2, Aidan Robertson3, Daniel Beaumont3
1NHS National Services Scotland, 2NHSCentral Registry National Records of Scotland, 3Scottish Government eHealth

Presenting Author: Janet Murray

Corresponding author email: janet.murray1@nhs.net

Robust and efficient proportionate governance is a vital enabler in the research infrastructure, aiming to meet the twin public interests of protecting confidentiality and sharing data for research and other legitimate uses.

A national Public Benefit and Privacy Panel (PBPP) being established in Scotland will provide governance of health data on behalf of data controllers, replacing three separate processes for governing data from different sources: NHS National Services Scotland’s Privacy Advisory Committee (National Datasets); National Caldicott Scrutiny Panel (data from multiple NHS Boards); and CHI Advisory Group (unique identifier CHI).

The PBPP will focus on public awareness and concerns and public benefit, with a commitment to protecting and promoting privacy as a public good and in the public interest.

It has three main objectives:

- To provide a single consistent and transparent scrutiny process enabling health and social care data to be used for a range of purpose including research
- To ensure the right balance between safeguarding privacy of all people in Scotland and the fiduciary duty of Scottish Public Bodies to make best possible use of the data collected
- To provide leadership on a range of complex privacy and information governance issues ensuring that the population benefits from use of data while addressing public concerns regarding privacy and promoting the protection of privacy as a public interest.

The two-tier panel will operate a system of proportionate governance developed from the SHIP model by IG Experts in NHS National Services Scotland. Tier 1, comprising IG experts from across NHS Scotland, will provide technical IG scrutiny applying criteria to approve applications or refer for further consideration by Tier 2. Tier 2, comprising senior leaders, Caldicott Guardians, researchers and public representatives, will consider wider privacy issues in regard to contentious applications or proposed policies.

Time has been required to establish national commitment to such a panel. A SHIP review of the IG landscape included the proposal 2011. Building on a consultation conducted by NSS Privacy Advisor in 2013, the IG Policy lead from SG led its development with IG experts from NHS Scotland and National Records of Scotland. Crucially stakeholders have been engaged throughout and the panel is drawn from organisations across Scotland.

Potential benefits include

- opportunity for transparent discussion of public benefit in use of data
- promotion of privacy as a public good
- engagement of all stakeholders
- consistency and efficiency in decision making
This paper will present data management and research governance practices built from experience from across the world, and will discuss the challenges faced, and possible options for optimising the timely provision of privacy protected data for policy makers and researchers.

Challenges include how best to facilitate and support Data Stewards/Data Custodians releasing their data into a data repository ready for subsequent approved analysis and research. How to optimise access and use of data that does not rely on personal relationships. How to ensure community engagement in both conducting research studies and in the establishment and operation of privacy protecting data integration infrastructure.

While Memos of Understanding (MOUs), other Data Access Deeds signed between entities at CEO/Secretary and Minister level, and Confidentiality Deeds are necessary to enable data access by researchers and policy makers, and to enable the linkage of datasets from different entities, high level sponsorship alone does not guarantee the timely provision of data.

Governance processes including the authorisation of the transfer of personal details, the use of an authorised data linkage unit to facilitate the linking of data across entities - especially where there is no common person identifier, and the use of privacy protecting practices including maintaining the separation of personal details from the content data being released for approved research studies and analysis.

As part of an 'Open Data' policy for the South Australian Government, and the need to implement efficient provision of data to researchers and policy makers, the option of legislation has been considered. Legislation to facilitate the establishment and operation of a data repository to receive, securely store, clean and extract anonymised data on government funded services and administratively collected data.

An evaluation of Ontario’s legislative approach to open government data has been undertaken, to determine whether a similar approach might be effective in South Australia. In accordance with Ontario’s Personal Health Information Protection Act (PHIPA) the Institute for Clinical Evaluation and Science (ICES) is one of a handful of organisations designated by Ontario’s Information and Privacy Commissioner (IPC), with appropriate privacy practices to receive health data.

The Ontario approach has PHIPA as an enabler, IPC as an authoriser and ICES as a trusted and authorised holder of health data for analysis and public good. The South Australian legislative approach looks to expand the responsibility and capability beyond health and medical data to include all data collected by the South Australian Government human services.
Abstract 1673

HARM VERSUS IMPACT IN CASES OF DATA BREACH AND MISUSE: EXAMINING LEGAL INSUFFICIENCIES AND WHAT LIES BEYOND THE REGULATORY SPACE?

Authors:
Christine Dobbs¹, L.A. Stevens², G Laurie³
¹Farr Institute CIPHER, Swansea University, ²University of Edinburgh School of Law, ³University of Edinburgh School of Law & Mason Institute

Presenting Author: Christine Dobbs

Corresponding author email: c.dobbs@swansea.ac.uk

Background: In this paper we present a sub-section of findings from an evidence review commissioned by the Nuffield Council on Bioethics (NCOB) Working Party on Biological and Health Data and the Wellcome Trust’s Expert Advisory Group on Data Access (EAGDA), entitled Review of evidence relating to harms resulting from security breaches or infringements of privacy involving sensitive personal biomedical and health data. A multi-disciplinary piece of work, this paper reports from the fields of Law and Social Psychology.

Purpose: One of the review’s purposes was to allow the NCOB and EAGDA to better understand the nature of the ‘actual harms’ resulting from data misuse or security breaches. However, we contend that ‘actual harm’ as understood in the legal context cannot be extrapolated to the psychosocial context, because here subjectivity is the key. Therefore we distinguish between ‘harm’ (legal) and ‘impact’ (psychosocial), the latter ranging from none at all, to mild irritation, to extreme distress.

Method: Given the expansive remit, this was a scoping exercise. (1) Systematic searches were undertaken to establish ‘hard evidence’ (UK Case Law; ICO) and from Twitter, and a narrative search in the grey literature (newspapers, NGOs). (2) Where possible, each piece of evidence was examined to establish the type of breach/misuse and cause of abuse, and the resultant harm/impact. (3) Findings were noted in a matrix and then triangulated to compare and contrast evidence from each source.

Results: We identified 51 pieces of hard evidence, 77 pieces from Twitter and 57 from the grey literature. We compare and contrast our findings in this paper.

Discussion: Based particularly on the grey literature-based evidence, it is clear that several factors need to be considered if we are to shed light on what harm and distress can mean in social reality terms, and if and how subjects can be protected, supported and, perhaps, compensated. There are limits in how the law recognises and compensates for harms. In context of data processing, legal redress is typically contingent on causal connection to an economic loss. This neglects the broader spectrum of impacts that the psychosocial dimension elucidates. Earlier and closer attention must be paid to identify the range of interests at stake when processing biological and health data, including risks and benefits involved with a proposed use of data. We propose more explicit engagement with all population groups and an enacted recognition of their sensitivities within governance mechanisms.
DATA GOVERNANCE IN A COMPLEX HEALTH RESEARCH ENVIRONMENT

Authors: Kathy Sykora, Felix Zhang, Jason Chai-Onn, Nicholas Gnizdziejko, Mahmoud Azimaee, J. Charles Victor
Institute for Clinical Evaluative Sciences, Toronto, Ontario, Canada

Presenting Author: Kathy Sykora
Corresponding author email: kathy.sykora@ices.on.ca

The Institute for Clinical Evaluative Sciences (ICES) is a not-for-profit organization conducting research to evaluate health care delivery and outcomes in Ontario, Canada. Established in 1992, ICES houses a vast and secure array of linkable, de-sensitized health and health-related data on more than 13 million Ontarians.

Ontario’s Personal Health Information Protection Act grants ICES special status, enabling it to receive fully identified data, de-sensitize it, and use it to carry out its mission. Naturally, this status is conditional on maintaining high standards to reduce privacy and security risks.

A Data Governance model is used to implement access to data. The challenge is to balance privacy, security and usability. At ICES, data access is guided by these principles:

1. Access to data is granted on a need-to-know basis. Fully identified data is only accessible to a handful of staff whose role is to remove or encode sensitive information, make data linkable to the rest of the repository, and perform data management tasks.
2. Certain sensitive fields are only accessible to those with a demonstrated need to use them.
3. Access to data is project driven. Every ICES project undergoes a Privacy Impact Assessment; authorizing the use of specific datasets for the project.
4. Logging and audits exist for reporting and ad-hoc investigations.

ICES data are stored as SAS datasets on a UNIX system accessed using SAS software. ICES uses a combination of tools to implement this model and control access to data:

1. UNIX integration with Active Directory (AD) provides flexibility in managing and auditing user and group memberships. SAS Metadata users and groups synchronize with specific AD organizational units enabling single sign-on throughout the system.
2. Users’ level of access groups determine the sensitivity of data they access, based on need.
3. SAS Metadata-Bound libraries secure and control Data Repository access.
4. AD groups are used to control access to data collected or created for the project.
5. A two-tiered group system is implemented: Each data library is assigned a data group, and a project group is defined for each project’s users.
6. Tasks are de-centralized. SAS Management Console is used by IT and data administration personnel. Assignment of users to project groups is done directly by users’ managers.

This paper presents ICES’ data governance structure from the information technology and security perspective, to provide other organizations with a model for handling sensitive data in a complex research environment.
Abstract 1760  

ECOUTER: ENGAGEMENT FOR GOVERNANCE AND POLICY IN DATA INTENSIVE BIOMEDICAL SCIENCE: TRUST AND DATA LINKAGE

Authors: Madeleine Murtagh¹, J T Minion², R C Wilson², A J Turner², O W Butters², C A Ochieng², B P P Murtagh²  
¹University of Bristol, ²Social and Community Medicine, University of Bristol

Presenting Author: Madeleine Murtagh

Corresponding author email: Madeleine.Murtagh@bristol.ac.uk

Engaging participant has become axiomatic; its rationales are both ethical and practical. Engagement, done well, is routinely understood both as marker of respect for participants and as a means of recruiting or retaining participant involvement. With the increasing complexity of data and biomedical sciences and rapidly changing social and technological context important but unforeseen policy issues may arise. In such settings a third rationale for engagement becomes crucial, i.e. input from participants to develop appropriate governance and policy responses to emerging issues. Existing deliberative democracy mechanisms can be useful but these rely on substantive existing knowledge, which in the case of new technological, legal or methodological approaches within the study design may be outside participant’s routine knowledge or experience.

ECOUTER [Employing COnceptual schema for policy and Translation Engagement in Research], French for ‘listen’, is a method for engagement which incorporates existing evidence that can help participants draw upon their own knowledge of cognate issues and interact on a topic of shared concern. An ECOUTER begins with an initial question posed by the organisers as well as links to a selection of key pieces of the existing evidence base. Participants are invited to respond and contribute additional ideas and links. The online discussions are analysed by social scientists to generate a conceptual framework of the phenomenon or issue considered. The results can then form the basis of recommendations for research, governance, practice and/or policy.

In this paper we present our experience of two ECOUTER events: First, the P3G ECOUTER, was run as an online ECOUTER involving ethical, legal and social science researchers across the globe which began with the question, “Can I trust you with my data?” Second, the Shopping Centre ECOUTER, was run as an interactive exhibition during the ESRC Festival of Social Science at a local inner city shopping centre in Bristol, involving the general public, asking the question “Your Health Records: Hand over or hands off?” The P3G ECOUTER produced a conceptual map including the themes: Defining trust and publics, the importance of context, governance, data security, consent, education and trust. The Shopping Centre ECOUTER produced a conceptual map including the themes: engagement with care.data, the benefits of research for health, privacy and confidentiality and data as a commodity. We present these conceptual maps and compare the responses of the researcher participants in the P3G ECOUTER and Shopping Centre ECOUTER.
Abstract 1714  
TOWARDS ENABLING PUBLIC TRUST AND ENGAGEMENT IN THE RE-USE OF HEALTHCARE DATA: PATIENT PERSPECTIVES ON SHARING DATA AND A DIGITAL SYSTEM FOR DYNAMIC CONSENT

Authors:  
Karen Spencer¹, Will G Dixon¹, Caroline Sanders¹, Edgar A Whitley², Jane Kaye³, David Lund⁴  
¹University of Manchester, ²Department of Management, London School of Economics and Political Science, ³HeLEX - Centre for Health, Law and Emerging Technologies, Nuffield Department of Population Health, University of Oxford, Oxford, United Kingdom, ⁴HW Communications, Lancaster, United Kingdom

Presenting Author: Will G Dixon

Introduction: Although recent studies have shown that patients support confidential access to their anonymised electronic patient record (EPR) for research, concerns have arisen surrounding security, privacy, potential misuse of data, willingness to share data with certain groups and clear processes for opt-out. Additionally, the lack of transparency and engagement regarding the re-use of healthcare data is thought to have undermined public trust as illustrated by response to plans for sharing primary care records in the UK. One proposed solution is to use technology to facilitate dynamic consent whereby patients can tailor preferences for who they share their data with, whilst also providing feedback about data recipients and research results. The purpose of this study was to undertake qualitative research to a) explore patient attitudes to the use of anonymised health data for research purposes, and b) to evaluate patient perceptions of a dynamic consent model.

Methods: 40 participants with a rheumatic condition were recruited to take part in one of three focus groups and/or a semi-structured interview (n=26). Data were analysed thematically using key techniques from grounded theory.

Results: Three broad themes characterised participants' views on the use of EPRs for medical research and their perceptions of dynamic consent: 1) 'The role of trust and perceived social responsibility to share health data'; 2) 'Transparency through dynamic consent and patient feedback, and the potential for enhanced empowerment and patient engagement' 3) 'Operational and technological scope and challenges for participation'. Participants had a high level of trust in the NHS system but this was threatened by concerns about certain non-NHS recipients and potential misuse of data. There were uncertainties about the degree to which they would want to express consent preferences, but they particularly valued potential for research feedback which is usually absent. Most participants found the system easy to use but a small number would not want to engage.

Discussion/Conclusion: This study has demonstrated a favourable view of a technical solution to help maintain trust in the re-use of health data and to meet the needs of recent national recommendations to bring greater transparency and patient engagement associated with this. The work has provided important insights for consideration when refining design of the intervention to enable implementation and piloting in practice.
DATA LINKAGE TO SUPPORT A MODEL OF CHILD CANCER SURVIVOR CARE

Authors: ML McBride¹, K Goddard¹, S Pritchard², C Fryer², F Walks¹, P Byron²
¹British Columbia Cancer Agency, ²British Columbia Children’s Hospital

Abstract 1459

Background/Objectives: Although more effective treatments and screening have led to increasing numbers of individuals surviving cancer, survivors may be at increased risk of long-term morbidity due to the cancer, treatment complications, or effect on existing comorbidities. High-quality evidence of late effects risks and their predictors, and risk-stratified models of survivor follow-up care, are needed for effective, efficient surveillance and management. Follow-up of adult-age childhood cancer survivors in British Columbia (BC) occurs mainly in the primary care setting, with little coordination and support. Results from a research program, linking clinical records of all child cancer cases diagnosed in BC from 1970 to mortality, cancer, and health administrative databases for up to 40 years, were used to inform recommendations for strategies and a model of care to meet ongoing healthcare needs of this survivor population.

Design/Methods: Results from the BC Childhood, Adolescent, Young Adult Cancer Survivor Research Program were used in a needs assessment; evaluation of ongoing healthcare demand and costs; identification of gaps in care; and determination of the size and characteristics of subgroups requiring different levels of care.

Results: There were approximately 3000 adult-age survivors. The number of survivors transitioning to adult-age care increased each year by 3%. Over 40% lived more than 70 km from the main pediatric hospital. The proportion of at-risk survivors receiving at least one recommended follow-up surveillance test ranged from 0.8% (thyroid-stimulating hormone) to 87% (complete blood count). By 20 years post-diagnosis, 20% of survivors had no conditions leading to hospitalization, and 35% had four or more types of these conditions. High users of later outpatient services included survivors of central nervous system (brain) tumours, bone sarcomas, or leukemia other than acute lymphatic leukemia; and those with previous cranial radiation, combination chemotherapy and radiation, combination surgery chemotherapy and radiation, or stem cell transplant. These data contributed to the development of risk-stratified models of care and a business case for implementation.

Conclusion: Working group recommendations led to funding for a formal risk-tiered adult-age childhood cancer survivor follow-up program in BC. This original research demonstrates that population-based longitudinal database linkages of clinical and health administrative data provide high-quality evidence to inform the development of models of quality, sustainable, cancer survivor care, for both children and adults, that can improve health care policy and practice. This approach can inform oncologists, primary care providers, program managers and policymakers of cost-effective strategies for long-term surveillance and care.
Abstract 1561  
RISK STRATIFICATION OF LONG-TERM MORBIDITY AMONG CHILDHOOD CANCER SURVIVORS: USING DATA LINKAGE TO INFORM A MODEL OF CARE

Authors:  
ML McBride, D Li, F Walks  
British Columbia Cancer Agency

Presenting Author:  
ML McBride

Corresponding author email:  
mmcbride@bccrc.ca

Background: Longterm risks of negative impacts of childhood cancer and its treatment vary considerably. Models of risk are needed to stratify survivors into different levels of intensity and healthcare setting needs, in order to provide both appropriate and efficient care.

Objective: Using longitudinal, linked, population-based individual cancer registry, clinical, and health administrative data for a childhood cancer survivor (CCS) cohort, we stratified CCS into levels of intensity of long-term morbidity, based on healthcare utilization, and determined the prevalence and predictors of each level.

Materials and Methods: Demographic records of CCS diagnosed in the province of British Columbia, Canada, under age 25 years between 1970 and 2005, surviving at least five years from diagnosis, were identified from the provincial cancer registry, and linked to clinical data and provincial health insurance administrative records. Healthcare utilization up to 20 years post-diagnosis was stratified into three levels of intensity of utilization. Multinomial logistic regression was applied to determine significant predictors of each stratum.

Results: 2508 (72.5%) CCS survived 5 years or more; at 10-, 15- and 20- years post diagnosis, numbers of CCS still alive were 1992 (79.4%), 1515 (60.4%), 1071 (42.7%), respectively. At the end of follow-up, healthcare utilization of 1757 survivors were obtained through linkage, with average 17 years follow-up. 251 (14.3%), 699 (39.8%), and 807 (45.9%) of survivors had high, moderate, and low levels of healthcare use, respectively. Older age at diagnosis, and being female (OR = 3.05 (2.23-4.18), were associated with high healthcare use. Central nervous system (brain tumour) survivors were four times as likely to be high users of healthcare (OR = 4.03, (2.03-8)) compared to the low-risk acute lymphoblastic leukemia survivors. Chemotherapy, radiotherapy, alkylators, growth stimulating factor, relapse or second cancer were also predictors of high healthcare use. Over the trajectory of cancer survivorship, CNS diagnosis, radiotherapy, early relapse and second cancer were all predictors of high level of use at 5-, 10-, 15-years post diagnosis. At 20-years post diagnosis, being female, having radiation to cranial-spinal axis or chest, and early relapse remained predictors of high levels of healthcare.

Conclusion: This study confirms that the CCS population can have significant late-occurring and longterm morbidity, that be stratified into risk groups based on level of risk of late effects, using linked longitudinal health data. Identification of predictors allows for development of models of care that will provide tailored care depending on intensity and types of risks in each group.

Abstract 1522  
CHILDHOOD PNEUMONIA IN ENGLAND: INTERVENTIONS IN PRIMARY CARE PRIOR TO HOSPITAL ADMISSION

Authors:  
Pia Hardelid, L Wijlaars, S Denaxas, R Gilbert  
University College London

Presenting Author:  
Pia Hardelid

Corresponding author email:  
p.hardelid@ucl.ac.uk

Background: Community-acquired pneumonia (CAP) remains a common reason for hospital admission in children in England, which might potentially be reduced by improved prevention in primary care through pneumococcal vaccination and targeted use of antibiotics. Hospital admissions for
pneumonia in the NHS are therefore considered ‘ambulatory-care sensitive’ and preventable through primary care interventions. We determined the proportions of children who had visited their GP in the week prior to hospital admission, and whether they were prescribed antibiotics.

**Methods:** We used a linked primary care (the Clinical Practice Research Datalink, CPRD) and hospital admission (Hospital Episode Statistics) dataset through the Cardiovascular Disease Research using Linked Bespoke Studies and Electronic Health Records (CALIBER) platform. We extracted all CAP-related primary care consultations and hospital admissions in children aged 0 to 19 years between 2007 and 2009, after the introduction of universal infant pneumococcal vaccination. We calculated hospital admission rates for CAP by age, gender and area-level deprivation per 1000 child-years (cy) with 95% confidence intervals (CIs) and examined independent risk factors using Poisson regression models. Using the first CAP admission per child, we determined the proportion of children who had a primary care consultation, and the proportion prescribed antibiotics in the week before admission.

**Results:** We extracted data for 551,415 children, who had 2637 episodes of CAP in primary and/or secondary care; 1560 (59.2%) of these were admitted to hospital. The overall incidence in primary care was 0.4/1000 cy (95% CI 0.4, 0.5) and 0.9/1000 cy (0.8, 0.9) in secondary care. There was a clear socioeconomic gradient in hospital admissions (incidence rate ratio comparing children living in most vs least deprived areas 1.41, 95% CI 1.17, 1.70). Of 1325 first CAP admissions, 61.1% (810 children) had had contact with primary care in the week before admission. Of them, 428 had been referred by their GP (32.3% of all admitted children). Overall, 13.1% of children had been prescribed an antibiotic in the week prior to admission.

**Conclusion:** Two thirds of children admitted to hospital with CAP see their GP in the previous week, and only 13% had an antibiotic prescribed. Further research is needed to determine the role of antibiotics, pneumococcal vaccine in at-risk groups, and what other primary care and community-based interventions, if any, can prevent hospital admissions for childhood CAP.

---

**Abstract 1565**

**CHANGING TRENDS IN CHILDREN ENTERING OUT-OF-HOME CARE IN ENGLAND AND BARRIERS TO EXPLORING THEIR EDUCATIONAL OUTCOMES**

**Authors:**

Louise Mc Grath-Lone¹, L Deardon², B Nasim¹, K Harron³, R Gilbert¹

¹ADRC-E, ²Institute of Education, ³London School of Hygiene and Tropical Medicine

**Presenting Author:** Louise Mc Grath-Lone

**Corresponding author email:** louise.mc-grath-lone.13@ucl.ac.uk

**Background:** On average 60,000 children are in out-of-home-care (OHC) annually in England, representing approximately 0.5% of the child population; however, the cumulative proportion entering OHC over their life course is not known. Policy makers acknowledge a strong interrelationship between health, education and social care, for example, mental and physical health problems are more common among looked-after children (LAC) and they have poorer educational outcomes. This study aimed to (1) describe the age-specific cumulative proportion of children entering OHC for the first time for non-respite reasons and (2) explore the feasibility of creating cohorts of children ever placed in care and followed up for school attainment using administrative data, namely the Department for Education’s one-third sample of Children Looked After (CLA) return from 1992 to 2012.

**Methods:** First, LAC (N=89,168) were grouped by year of birth and rates of first entry to OHC by age, gender and ethnicity were calculated using denominator data derived from population estimate averages. The coverage of Unique Pupil Number (UPN) among a 1992 birth cohort of LAC was also described. UPN, assigned on enrollment to a state-maintained school or nursery, is the identifier used by DfE to link children’s social care and educational records.

**Results:** Overall 3.4% of children born 1992-94 entered OHC by age 18. Variation in this cumulative proportion was observed by gender (3.6% boys vs. 3.1% girls) and ethnicity (4.8% black vs. 1.6% white children). Of 7,501 children born 1992 who entered care by age 18, 26% had missing UPN as they exited...
before school entry. A further 25% in care during school years did not have UPN recorded. Name is not recorded in CLA and other available identifiers (i.e. date of birth, sex and postcode at entry to care) were not sufficiently specific for probabilistic matching.

Conclusion: Incomplete coverage of UPN in CLA means it cannot be used to explore the potential effects of early childhood experiences of care on educational outcomes, primarily because the majority of LAC enter and leave care before school entry. A change in policy to allow the collection of names in CLA should be considered to link children’s social care and educational records. Policy makers should also consider the increasing proportion of children entering OHC when planning health services as each LAC will require an individual health plan and regular assessments.

Abstract 1611

DESCRIBING THE EPIDEMIOLOGY OF RESPIRATORY SYNCYTIAL VIRUS IN INFANTS AND YOUNG CHILDREN IN ENGLAND USING A ROUTINE LABORATORY DATABASE

Authors:
Rachel Reeves1, Ruth Gilbert2, Pia Hardelid2, Joanna Ellis3, Richard Pebody3
1Farr Institute of Health Informatics, 2University College London, 3Public Health England

Presenting Author: Rachel Reeves

Corresponding author email: rachel.reeves.14@ucl.ac.uk

Background: Respiratory syncytial virus (RSV) is the most frequent cause of lower respiratory tract infections in infants and young children worldwide. A range of RSV vaccines are currently in development. Potential target groups for a future licensed RSV vaccine include pregnant women, infants 4-6 months of age and children >6 months of age. However, infection in infants and young children has not been explored in sufficient detail to inform decisions on the optimal age and target groups for immunisation in England. This study examines variation in the number of laboratory confirmed RSV infections in children <5 years of age in England from 2009-2014 by age, season and birth month, to contribute to informing potential vaccine strategy.

Methods: Weekly laboratory reports of RSV positive respiratory samples in children <5 years of age in England were extracted from Labbase2, held by Public Health England, for the period July 2009 to July 2014. For patients with multiple records during a single episode of infection (within 14 days), only the first record was used in the analysis. For each birth month, cumulative risk of laboratory confirmed RSV in the first year of life was calculated using Office of National Statistics data on live births per month per year as the denominator. We calculated risk ratios with 95% confidence intervals comparing infants by birth month, using infants born in February as the baseline.

Results: We identified 38,560 RSV positive episodes. Each season peaked at week 52, excluding seasons 2012/13 and 2013/14 which peaked at week 48 and 50 respectively. 82% (n=31,676/38,560) of reported episodes were in infants. 62% (n=23,890/38,560) were in infants <6 months of age: 27% (10,542/38,560) within the first two months of life, 22% (8,484/38,560) in months 2-3 and 13% (4,864/38,560) in months 4-5. On average, infants born in September, October and November had the highest risk of laboratory confirmed RSV compared to infants born in February, with risk ratios of 2.9 (95% CI, 2.7-3.1), 3.6 (95% CI, 3.3-3.9) and 3.6 (95% CI, 3.3-3.8) respectively.

Conclusions: The results suggest a high burden of severe RSV infection in the first 6 months of life, particularly infants born around the beginning of RSV season. Further work is planned to analyse RSV-related hospitalisations, including linkage of laboratory and hospital records, to determine whether these results reflect the true burden of RSV in early life. This could have important implications for the design of future vaccine programmes.
WHY DO MORE CHILDREN DIE IN ENGLAND THAN IN SWEDEN? AN INTER-COUNTRY COMPARISON OF UNDER-5 MORTALITY USING LINKED NATIONAL REGISTERS.

Authors:  
Anna Zylbersztejn1, Ruth Gilbert2, Pia Haridelid2, Anders Hjern3 
1The Farr Institute of Health Informatics Research, 2Institute of Child Health, UCL, 3Centre for Health Equity Studies (CHESS), Stockholm University

Presenting Author: Anna Zylbersztejn

Corresponding author email: ania.zylbersztejn.14@ucl.ac.uk

Background: England has one of the highest childhood mortality rates in Western Europe, while Sweden has one of the lowest. Comparison of cohorts that are similar in terms of key risk factors at birth, such as low birth weight and gestational age, could show where such discrepancies originate and inform policy makers as to what the focus of preventive strategies should be. The aim of this study is to investigate whether children born in England have an increased risk of death compared to children born in Sweden with similar characteristics at birth and to assess which groups are most at risk.

Methods: We examined mortality rates and rate differences per 1000 births with 95% confidence intervals by birth weight category (>1000g, 1000-1499g, 1500-2499g, 2500-3499g, 3500-4499g, ≤4500g) and gestational age category (22-27, 28-31, 32-34, 35-36, 37-41, 42-45 weeks). Mortality rates were investigated at 0-6 days, 7-27 days, 28-364 days and 1-4 years. Differences in the rates between England and Sweden were tested using two-proportion z-test.

Results: We identified all births for 2002-2006 using the Swedish Medical Birth Register (n=477,153) and Hospital Admission Statistics (n=2,787,807). Information on deaths was obtained through linked data from the Swedish Cause of Death Register (n=1,492) and the Office for National Statistics (n=12,928). In the English cohort, 44.6% and 48.3% of birth records were missing information on birth weight and gestational age respectively. The issue was minor in the Swedish cohort (<1% missing).

Total mortality rates were higher in England than in Sweden for all age groups. For deaths at 7-27 days, the differences in rates were significant for both gestational age (0.13, 95% CI: 0.06-0.19, p=0.001) and birth weight (0.22, 95% CI: 0.16-0.29, p=0.000). Analysis of cohorts subdivided by gestational age and birth weight, however, showed insignificant results, suggesting that the disparities are driven by other risk factors in this age group. For deaths at 28-364 days the differences were most significant for ‘low-risk’ babies – those born at 37-41 weeks of gestation (0.24, 95% CI: 0.15-0.33, p=0.000) and those weighting 2500-3499g (0.28, 95% CI: 0.13-0.42, p=0.001), suggesting that improvements could be achieved for this group.

Discussion: Use of linked national registers shows great potential for identifying sources of discrepancies in childhood mortality between England and Sweden. The results, however, are preliminary as further investigation of missingness mechanisms and potential bias due to miscoding of gestational age and birth weight in the datasets is ongoing.
**Abstract 1742**

**USING LINKED DATA TO ASSESS EQUITY ACROSS CHILD AND YOUTH MENTAL HEALTH AND ADDICTIONS INDICATORS IN ONTARIO, CANADA**

**Authors:** Julie Yang¹, Sima Gandhi¹, Saba Khan¹, Kelvin Lam¹, John Cairney², Paul Kurdyak³, Astrid Guttman⁴

¹Institute for Clinical Evaluative Sciences, ²McMaster University; Institute for Clinical Evaluative Sciences, ³Centre for Addiction and Mental Health; Institute for Clinical Evaluative Sciences, ⁴The Hospital for Sick Children; Institute for Clinical Evaluative Sciences

**Presenting Author:** Julie Yang

**Corresponding author email:** julie.yang@ices.on.ca

**Background:** The social determinants and their link to mental health have been studied extensively. Exposure to social disadvantage in childhood and adolescence is particularly important as it is an important developmental period. Social disadvantages are also of interest as they are potentially modifiable risk factors. As a part of a government-wide mental health and addictions strategy in Ontario, Canada, the Institute for Clinical Evaluative Sciences (ICES) developed a scorecard for child and youth mental health. One of the strategy themes was closing critical gaps for vulnerable populations. As such, equity was a key domain addressed.

**Methods:** An equity lens was applied to all scorecard indicators instead of having a separate category of equity indicators. This was done by stratifying each of the 25 indicators by immigration status from the Citizenship and Immigration Canada register, by neighbourhood income quintiles calculated through the 2006 Canadian Census data, and by age, sex and geography across Ontario. These data were all linkable to health administrative data held at ICES using an individual, linkable unique identifier for the entire 13 million people in Ontario held within the ICES Data Repository. Three-year average rates were calculated for all equity lenses.

**Selected Findings:**

- The treated prevalence of schizophrenia was higher in refugees and immigrants than in non-immigrants (26.3 versus 18.0 versus 11.0 per 10,000 population, respectively). However, refugees and immigrants were less likely to have alcohol or drug problems.

- Income was inversely related to high service need. Compared to children and youth in the highest-income neighbourhoods, those from the lowest-income neighbourhoods had the highest rates of emergency department visits (11.1 versus 7.4 per 1,000) and hospitalizations (2.4 versus 1.7 per 1,000).

- Children and youth from rural and remote regions were three times more likely to use acute care services for mental health and addictions but less than one-fourth as likely to use outpatient physician services, suggesting a misalignment between service provision and need.

**Conclusions:** The ability to apply an equity lens to all indicators allowed for a richer description of child and youth mental health than having a separate category of equity indicators. Health services use and outcomes varied according to the equity lenses used, signalling inequities in the system or differences in vulnerability and need within the population. This information is implicitly useful to inform policy that will help to improve child and youth mental health care and outcomes in Ontario.
Abstract 1764  UNRAVELLING 33 YEARS OF PAEDIATRIC ADMISSIONS FOR EPILEPSY IN SCOTLAND USING ROUTINE ADMINISTRATIVE DATA (1981 – 2013)

Authors:  'Tomi Ajetunmobi1, R F Chin2, C J Weir2
1University of Edinburgh/Information Services Division, Scotland, 2University of Edinburgh

Presenting Author:  'Tomi Ajetunmobi

Corresponding author  
email: tomi.ajetunmobi@ed.ac.uk

Background: There is a need to improve paediatric health outcomes in the UK, particularly for chronic conditions such as epilepsy. Recent hospital audits for children with epilepsy have highlighted several areas for improvement. However, little is known of the patterns of health care utilisation, in children with a diagnosis of epilepsy that may inform better health care.

Methods: An anonymised extract of retrospective acute care records for children aged less than 18 years admitted to Scottish hospitals between 1981 and 2013 (inclusive), was obtained from the Information Services Division (ISD), Scotland. It comprised demographic and clinical management information for each admission record. Hospital admissions with a primary and secondary diagnosis of epilepsy were identified using ICD9 and 10 codes (345 and G40/G41 respectively). Neonatal admissions were excluded. Approval for the study was obtained from the Privacy Advisory Committee, NHS National Services Scotland. We describe trend analyses conducted to explore patterns of hospital admission in the cohort.

Results: Over the study period, there were 56,495 paediatric records of children admitted to 138 Scottish hospitals with a diagnosis of epilepsy (16,463 children). Overall, 6% of the admissions were of infants aged less than 1 year, 26% of children aged 1-4 years, 28% aged 5-9 years, 25% aged 10-14 years and 15% of the admissions were of young people aged 15-17 years. Males made up more than half of the admissions (52%). Short stays of up to one day and emergency events comprised 54% and 69% of the admissions respectively.

Between 1981 and 2013, the rates of hospital admission with a diagnosis of epilepsy increased from 84.9 to 214.4 admissions/100,000 children. This increase was particularly marked in the rates of hospital admissions with a secondary diagnosis of epilepsy (from 23.8 to 110.5 admissions per 100,000 children between 1981 and 2013). There was also a significant increase in the rates of emergency admissions with a primary diagnosis recorded by 43%, particularly amongst children aged less than one year (by 45%).

Conclusions: There was a marked increase in hospital admissions amongst children with a diagnosis of epilepsy; our analysis suggests that improved recording of comorbidities may play a role. Better information is needed to support the integration of paediatric health care systems for children with a diagnosis of epilepsy.
Abstract 1369  PATIENT-LEVEL VARIATION IN ANTIBIOTIC USE IN PRIMARY CARE: AN EPIDEMIOLOGICAL STUDY USING ELECTRONIC HEALTH RECORDS

Authors:  Laura Shallcross, Nick Beckley, Irene Petersen, Greta Rait, Andrew Hayward
UCL

Presenting Author:  Laura Shallcross

Corresponding author email:  l.shallcross@ucl.ac.uk

Background:  Over the last 70 years antibiotics have revolutionised how we practise modern medicine. Unfortunately most antibiotics have a finite lifespan because drug resistant strains emerge; an inevitable consequence of bacterial evolution. Resistance is mainly driven by inappropriate antibiotic use and although the majority of highly drug resistant infections emerge in hospitals, 80-90% of antibiotics for human use are prescribed in primary care. Most commonly antibiotics are prescribed for respiratory tract, urinary tract or skin and soft tissue infections. Despite guidance on the management of these syndromes there is major variation in rates of antibiotic prescribing between practices within the UK, suggesting scope for optimising antibiotic use in primary care.

To date, initiatives to improve prescribing have been aimed at all patients in primary care, irrespective of whether they are rarely or frequently prescribed antibiotics.

Methods:  We included data from patients registered with a practice that participated in The Health and Improvement Network (THIN) scheme of data collection between 2011 and 2013. Data on antibiotic prescriptions was used to determine the average number of antibiotic prescriptions per person-year for each class of antibiotics. We plotted frequencies of the number of antibiotics per person-year for each antibiotic class using a general inverse power model and identified cut-off values to identify the 10% of patients with the highest antibiotic consumption.

Results:  In primary care 10% of patients have on average at least 1.6 prescriptions for a broad spectrum penicillin per year. Five per cent of registered patients have more than 2.2 prescriptions for broad spectrum penicillin per year on average, (figure 1). At least one percent of patients have more than 4 prescriptions for broad spectrum penicillin per year.

Discussion:  In primary care most patients have less than one antibiotic prescription per year. 95% of patients have fewer than two prescriptions per year for broad spectrum penicillin, which is the class of antibiotics most commonly prescribed for respiratory tract infections in the community. One percent of patients are prescribed at least four courses of broad spectrum antibiotics per year, equivalent to 530,000 patients based on an estimated population of 53 million people in England.

Current public health initiatives aim to reduce total antibiotic use across the population, yet most individuals are prescribed less than one antibiotic per year. Reductions in antibiotic use may be more easily achieved by targeting public health interventions at patients who are frequent users of antibiotics.
Abstract 1470  ORAL ANTICOAGULANTS IN SCOTLAND – UTILISATION, CLINICAL EFFECTIVENESS AND SAFETY

Authors: Tanja Mueller¹, S Alvarez-Madrazo¹, S McTaggart², B Johnston¹, C Robertson¹, M Bennie¹
¹University of Strathclyde, ²NHS National Services Scotland

Presenting Author: Tanja Mueller

Corresponding author email: tanja.muller@strath.ac.uk

Background: Thromboembolic events represent a substantial burden of disease in many countries, including Scotland. Anticoagulants are a mainstay of treatment in clinical practice, especially in atrial fibrillation where anticoagulant drugs are used long-term. Warfarin has been the established therapy, and its utilisation and effects have been widely studied; however, similar information regarding the newly approved drugs dabigatran, rivaroxaban and apixaban is still limited.

Objectives: To describe and analyse utilisation and clinical outcomes of the new oral anticoagulants (NOACs) in Scotland, compared to traditional warfarin treatment.

Methods: A retrospective longitudinal cohort study will be used to a) analyse prescribing trends and utilisation of new oral anticoagulants over time and b) determine clinical effectiveness and adverse events associated with different treatment options in patients with a confirmed diagnosis of atrial fibrillation (AF).

Two cohorts, comprising users of oral anticoagulants and AF patients respectively, have been identified in databases of routinely collected data: the Prescribing Information System (PIS) and the Scottish Medical Records/Inpatient and day care dataset (SMR01). Eligible anticoagulant treated patients within PIS have been identified using the British National Formulary (BNF) classification for oral anticoagulants (02.08.02.), while AF patients have been identified using the designated ICD-10 code for atrial fibrillation (I48). A subsequent record linkage based on the Scottish Community Health Index (CHI) facilitates patient follow-up from April 2009 to June 2014; PIS, SMR01 and two additional datasets – the Scottish Medical Records/Outpatient attendance dataset (SMR00) and the National Records of Scotland (NRS) – will be used for this linkage to ensure availability of a wide range of demographic, socio-economic and medical information for each cohort participant.

Results: Analysis of a subset of the data showed that during the first six months of 2014, 7.1% of all prescriptions for oral anticoagulants in Scotland were issued for NOACs, ranging from 1.5% to 18.6% dependent on Health Board. Overall, the most commonly prescribed drug was rivaroxaban with 45.8 prescriptions per 10000 inhabitants; figures for apixaban and dabigatran were 9.1 and 7.2, respectively. 55.0% of all patients receiving oral anticoagulation were male, and the patient’s mean age was 72.0 years (standard deviation 12.8 years). However, variations with regards to age and gender between the different drugs have been identified.

Conclusions: This observational study will provide previously unavailable population-based information from real-world settings about the utilisation and effectiveness of new oral anticoagulants, in addition to existing data based on clinical trials.
APPLICATION OF NATURAL LANGUAGE PROCESSING METHODS TO EXTRACT CODED DATA FROM FREE TEXT PRESCRIPTION DOSE INSTRUCTIONS

Authors: Clifford Nangle1, David Robertson2, Stuart McTaggart1, Elizabeth Nicholson1, Jackie Caldwell1, Margaret MacLeod1, Marion Bennie1

1Public Health Intelligence, NHS National Services Scotland, Edinburgh, 2School of Informatics/Usher Institute, University of Edinburgh

Presenting Author: Clifford Nangle

Corresponding author email: c.nangle@nhs.net

Background: Medicines are the most frequently used health technology (accounting for approximately 15% of NHS Scotland costs) and their safe and effective use, supported by pharmacoepidemiological studies, is a key part of routine clinical practice.

The Prescribing Information System (PIS) data mart, hosted by NHS National Service Scotland receives around 90 million electronic prescription messages per year from GP practices across Scotland. Prescription messages contain information held on the GP10 prescription form (e.g. drug name, quantity and strength) stored as coded, machine readable, data while the prescription instructions are unstructured free text and difficult to interpret and analyse in volume.

The Farr Institute Scotland Pharmacoepidemiology Workstream has identified that storing drug dosage (quantity / unit of measurement) and consumption frequency, as coded data in PIS, is a priority to make the system analysis ready for clinical studies.

Aim: To automate the extraction of drug dose, unit, and frequency metadata from prescription instructions using Natural Language Processing (NLP) methods.

Method: An NLP algorithm was developed using the Ciao implementation of Prolog (Bueno, et al., 2011) to extract drug quantity, unit, and frequency metadata from prescription instructions. An accuracy estimate was obtained by randomly sampling source records and comparing the source text with the metadata extracted by the algorithm.

Results: Applying the NLP algorithm to a study on Methadone prescribing in Scotland between July 2009 and June 2013, consisting of 554,598 prescriptions, the NLP algorithm extracted dose, unit, and frequency metadata from 488,267 (88%) prescriptions. The accuracy of the metadata extracted varied from 90% to 99%.

Next Steps: Work is being undertaken to generalise the algorithm so that it can be applied against all prescription instructions held in PIS with improved success rates and accuracy. During the development of the prototype algorithm, it became apparent that for large studies:

1. Statistical methods need to be used to measure the validity and accuracy of the data extracted by the NLP algorithm to quantify bias; and
2. The accuracy and validity of the metadata extracted should be published in order to support research led decision making

References
METHADONE PRESCRIBING IN SCOTLAND: JULY 2009 TO JUNE 2013

Authors: Sheila Bird1, P Dimitropoulos2, R Turner2, S McTaggart3, L Nicholson2, M Bennie3
1 MRC Biostatistics Unit, Cambridge CB2 0SR, 2 MRC Biostatistics Unit, CAMBRIDGE CB2 0SR, 3 National Services Scotland, EDINBURGH EH12 9EB

Presenting Author: Sheila Bird

Corresponding author email: sheila.bird@mrc-bsu.cam.ac.uk

Background: The World Health Organization’s (WHO) recommended daily dose of methadone is 60ml. The Scottish press has complained that methadone’s cost is not properly accounted for because Scotland does not know the number of persons per annum for whom methadone prescriptions are written; nor is information published on how the average daily dose of prescribed methadone compares with WHO’s recommendation of 60ml per day or varies by region, gender and age-group. Scotland’s drug-related death-rates per 1000 problem drug users varied importantly by region in 2009-2013.

Aim: To estimate how many persons in 2009/10 versus in 2012/13 were prescribed methadone in Scotland, and compare how the quantity prescribed varies by gender and age-group when region, year and prescription-source (GP, other) are taken into account. Gender and age-group are known only if the prescription had an associated Community Health Index (CHI).


Methods: Descriptive statistics backed by Bayesian hierarchical modelling of quantity prescribed to take account of repeat prescriptions for the same client.

Findings: Of 515079 methadone prescriptions in 2009/10, 61.4% were written by GPs. Gender and age-group were available for 247529 (78.3%) of GP-prescriptions but for only 22.5% of 198922 other-source prescriptions. Of 478094 methadone prescriptions in 2012/13, 55.6% were written by GPs. Gender and age-group were available for 229204 (86.2%) of GP-prescriptions but for only 47.2% of 212137 other-source prescriptions.

Mean quantity of methadone per prescription decreased in 2012/13 (versus 2009/10) by:
- 46 ml (95% CI: 40 to 52 ml) for GP-prescriptions with gender & age-group
- 83 ml (95% CI: 74 to 92 ml) for other prescriptions with gender & age-group

Prescriptions for females outnumbered those for males in the youngest age-group (15-24 years), the reverse of three older age-groups. For each age-group, mean quantity of methadone was significantly lower for females than males and, for GP-prescriptions, increased as the client aged but, by 2012/13, the age-related increase in mean quantity per other-source prescription was stable for the oldest age-group (45+ years).

Interpretation: The quantity per prescription reduced significantly in 2012/13, and differed by prescription-source; quantity prescribed varied importantly by gender (lower for females) and age-group. Detailed estimates from Bayesian hierarchical modelling will be presented.

However, analysis of prescription details beyond quantity and instalments (needed for reimbursements) is essential for governance. And, with fewer than 80% of prescriptions CHI’d, estimates for Scotland’s number of methadone-clients range widely: 30,000 to 50,000.
Abstract 1721

ESTIMATING THE ASSOCIATION BETWEEN ANTIMICROBIAL PRESCRIPTION AND HEALTHCARE ASSOCIATED INFECTIONS USING DATA LINKAGE

Authors: Jiafeng Pan¹, Kimberley Kavanagh¹, Charis Marwick², Camilla Wiuff³, Peter Davey⁴, Scott Bryson⁴, Chris Robertson⁵, Marion Bennie⁶

¹University of Strathclyde, ²University of Dundee, ³Health Protection Scotland, ⁴NHS Greater Glasgow and Clyde, ⁵University of Strathclyde; Health Protection Scotland

Presenting Author: Jiafeng Pan

Corresponding author email: pan.jiafeng@strath.ac.uk

Introduction: Associations between antimicrobial prescribing, antimicrobial resistance (AMR) and Clostridium difficile infection (CDI) are well documented but there are limited good quality studies quantifying the relationships between primary care prescribing and the prevalence of CDI and resistance. Quantification of these effects through national, individual-level linked datasets would support antimicrobial stewardship policy in improving antimicrobial use.

Methods: This study investigates (i) the effect of community prescription of any antimicrobial including the known high risk antimicrobials for CDI – clindamycin, cephalosporins, fluoroquinolones (ciprofloxacin) and co-amoxiclav - in the months prior to a diagnosis of community-associated CDI (CA-CDI); (ii) the cumulative effect of antimicrobials on CA-CDI.

We have conducted a matched case control analysis via linkage of three Scottish patient level data sets: laboratory confirmed infections (ECOSS), prescriptions for antimicrobials in primary care (PIS) and all hospital admissions (SMR01). From the ECOSS data we identified all cases of CDI reported in the period August 2010 to July 2013. Each CDI case has been assigned up to 6 population-based controls on the basis of age, sex and location, with their corresponding records retrieved from each data set. Conditional logistic regression was applied to assess the impact of antibiotic exposure on the risk of CA-CDI. Potential confounding variables – exposure to Proton Pump Inhibitors or H2 antagonists, previous hospital admissions, comorbidity, care home residency status and social deprivation are included and adjusted for in the analysis.

Results: In the study period, there were a total of 1557 CA-CDI cases without recent (prior 3 months) hospitalisation, with 8560 matched controls. Exposure to any antimicrobial in the previous 6 months increased the risk of CA-CDI (adjusted OR 2.63, 95% CI 2.25-3.07). Exposure to any high risk antimicrobial, compared with no antimicrobials, had higher associated risks (adjusted OR 5.69, 95% CI 4.46-7.27). Risk of CA-CDI increased with cumulative antibiotic exposure: compared with no antibiotic exposure, up to 1 week of cumulative exposure to any antibiotics approximately doubled the associated odds, while >4 weeks quadrupled the odds.

Conclusions: This study provides a national linked patient level data set and establishes the scale of the risk of CA-CDI given antimicrobial exposure and, in particular, cumulative exposure. This methodology allows for expansion to other infections and AMR and is an exemplar of the use of NHS Scotland’s developing Infection Intelligence Platform which will place Scotland as a world leader in the use of health informatics to support infection control and antimicrobial stewardship.
Abstract 1452

SYSTEMATIC EVALUATION OF THE IMPACT OF BIAS IN A PHARMACOEPIDEMIOLOGICAL STUDY OF STATIN USE AND CANCER RISK

Authors: M Ranopa1, I Douglas1, T van Staa2, L Smeeth1, K Bhaskaran1
1Department of Non-Communicable Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, United Kingdom, 2Institute of Population Health, University of Manchester, Manchester, United Kingdom

Presenting Author: M Ranopa

Corresponding author email: michael.ranopa@lshtm.ac.uk

Background: There has been a wealth of pharmacoepidemiological studies investigating the association of statin use and the risk of cancer. However, there have been conflicting findings, which may have been influenced by biased study designs.

Objective: We aimed to measure and compare the impact of five different biases in the context of the statin-cancer association utilising the UK Clinical Practice Research Datalink (CPRD).

Methods: We utilised the CPRD to estimate the relative risk (RR) of breast, colorectal, lung, and prostate cancer (four of the most common cancers in the UK) among patients prescribed statins. We assumed no causal association exists between statin use and cancer risk based on previous published literature. Five commonly cited biases were examined, namely: immortal time, protopathic, healthy user, prevalent user, and time window bias. Two main analyses were conducted for each bias: (i) a biased analysis, and (ii) a corresponding corrected analysis. We measured the level of impact of such biases by estimating the percentage difference of risk attributable to bias (PDRAB).

Results: The impact of immortal time bias on the statin-cancer association was marginal. PDRAB ranged from 0.9% (prostate cancer) to 1.9% (lung cancer). Confidence intervals for both biased and corrected analyses across all cancers included 1. Similar results were observed for protopathic bias: PDRAB ranging from 0.9% (prostate cancer) to 1.0% (lung cancer). For healthy user bias, PDRAB ranged from 1.9% (prostate cancer) to 15.0% (colorectal cancer). Again confidence intervals for both biased and corrected analyses included 1. Prevalent user bias yielded PDRAB estimates ranging from 0.0% (prostate cancer) to 6.1% (breast cancer). Time window bias yielded the most variability in terms of PDRAB, which ranged from 38.5% (lung cancer) to 45.8% (breast cancer). Biased analyses yielded statistically significant protective effects across all four cancer types, while corrected analyses showed no association for all four cancer types.

Conclusions: We assessed the impact of five commonly cited biases which have been previously noted as potential drivers of discrepant results. In the context of the statin-cancer association, the impact of 4 of the 5 biases examined were marginal; only time widow bias showed a consistent change in effect from protective (biased analysis) to a null association (corrected analysis). More attention is needed to consider how such biases may impact on results in order to ensure confidence in findings and to rule out possible factors that may drive conflicting results between studies.

Abstract 1467

THE VALUE OF A DISTRIBUTED NETWORK OF DATA CENTRES IN QUANTIFYING UNCOMMON ADVERSE EFFECTS OF DRUGS

Authors: David Henry1, Samy Suissa2
1ICES, 2McGill University, Montreal, Canada

Presenting Author: David Henry

Corresponding author email: david.henry@ices.on.ca

Uncommon adverse effects of drugs are increasingly being investigated by distributed analyses of linked population health data held at multiple sites. Distributed analysis overcomes the challenges of
pooling multi-jurisdictional data. Two examples of distributed drug safety networks in North America are Sentinel (USA) and CNODES (Canada). CNODES, part of the Drug Safety and Effectiveness Network (http://www.cihr-irsc.gc.ca/e/46981.html), was established in 2011 with funding and support from Health Canada and the Canadian Institutes for Health Research. CNODES comprises over 60 researchers and analysts working in 7 Canadian provinces with additional data obtained from the Clinical Practice Research Database (UK) and Marketscan (USA). The large population base (>44 million) has enabled assembly of large cohorts to quantify uncommon outcomes (e.g. 2 million users of statins). CNODES has developed rigorous procedures to ensure consistent analyses across multiple sites, to minimize selection bias, misclassification and confounding and to eliminate selective outcome and analysis reporting. Results from each provincial team are collected and meta-analyzed by a separate Methods team. Experience to date has included the investigation of: high-potency statins (renal damage, diabetes), proton-pump inhibitors (community-acquired pneumonia), antipsychotics (hyperglycemic emergencies), oxycodone dispensing near the US border crossings (drug trafficking), serotonin norepinephrine reuptake inhibitors (acute kidney injury), isotretinoin (occurrence of pregnancy during treatment), domperidone (serious cardiac events); quetiapine (cardiomyopathy, acute liver injury), incretin-based therapies (pancreatic cancer, pancreatitis, heart failure). This experience has provided useful insights into the challenges and opportunities presented by distributed data networks, which will be discussed at the conference. The experience of CNODES and Sentinel (and other international networks) indicates that this will become the standard approach for investigating serious adverse effects of therapeutic drugs after marketing.

Abstract 1376 MEANINGFUL EVENTS FROM EHR PRESCRIPTION DATA
Authors: R Williams, B Brown, N Peek, I E Buchan
MRC Health eResearch Centre; NIHR Greater Manchester Primary Care Patient Safety Translational Research Centre, University of Manchester
Presenting Author: R Williams
Corresponding author email: richard.williams2@man.ac.uk

Introduction: Primary care research databases derived from EHRs contain coded prescription information. Much of this information is unstructured or unordered, and interpretation is required to assign clinical meaning to it. A common approach is to consider the dosage or presence of a drug as a single covariate, however, to contextualize prescribing to the clinical decisions made on care pathways it is necessary to convert the EHR data into events, such as when therapy is commenced, terminated or changed. These events are useful for: care pathway analysis; process mining; next-generation phenotyping; realistically-complex quality indicators; and audit. We focus on antihypertensive medications.

Method: We used an anonymized extract of primary care data from 53 general practices in Salford, UK (population 234k). All prescriptions of drugs recommended by NICE for the treatment of hypertension were extracted. We created a mapping between each drug code, the active ingredient(s) and the tablet dose (mg). We found 653 Read codes and 199 vendor-specific codes, covering 178 brand names and 70 generic names. Text mining, using regular expressions, on 216,101 distinct patient instructions yielded the number of tablets taken per day. We then iteratively developed an algorithm to take the amount, frequency, duration and type of medication to extrapolate events. Two authors (RW and BB, a clinician) independently reviewed a random sample of 100 patients to determine if the correct sequence of events had been extracted, with disagreements resolved through discussion.

Results: The algorithm was developed over 6 iterations involving the examination of 179 patient records. 10,311,973 prescriptions were extracted for 81,096 patients between 1977 and 2014 (54% female, median age 64, IQR [50, 77]). The algorithm produced 850,028 events (28% starts, 34% stops, 15% restarts, 16% dose increases, 8% dose decreases, 0.02% unclassified). During validation the algorithm achieved a PPV of 92% (95% CI 85%-96%). Of the 100 records reviewed only 8 had incorrect sequences and these were still largely correct. Four missed a single stop event, two were false increases due to an
erroneous prescription, one had an extra stop, and one misclassified a switch from 2.5mg indapamide to 1.5mg modified release (clinically identical) as a decrease.

**Discussion:** We have developed a method for transforming unstructured EHR prescription data into clinically meaningful events on a care pathway. From these events we can determine: when a patient is taking medication; adherence issues; when an intervention is/isn’t made by a physician; and when guidelines are followed correctly.

---

**Abstract 1330**

**ADOLESCENT CANNABIS AND TOBACCO USE AND EDUCATIONAL OUTCOMES AT AGE 16: BIRTH COHORT STUDY**

**Authors:** Alex Stiby¹, Matthew Hickman¹, Marcus R Munafo², Jon Heron¹, Vikki L Yip¹, John Macleod¹

¹School of Social and Community Medicine, University of Bristol, ²MRC Integrative Epidemiology Unit, University of Bristol

**Presenting Author:** John Macleod

**Corresponding author email:** A.Stiby@bristol.ac.uk

**Background:** Cannabis use has been consistently shown to be associated with poor educational outcomes but this association may not be causal. Issues of bias, reverse causation and residual confounding can make causal inference difficult. Record linkage can help address these issues.

**Aims:** To investigate the relationship between cannabis and tobacco use by age 15 and subsequent educational outcomes.

**Design:** Birth cohort study.

**Setting:** England.

**Participants:** The sample was drawn from the Avon Longitudinal Study of Parents and Children; 1,155 individuals had complete information on all variables

**Measurements:** Exposures were cannabis and tobacco use at age 15 assessed by computer assisted questionnaire and serum cotinine. Outcomes were performance in GCSE examinations in English and Mathematics at age 16, attainment of 5 or more GCSEs at grade C or higher and leaving school with no qualifications all ascertained through linkage to the English National Pupil Database. Covariates considered were maternal substance use, life course social position, child sex, month and year of birth, child substance use prior to age 15 and child conduct problems all from study data and child educational attainment at age 11 from record linkage.

**Findings:** In fully adjusted models both cannabis and tobacco use at age 15 were associated with subsequent adverse educational outcomes. In general the dose response effect seen was consistent across all educational outcomes assessed. Weekly cannabis use was negatively associated with English GCSE results (Grade Point Difference [GPD], -5.93, 95% CI, -8.34, -3.53) and with Mathematics GCSE results (GPD, -6.91, 95% CI, -9.92, -3.89). Daily tobacco smoking was negatively associated with English GCSE results (GPD, -11.9, 95% CI, -13.47, -10.33) and with Mathematics GCSE (GPD, -16.72, 95% CI, -18.57, -14.86). The greatest attenuation of these effects was seen on adjustment for other substance use and conduct disorder. Following adjustment tobacco appeared to have a consistently stronger effect than cannabis.

**Conclusions:** Both cannabis and tobacco use in adolescence are strongly associated with subsequent adverse educational outcomes. Given the non-specific patterns of association seen and the attenuation of estimates on adjustment it is probable that these effects are not causal though a causal explanation cannot be discounted. Prevention of adolescent tobacco and cannabis use is a public
health priority but may not lead to improved educational outcomes in young people.

Abstract 1416

AN EVALUATION OF THE IMPACT OF ‘LIFESKILLS’ TRAINING ON ROAD SAFETY, SUBSTANCE USE AND HOSPITAL ATTENDANCE IN ADOLESCENCE

Authors: A Teyhan¹, R Cornish¹, J Macleod¹, A Boyd¹, R Doerner¹, M Sissons Joshi²

¹ALSPAC, University of Bristol, ²Department of Psychology, Social Work & Public Health, Oxford Brookes University

Presenting Author: A Teyhan

Corresponding author email: alison.teyhan@bristol.ac.uk

Background: The Lifeskills safety education centre in Bristol is built as a realistic village and provides safety training to school children aged 10-11 years (Year 6). There is the capacity for 60% of schools in the Lifeskills catchment area (Bath and North East Somerset, Bristol city, North Somerset, and South Gloucestershire) to attend each year.

Aim: To determine whether children who attend Lifeskills engage in safer behaviours, or suffer less accident and injury, in adolescence than a comparable group of children who do not participate in this training.

Methods: The sample are from the Avon Longitudinal Study of Parents and Children (ALSPAC), which is based in the same geographic area as Lifeskills. The National Pupil Database Year 6 school registration details of the ALSPAC participants were linked to the Lifeskills attendance register; where a link was established, the child was classified as having attended Lifeskills, where no link was established the child was classified as not having attended Lifeskills.

At 14-15 years, the participants (n=approximately 3000, varies by outcome) self-reported road safety behaviours and accidents, and perceived health effects and use of alcohol, cannabis, and tobacco. Additional outcomes from linkage to Hospital Episodes Statistics were available for a sub-sample of participants (n=1768): hospital admittance (for an accident-related reason, from age 11-16 years) and A&E attendance (for any reason, from April 2007 to age 16 years). Analyses adjusted for child age and sex, family socioeconomic position, and school and neighbourhood clustering.

Results: Children who attended Lifeskills were more likely to report using pedestrian crossings on their way to school than children who did not attend (59% versus 52%). Lifeskills attendance was unrelated to the ownership of cycle helmets, or the use of cycle helmets, seat belts, or reflective/fluorescent clothing. Use of cycle helmets (37%) and reflective/fluorescent clothing (<4%) on last cycle was low irrespective of Lifeskills attendance. Children who attended Lifeskills were less likely to report recent smoking (14% versus 18%) or occasional cannabis use (8% versus 11%) but attendance was generally unrelated to perceptions of the health impact of substance use. A&E attendance was not associated with Lifeskills attendance. Very few children were admitted to hospital for an accident-related reason, preventing analysis by Lifeskills attendance status.

Conclusions: Lifeskills attendance was associated with some safer behaviours in adolescence. The overall low use of cycle helmets and reflective/fluorescent clothing is evidence of a need for more effective promotion of some safer behaviours.
Abstract 1435  
CANCER RISK FOLLOWING HOSPITAL ADMISSIONS FOR PNEUMONIA  
Authors: TienYu Owen Yang, B J Cairns, G K Reeves, J Green, V Beral  
Nuffield Department of Population Health, University of Oxford  
Presenting Author: TienYu Owen Yang  
Corresponding author email: tienyu.owen.yang@gmail.com

Background: Some investigators have reported excess risks of lung cancer in the short-term and also in the long-term following an episode of pneumonia. The risks of other cancers following pneumonia are less studied. We report here on the short-term and long-term incidence of 12 site-specific cancers following hospital admission for pneumonia in a large cohort of women for whom information on risk factors for cancer, such as smoking, body mass index, and alcohol consumption, were recorded.

Methods: In the prospective Million Women Study, 133,574 incident cancers were identified among 1.3 million women during an average of 13 years of follow-up. Hospital admissions for pneumonia were identified through linkage to routinely collected National Health Service hospital admission data. Cox regression was used to estimate relative risks for each of the 12 cancers 0-1, 2-4, and 5+ years after the first known hospital admission for pneumonia.

Results: For 8 of the 12 cancers investigated, excess risks were found in the two years after hospital admission for pneumonia (after adjustment for age, year of birth, year of recruitment, region, socioeconomic status, adult height, body mass index, smoking status, and alcohol consumption; p<0.05 after correction for multiple comparisons). All cancer risks declined substantially more than 2 years after hospital admission for pneumonia, but the risks of lung cancer and myeloma remained significantly elevated 5 or more years after an admission for pneumonia. Risks estimates for lung cancer were sensitive to adjustment for smoking, and the association observed even five years after hospital admission for pneumonia may well be due to residual confounding. The risk of myeloma was not sensitive to adjustment for smoking or other factors and may reflect sub-clinical disease even five or more years before clinical diagnosis.

Conclusion: Most of the observed excess risk of cancer after an episode of pneumonia may be due to the existence of sub-clinical cancer or to residual confounding by smoking.

Funding: The Million Women Study is funded by Cancer Research UK and Medical Research Council.

Abstract 1445  
RISK OF VENOUS THROMBOEMBOLISM IN WOMEN WITH CANCER  
Authors: Sian Sweetland1, Valerie Beral1, Toral Gathani1, Jane Green1, Cliona Kirwan2, Gillian Reeves1  
1University of Oxford, 2University of Manchester  
Presenting Author: Sian Sweetland  
Corresponding author email: sian.sweetland@ceu.ox.ac.uk

Background: It is known that the incidence of venous thromboembolism (VTE) is increased in people with cancer, but less is known about how the risk varies by cancer type or over time since initial cancer diagnosis. This study aims to describe the incidence of VTE in women after cancer and by type of malignancy.

Methods: 1.3 million UK women were recruited into the Million Women Study in 1996-2001. These women are being followed by record linkage to routinely collected inpatient and day-case NHS hospital admissions data, cancer registrations and death records. The relative risks (RRs) and 95% confidence intervals (CIs) of an inpatient/day-case hospital admission or death record for VTE (ICD10 codes I26, 180-182) were estimated in relation to a variety of malignancies. Analyses using Cox regression were adjusted for age, region of residence, socio-economic group, body mass index, menopausal hormone
therapy, smoking and alcohol consumption.

**Results:** Among 1,198,471 women with no previous VTE or cancer and an average of 11.7 years of follow-up per woman, 108,691 (9%) had a hospital admission for incident cancer and 20,451 (1.7%) had a hospital admission or death from VTE, including 5,153 women whose VTE followed cancer. Overall, women were nine times more likely to have a hospital admission for VTE after cancer than without (or before) any cancer diagnosis (adjusted RR=9.1, 95%CI:8.8-9.4). There was a more than 40-fold increased risk of VTE during the first three months, which decreased to less than 10-fold after one year, but still remained significantly elevated (around 2-fold) many years later. Cancers of the brain and pancreas were associated with the highest risks, at about 70 times the risk without cancer.

**Discussion:** The risk of VTE is significantly increased following a cancer diagnosis, and there was considerable variation over time and by cancer type. We are now investigating the risks associated with different stages of disease and with different cancer treatments.

---

**Abstract 1439**

**BOWEL CANCER SCREENING: LINKED COHORT STUDY OF FACTORS AFFECTING PARTICIPATION AND OUTCOMES IN WOMEN**

**Authors:**
Roger Blanks1, Vicky Enson1, Rupery Alsion2, Anna Brown1, Gillian Reeves1, Valerie Beral1, Julietta Patrick3, Jane Green1

1Cancer Epidemiology Unit, University of Oxford, 2Cancer Epidemiology Unit, Nuffield Department of Population Health, University of Oxford, UK, 3NHS Cancer Screening Programmes, Public Health England

**Presenting Author:** Jane Green

**Corresponding author email:** roger.blanks@ceu.ox.ac.uk

**Background:** In 2006, the National Health Service Bowel Cancer Screening Programme in England (NHSBCSP) began offering routine population-based biennial faecal occult blood testing (FOBt) at ages 60-69. There is, however, limited information on how characteristics of individuals affect participation and outcomes of screening, and we studied this association by linking individual-level NHSBCSP and study data in a large prospective cohort of women.

**Methods:** Electronic linkage of the NHSBCSP and Million Women Study records identified 899,166 women in the study cohort with at least one invitation for screening. NHSBCSP provided information on screening acceptance, FOBt results, screen-detected colorectal cancer and other outcomes. The Million Women Study provided prospectively collected information on personal and lifestyle factors. Multiple regression was used to estimate relative risks (RRs) of factors associated with acceptance and outcomes of screening.

**Results:**
Overall, 70% of women (628,976/899,166) accepted their first invitation for bowel cancer screening, of whom 9133 (1.5%) were FOBt-positive, 743 (0.1%) had screen-detected colorectal cancer and 3056 (0.5%) had screen-detected colorectal adenoma. Acceptance was lower in women from the most than the least deprived tertile, in South Asians and in Blacks than in Whites, in current than in never smokers and in obese than in normal weight women: adjusted RRs (95% confidence interval) for acceptance vs not, 0.90 (0.90-0.90); 0.77 (0.75-79); 0.94 (0.92-0.96); 0.78 (0.77-0.78); and 0.88 (0.88-0.89), respectively; P<0.001 for each. These factors were also associated with an increased risk of being FOBt-positive and of having screen-detected adenoma, but were not strongly associated with the risk of screen-detected colorectal cancer. Relative risks for screen-detected adenoma were 1.22 (1.12-1.34), 2.46 (1.75-3.45), 1.61 (1.05-2.48), 1.53 (1.38-1.68) and 1.77 (1.60-1.95), respectively (P<0.001 for all, except for Blacks vs Whites P=0.03). Use of hormone therapy for menopause was associated with reduced risk of screen-detected adenoma, RR ever vs never use, 0.87 (0.81-0.93), P<0.001 and colorectal cancer, 0.78 (0.68-0.91), P=0.001.

**Interpretation:** Among women in England, socioeconomic and lifestyle factors strongly affect participation in routine bowel cancer screening, risk of being FOBt-positive and risk of having screen-detected colorectal adenoma. However, screen-detected colorectal cancer risk is not strongly related to these factors.
THE IMPACT OF VISION ON EARLY DEVELOPING LITERACY SKILLS

Authors: Alison Bruce, Fairley Lesley
1Bradford Institute for Health Research, 2University of Leeds

Presenting Author: Alison Bruce
Corresponding author email: alison.bruce@bthft.nhs.uk

Purpose: To estimate the impact of presenting visual acuity at age 4-5 years on early developing literacy skills.

Setting: Born in Bradford (BiB) is a multi-ethnic birth cohort of children born in Bradford, UK, between 2007 and 2011, (50% of South Asian origin). Baseline questionnaire (collected during pregnancy) birth data and literacy measures (4-5 years) have been collected and linked to the data from the school vision screening programme along with literacy data from a separate “Starting Schools” study.

Methods: Results of the universal school vision screening programme were prospectively recorded and linked to epidemiological data and literacy measures collected from the BiB cohort over two consecutive school years (2012 – 2014).

A standardised literacy measure; the Woodcock Reading Mastery Tests-Revised: Letter Identification (Letter ID) was used to measure the child’s ability to identify single letters, an essential precursor to reading. The effect of presenting visual acuity (PVA) (best acuity right or left eye, with glasses if worn) on literacy (Letter ID) was assessed using multi-level regression models with individuals nested within schools and adjusting for maternal and child characteristics.

Results: Unadjusted analysis of the BiB children (n=2025) showed that the letter ID score was associated with the level of VA, with the literacy score (Letter ID) reducing by 2.42 points for every 1 line (0.10logMAR) reduction in PVA (CI 2.98 to 1.87) p=<0.001. When adjusted to account for either demographic factors or socio-economic factors the impact of PVA remains significant and continues to remain significant in the multivariable model after all factors are accounted for with the letter ID score reducing by 1.65 (CI 2.17 to 1.13) p=<0.001 for every 1 line (0.10logMAR) reduction in PVA.

Conclusions: In order to explore a causal relationship between vision and literacy the effect of potential confounding socioeconomic factors must be taken into account, this requires comprehensive data collection. The ability to use linked data from the universal vision screening programme along with the epidemiological data from the birth cohort has enabled this study to reliably demonstrate that poor visual acuity in young children is associated with reduced early developing literacy, even after other relevant factors are taken into account.

Financial Disclosure
Dr Alison Bruce is funded by the National Institute for Health Research Post-doctoral fellowship Programme (NIHR PDF).
Abstract 1436

A TOOL TO ENCOURAGE ONGOING PARTICIPATION IN AN ONLINE LONGITUDINAL STUDY

Authors: Hazel M Lockhart-Jones, Rodden M Middleton, Katie Tuite-Dalton, David V Ford, Simon Ellwood-Thompson, Kerina H Jones
College of Medicine, Swansea University

Presenting Author: Hazel M Lockhart-Jones

Corresponding author email: hazel@chi.swan.ac.uk

Background: The UK MS Register gathers data from 3 sources; Clinicians, routine data and directly from people with MS via a web portal. Currently more than 11,000 participants are involved in the longitudinal study and asked to return quarterly to answer Patient Reported outcome Measure (PRoMS) questionnaires. As with all longitudinal studies attempting to keep participants engaged with the research is challenging.

Objective: To categorise the responsiveness of participants answering PRoMS promptly and comprehensively to be monitored by the team and reported back to the participants to encourage them to keep participating.

Methods: The participants were categorised so that activity could be monitored weekly. A SQL stored procedure was deployed calculating user activity from registration to present and then assigning states to users, these were:

- New user - participant who has joined the Register in the last week.
- Registered Only - participant who joined the Register more than a week ago, completed no questionnaires.
- Super Star - participants who filled in their questionnaires including core ones within ten days of their email reminder.
- Very Active - participants who return to fill in their questionnaires within three months of their email reminder.
- Active - participants who return within 6 months to fill in their questionnaires
- Hibernating - participants who have not completed any questionnaires for 6 months.

These categories were then displayed to the participants as a “Star-Bar”. An email system was created to tell users their status and when to return back to the Register. These reminders were automated to try and keep people in the Super Star, Very Active and Active categories.

A web based visual chart was created to show the total number of users by state and the vectors between them eg Active participants to Hibernating. The interactive diagram and Star-Bar uses new JavaScript and HTML5 techniques.

Results: Once categorised and email reminders sent out, a substantial number of previously inactive and potentially disengaged participants responded. This movement was clearly illustrated by the new instrument. The numbers in January 2013 before implementation of Platinum, Very Active and Active have increased measurably from a sum of 2567 to 3972.

Conclusions: The tool is now used daily to monitor the effectiveness of marketing initiatives and new questionnaires to see if they engage with the users to participate. Results are used for management reporting. The automated emails have meant that retention to the study is higher, which results in richer data to link with.
The BB trial [ISRCTN 23019866] assessed the short-term impact for 1600 teenage mothers and their children from a nurse-delivered home visiting programme that has existing evidence of longer term benefits. This NIHR Public Health Research programme funded follow-on study, BB:2-6, will commence the work to assess the programmes’ impact on longer-term benefits for mothers and children, through the linkage of routinely collected data, with a particular emphasis on the programmes’ impact upon child maltreatment.

Using routine data, we will follow-up participants of the trial for an additional four years, until their child is six, to determine the longer-term impact of the FNP home visiting intervention upon objective indicators of child maltreatment when compared to usually provided health and social care services alone. Follow up will be by linked anonymous data abstraction from the Health and Social Care Information Centre (HSCIC) and Department for Education, National Pupil Database (NPD).

Whilst this model of linkage offers the possibility of long term evaluation at a lower cost, what are the associated problems in measuring the proposed outcomes and answering the objectives of the study? This presentation will examine in detail the primary and secondary outcomes of the study and how linking to multiple sources (HSCIC, NPD, and using existing data other sources from the BB trial) can hopefully answer the objectives proposed by the study and what are the potential gaps in knowledge. We will discuss the gains and losses of using these sources versus other available data sources (such as social care data arising from local authorities (LAS)) highlighting the consequences for certain planned analyses (such as Markov Chain Modelling). We will also discuss what is known of the quality of these data sets, the possibility of auditing the data e.g. by sampling local authorities and examining social care data and comparing to NPD data, and what could be achieved by performing this exercise.
Kidney disease is a substantial public health problem with chronic kidney disease (CKD) affecting approximately 10% of the population in the UK, and acute kidney injury (AKI) affecting up to 7% of all hospital admissions. Internationally, and in the UK, there are many challenges to optimising kidney care: raising awareness of the significance of kidney disease; timely detection; adequate surveillance; coordinating care across primary, secondary and social care; involving patients in disease management; and building a robust evidence base for effective treatment.

National Renal Registries provide internationally recognised exemplars for longitudinal monitoring of outcomes in CKD patients requiring renal replacement therapy (i.e. dialysis and transplantation). With the UK Renal Data Collaboration (UKRDC), renal services in the UK have a unique health data infrastructure available. In addition, PatientView, with over 5,000 active users, gives people with CKD web-access to their renal electronic health record.

Optimising the use of available health data and information technology provides a growing opportunity to improve our understanding of kidney disease and effective treatments, as well as health care services and outcomes for kidney patients. In this context, the renal registries and the Farr centres have a strong track record in renal research. Examples include novel methods to link biochemistry with hospital episode and renal registry data to develop tools to support CKD and AKI care, risk prediction and planning (Farr@Aberdeen); external validation of prediction models for CKD onset using primary care data (Farr@HeRC); and clinical trials evaluating the effect of vitamin D on dialysis patients' outcomes and effectiveness of electronic alerts in CKD and AKI using registry data (Cambridge CTU, UK Renal Registry). Furthermore, the Farr Institute has facilitated cross-centre working with Farr@London and Farr@Scotland working together on AKI to test algorithms for AKI definition in NHS laboratory data and understand the association between AKI and prescribing.

Further strengthening the collaboration among Farr centres, and between the Farr and national renal data partners will capitalise on available data, expertise, and willingness to bring health informatics and kidney research in the UK to the next level. Kidney Disease@Farr is the initiative that brings together a multidisciplinary group of health informaticians, computer scientists, statisticians and clinical researchers from across the four Farr centres and the renal data partners to develop and share research ideas, foster collaborations and utilise the Farr infrastructure to improve kidney patient care and outcomes through health informatics research.
Abstract 1773 DOES THE PAST PREDICT THE FUTURE? IMPLICATIONS OF PRIOR RENAL FUNCTION ON FUTURE RENAL OUTCOMES

Authors: A Marks, L Robertson, C Black
University of Aberdeen

Presenting Author: A Marks

Corresponding author email: a.marks@abdn.ac.uk

Introduction: Chronic kidney disease (CKD) is common (~10% population) and has serious implications (high mortality and morbidity, including the need for costly renal replacement therapy (RRT)). Current predictions of future population-level mortality and RRT service requirements are based on the population prevalence of single measurements of excretory renal function and proteinuria. Outcome prediction models are also currently based on single measurements. Traditionally, clinical estimates of when a given patient will need to initiate RRT were based on decline in renal function over time. The aim of treatment for CKD is also to slow and halt that decline. Would future predictions of population-level RRT service requirements be better based on a population’s past renal function decline rather than a single measure of current function?

Methods: Data-linkage of a cohort of patients with advanced CKD has been used to investigate and predict outcomes in CKD. Data-linkage of routine healthcare data from biochemistry laboratories (1999 to 2009); health care admission records (diagnoses); specialist registries (Scottish renal registry for RRT) and mortality records for a cohort with CKD in 2003 has allowed exploration of the effects of baseline comorbidity and renal function on healthcare events. Those who have progression (15ml drop in renal function) have been described. The performance of models to predict outcome have also been described. Combining these measures of progression with prediction models in a larger cohort with a wider range of CKD will investigate whether this is a useful undertaking for those involved in the planning of care for renal patients.

Results: Published prediction models performed well with single measures of renal function (c-statistic 0.936). However in 18687 with CKD, 3278 were falsely predicted to start RRT within 5 years. ‘Progression’ was more common in males, but stable across CKD stages. Results exploring the addition of measures of prior progression to prediction models will be presented.

Discussion: As knowledge of the epidemiology of CKD increases worldwide, estimates of future RRT needs and estimates of deaths due to lack of RRT facilities are being made in both the developed and developing world. However these predictions are currently based on information gleaned from studies using single measures of renal function. Research into the effect of prior decline of renal function should refine these estimates.
almost all interventional trials to avoid confounding, and trial results can therefore not easily be extrapolated to kidney failure patients. Carrying out trials in this population is technically challenging and often prohibitively expensive. Novel approaches to clinical trials are urgently needed to advance the treatment of patients with kidney failure.

**Intervention:** Vitamin D deficiency is highly prevalent in dialysis patients and is associated with increased CV and all-cause mortality in the general population. Its supplementation in dialysis patients is safe, and may reduce mortality.

Native Vitamin D (Cholecalciferol) is a low risk intervention and therefore provides the ideal test bed for novel trials methodologies.

**Design:** The SIMPLIFIED trial (Survival Improvement with Cholecalciferol In patients with ESRD on Dialysis) is a large (n=4,200) UK-only pragmatic prospective randomised open-label blinded endpoint (PROBE) trial of cholecalciferol versus standard care in patients on dialysis. The primary endpoint is all-cause mortality, and secondary endpoints include hospitalisation-requiring CV events, infection, and cancer incidence.

**Follow-up:** SIMPLIFIED will have no dedicated point of care trial visits, but will capture all data using routine data sources (ONS, HES, UK Renal Registry, PatientView). The UK Renal Registry (UKRR) captures detailed data on all UK dialysis patients, including blood results (within 24h of collection), and undertakes linkage with HES and ONS. The UKRR also links with the patient portal PatientView, which allows renal patients to view and update their data directly.

Within SIMPLIFIED, patients will consent to their data being directly accessible in real time to the trial team via PatientView. A bespoke trial database will incorporate regular data tranches from the UKRR, HES and ONS. This approach could serve as the basis for future trial design.

**Conclusions:** SIMPLIFIED will be the first UK nephrology trial to use routine data exclusively for trial follow-up, and should contribute to the development of expertise and infrastructure to carry out affordable clinical trials within the NHS.

**Abstract 1603**

**REUSING ROUTINELY COLLECTED DATA TO ASSESS PERFORMANCE OF CHRONIC KIDNEY DISEASE PREDICTION MODELS IN THE UK PRIMARY CARE: AN EXTERNAL VALIDATION STUDY**

**Authors:** Paolo Fraccaro¹, Sabine Van der veer², Donal O’Donoghue³, Benjamin Brown², Mattia Prosperi², Niels Peek¹, Iain Buchan¹

¹HeRC & GM Primary Care Patient Safety Translational Research Centre, The University of Manchester, ²HeRC, The University of Manchester, ³Salford Royal NHS Foundation Trust

**Presenting Author:** Paolo Fraccaro

**Corresponding author email:** paolo.fraccaro@postgrad.manchester.ac.uk

**Background:** Adequate identification of patients with an increased risk to develop chronic kidney disease (CKD) may allow preventive actions to delay disease onset. CKD is mainly detected by using a creatinine-based estimation of renal function, the estimated glomerular filtration rate (eGFR), which in the UK is automatically calculated every time creatinine is measured since 2006. However, eGFR tests are difficult to interpret and act on. Several CKD prediction models that could support eGFR value interpretation have been developed, but their performance outside the context in which they were developed is largely unknown. This study aimed to assess the performance of existing CKD prediction models within a UK primary care population.

**Methods:** From an anonymized primary care electronic health record database used in Salford (UK), we included adult patients with at least one eGFR measurement in 2009 and a complete follow-up until 2014. We stratified our analysis for presence of CKD risk factors, and defined CKD as eGFR<60ml/min/1.73m² for three months or longer. For each model, we evaluated calibration as the difference between predicted and observed risk, and discrimination with the area under receiver
operating characteristic curve (AUC). Performance of risk scores derived from included models were assessed based on AUC, positive predictive value (PPV) and sensitivity.

**Results:** We evaluated nine models and seven risk scores selected from two recent systematic reviews. We included 24,591 with and 10,217 patients without CKD risk factors; the number of CKD cases developed in the study period was 13.3% and 4.7%, respectively. The majority of the models needed recalibration. The included model that used eGFR as predictor had the best AUC performance with values of 0.81 and 0.85 for the cohort with and without CKD risk factors, respectively. Performance of models that did not include eGFR ranged between 0.68-0.75 for patients with CKD risk factors and 0.76-0.80 for those without. Risk scores’ performance was similar to the models they were derived from. Sensitivities ranged between 0.66-0.79 for patients with CKD risk factors and 0.73-0.83 for patients without. The model that included eGFR as predictor had best PPVs, which overall varied between 0.15-0.31 for patients with CKD risk factors and 0.1-0.16 for patients without.

**Conclusions:** Although the included CKD prediction models scored moderate discriminative performance, they did not seem adoptable in practice (i.e. poor PPV and lack of calibration) when focusing at patients with and without CKD risk factors in UK primary care.

---

**Abstract 1527**

**DIAGNOSING AKI IN ROUTINE HEALTHCARE – WHY THE DEFINITION MATTERS**

**Authors:** Simon Sawhney¹, Angharad Marks¹, Nick Fluck², Gordon Prescott¹, Corri Black¹

¹University of Aberdeen, ²NHS Grampian

**Presenting Author:** Simon Sawhney

**Corresponding author email:** simon.sawhney@abdn.ac.uk

**Background:** Acute kidney injury (AKI) is a serious condition complicating 1 in 7 hospital admissions and often life-threatening. It coexists with severe illness and happens in any clinical setting so all clinicians must be aware of AKI. True diagnosis of AKI is clinical, incorporating rapid changes in blood tests (serum creatinine) and urine output.

Unfortunately, recognition and care in AKI is often delayed leading to preventable harm. Initiatives to improve its timely diagnosis include automated AKI alerts to clinicians when serum creatinine changes suggest AKI.

World Health Organisation International Classification of Disease (ICD-10) codes for diagnosis are also routinely collected in the UK for use in healthcare planning. They rely on both clinical recognition and subsequent recording on hospital discharge.

The incidence of AKI is believed to be rising. This has implications for healthcare planning, but current evidence is based on ICD-10 recording that may reflect increased recognition rather than true changes.

This presentation will explore the implications of these three methods commonly used to define AKI in research - clinical diagnosis, automated creatinine algorithms, and ICD-10 coding.

**Methods:** We used the NHS England automated AKI detection algorithm to identify AKI using all serum creatinine results in Grampian, 2003-2008 (adult population ~450,000). We compared the estimated annual incidences of AKI using automated detection with the mid-year Grampian populations and blood testing practices, 2003-2008.

We collected ICD-10 AKI coding and compared this with automated AKI detection. Finally, we used a subgroup of 4464 patients with renal impairment characterised by nephrologist review of case-notes to provide a comparison with the clinical diagnosis.

**Results:** From 2003 to 2008 the adult population of Grampian rose 6% and serum creatinine testing rose 26%. By automated AKI detection, the annual incidence of AKI also rose 14% from 5565 to 6359, but was unchanged after the rise in population and blood testing were accounted for. In contrast, ICD-10 AKI recognition rose almost twofold.

ICD-10 AKI was a minority of all automated AKI detection cases (6%), but a greater proportion of the
most severe (stage 3) automated detection cases (28%). Against case-note review, automated detection did identify patients with a definite clinical diagnosis, but also included patients without AKI (20%) and with an uncertain diagnosis (29%).

Conclusions: The incidence of AKI is increasing due to increased testing and recognition. Different definitions select a different AKI case-mix and this requires careful thought when planning and interpreting AKI clinical research.

Abstract 1755  USING DATA LINKAGE TO EVALUATE RATES OF ACUTE KIDNEY INJURY AND CLOSTRIDIUM DIFFICILE INFECTION FOLLOWING A CHANGE IN ORTHOPAEDIC PROPHYLACTIC ANTIBIOTIC POLICY

Authors: Heather Kenny1, Andrea Patton2, Gwen Bayne3, Charis Mariwck4, Samira Bell4
1NHS TAYSIDE, 2Scottish Antimicrobial Prescribing Group and University of Dundee, 3Scottish Antimicrobial Prescribing Group, 4NHS Tayside and University of Dundee

Presenting Author: Andrea Patton

Corresponding author email: hkenny1@nhs.net

Background: Prophylactic antibiotics are used in orthopaedic surgery to reduce rates of surgical site infections. Following a change in national antibiotic policy within NHS Scotland aimed at reducing Clostridium difficile infections (CDI), there was evidence that a regimen of flucloxacillin and gentamicin in orthopaedic patients was associated with a significant increase in acute kidney injury (AKI). As a result, the national antibiotic policy recommendation for orthopaedic surgical prophylaxis in Scotland, was changed from flucloxacillin and gentamicin to co-amoxiclav. Patients undergoing a neck of femur (NOF) repair received co-amoxiclav throughout the study period. This study aimed to assess whether this change in orthopaedic antibiotic prophylaxis policy was associated with changes in the rates of post-operative AKI and CDI in adult patients undergoing orthopaedic surgery.

Methods: An observational cohort study was performed, assessing the rates of AKI and CDI among patients who had undergone orthopaedic surgery. Data were linked through the Health Informatics Centre, University of Dundee between the following datasets: Scottish Morbidity Record of hospital admissions, OPCS-coded procedures, laboratory results, medicines dispensed by community pharmacies, General Register Office death database and Scottish Care Information – Diabetes Collaboration. 14,563 adults aged ≥18 years who resided, or died, in the NHS Tayside region and who underwent an orthopaedic surgical procedure between October 1, 2008, and December 31, 2013 were included in the analysis. Interrupted time series (ITS) was used to assess rates of AKI over the time period of October 2008 to May 2012 when flucloxacillin and gentamicin was used compared with June 2012 to 2013, when the policy was changed to co-amoxiclav. Cases were split into two groups, NOF operation group and other orthopaedic surgeries group.

Results: ITS analyses showed that 18 months following the change in policy there was a statistically significantly relative decrease in rates of all AKI by -63% (95% CI -77%, -49%) in the other orthopaedic surgeries group. The incident rate ratio (95% CI) of CDI was 0.29 (0.09 to 0.96).

Conclusion: The change in orthopaedic antibiotic prophylaxis policy to co-amoxiclav led to reduced rates of AKI compared to the previous policy of flucloxacillin and gentamicin and was not associated with an increased rate of CDI.
Abstract 1723

ROUTINE COLLECTION AND USE OF PATIENT-REPORTED INFORMATION IN CHRONIC KIDNEY DISEASE: CURRENT STATUS AND FUTURE CHALLENGES

Authors: Sabine N van der Veer1, Fergus J Caskey2
1Institute of Population Health, 2UK Renal Registry, Bristol, UK

Presenting Author: Sabine N van der Veer

Corresponding author email: sabine.vanderveer@manchester.ac.uk

Background: Chronic kidney disease (CKD) is a public health problem, with a prevalence in the UK of 6.5%. It also impacts on the individual: CKD patients often have a lower quality of life than those with cancers and other long-term conditions. Many become chronically tired, are depressed, and suffer pain; these symptoms frequently go unrecognised. Optimal use of information on how kidney patients experience these consequences of the condition on their life and health might improve symptom management, quality of life, self-efficacy, and satisfaction with services.

Current status: As in other long-term conditions, routine collection and thus use of patient-reported experiences in the context of CKD is still under-developed. Therefore, NHS England has initiated the Transforming Participation in Chronic Kidney Disease (TP-CKD) programme, involving patients, clinicians and researchers. TP-CKD aims to empower kidney patients to manage and make decisions about their care, and to enhance their quality of life.

The TP-CKD Measurement workstream focuses on exploring acceptable and feasible methods to routinely capture patient-reported measures, such as symptom burden, and patient activation. Data collection will start later this year, building on the national IT infrastructure for renal services. This includes the UK Renal Registry, which automatically extracts data from renal units’ IT systems, with 100% coverage of dialysis and transplant patients in the UK; and PatientView, a portal giving kidney patients online access to their renal electronic health record.

Future challenges: Development of successful methods for collecting patient-reported experiences in the context of CKD and other long-term conditions faces several challenges:

1. Minimal administration burden: Since people with long-term conditions will be asked repeatedly to report experiences throughout the course of their disease, minimising the administration burden is essential. Although TP-CKD will mainly employ traditional paper-based questionnaires, ubiquitous technology provides an opportunity to use digital, more flexible alternatives that are tailored to individual patients’ responses.

2. Equally involving all patients: Instead of a ‘one size fits all’ approach, TP-CKD will explore data collection needs of different patient groups, to also enable participation of those with, for example, lower literacy levels. This will be done by comparing characteristics of responders and non-responders, and investigating reasons for non-response.

Optimal use of collected data: TP-CKD will make the collected patient-reported data available in local renal IT systems to allow it to inform treatment decisions. Also, patients will be able to monitor their reported experiences in PatientView with the aim to improve self-efficacy and support self-management skills.
Abstract 1774  
ROUENTLY COLLECTED CLINICAL DATA SUPPORTING CLINICAL TRIALS: 20 YEARS OF WOSCOPS

Authors:  
Colin McCowan, Murray Heather, Chris Packard, Sharon Kean, Ford Ian  
University of Glasgow

Presenting Author: Colin McCowan

Corresponding author email: colin.mccowan@glasgow.ac.uk

Aims: The benefits of record linkage to provide long-term follow up of trial participants have been known for some time and are becoming more routinely used in the prospective conduct of trials. The West of Scotland Coronary Prevention Study (WOSCOPS) was one of the first studies to illustrate this methodology. We evaluated benefits over 20 years of follow-up from 5 years of pravastatin (40mg/d) vs. placebo use in men with hypercholesterolemia with no history of myocardial infarction (MI) using record linkage.

Methods: In WOSCOPS, regular applications for hospital admissions, cancer registrations and deaths to allow follow-up using a national electronic record linkage have resulted in follow-up over a 20-year period. Effects of randomized treatment assignment on events of interest including cardiovascular events, cardiovascular mortality, other cause mortality, other cause hospitalisations and cancers were analyzed using survival analysis models.

Results: There was a significant difference in cardiovascular mortality over the 20 year follow-up in favour of the pravastatin group (12% vs 15% HR=0.79, 95%CI 0.69-0.90, p=0.0004) and also in number of cardiovascular events (1838 vs 2286 p<0.0001). There was no difference in other cause mortality, other cause hospitalisations or cancer incidence.

Conclusions: Computerized record linkage to routine health records has provided new insights to the effects of pravastatin long after trial cessation. Scotland, and the rest of the UK, are ideally positioned to maximise findings from long-term follow-up from other clinical trials.

Abstract 1631  
USING LINKED HEALTHCARE DATA TO CREATE A CLUSTER RANDOMISED CONTROLLED TRIAL: THE RAPID TRIAL (REDUCING ANTIBIOTIC PRESCRIBING IN DENTISTRY)

Authors:  
Paula Elouafkaoui1, Jan E Clarkson1, Linda Young2, Rumana Newlands3, Eilidh Duncan3, Craig R Ramsay3  
1NHS Education for Scotland/University of Dundee, 2NHS Education for Scotland, 3HSRU, University of Aberdeen

Presenting Author: Paula Elouafkaoui

Corresponding author email: paula.elouafkaoui@nes.scot.nhs.uk

Background: Antimicrobial resistance is a serious threat to global public health and patient safety. Inappropriate use of antibiotics is a major contributor to the spread of antimicrobial resistance. In Scotland dentists prescribe 9% of all antibiotics in primary care. Despite the widespread usage of the Scottish Dental Clinical Effectiveness Programme’s (SDCEP) Drug Prescribing for Dentistry guidance, the total number of antibiotics prescribed by dentists increased steadily up to 2013.

Methods: The RAPID partial factorial cluster randomised controlled trial used national, routinely
collected dental prescribing and treatment data to compare the effectiveness of individualised audit and feedback (A&F) strategies for the translation into practice of SDCEP recommendations on antibiotic prescribing in primary care dentistry in Scotland. Data was used in five aspects of the trial:

1) **Identify participants**: all currently practicing General Dental Practitioners (GDPs) in Scotland.
2) **Apply inclusion/exclusion criteria**: based on dental practice contract status and a minimum level of recent NHS treatment provision.
3) **Carry out stratified randomisation**: all eligible dental practices in Scotland were simultaneously randomised at baseline to control or to one of two audit and feedback interventions. Randomisation was stratified by single-handed/multi-handed practices. Intervention practices were further randomised using a factorial design.
4) **Generate the trial intervention**: individualised graphical feedback on antibiotic prescribing. The initial feedback report contained 14 months retrospective antibiotic prescribing data.
5) **Analyse trial outcomes**: the primary outcome is the total antibiotic prescribing per 100 courses of treatment over the year following the delivery of the baseline intervention.

**Results**: The primary analysis revealed a 6% reduction in the rate of antibiotic prescribing by dentists exposed to the A&F intervention, compared to the control group. Subgroup analyses demonstrated that the prescribing rate for dentists who received the graphical feedback plus a written behaviour change intervention was lower (6%) than dentists who received the graphical feedback only. For dentists who were provided with a health board comparator as part of their feedback, a 4% reduction in the prescribing rate was observed, compared to dentists who did not receive the comparator.

**Discussion**: RAPiD demonstrates that linked administrative datasets have the potential to be used efficiently and effectively across all stages of a randomised trial. This is a relatively straightforward public health and patient safety intervention that can be delivered at low cost and at scale. It has the potential to help the dental profession, and other healthcare professions, address the increasing challenge of antimicrobial resistance.

---

**Abstract 1636**

**THE DESIGN OF POINT-OF-CARE TRIALS THAT USE ROUTINELY COLLECTED DATA: EVALUATION OF THE TRIALS WITHIN COHORTS (TwiCs) DESIGN**

**Authors:**
Tjeerd van Staa\(^1,2\), Alexander Patt\(^1,2\), Jane Candlish\(^1,2\), Matthew Sperrin\(^1,2\)

\(^1\)Institute of Population Health, University of Manchester \(^2\)Farr Institute @ HeRC

**Presenting Author:** Alexander Patt

**Corresponding author email:** alex.pate@manchester.ac.uk

There is a major need for randomised trials that are conducted at the point of care and that evaluate interventions in usual care settings (i.e., pragmatic trials that test effectiveness rather than efficacy in strictly controlled and selective settings). The design of these trials could be simplified. An example is the Trials within Cohorts (TwiCs) design in which a random sample within a well-defined cohort is offered the new intervention while the remaining cohort follows usual care and is not formally recruited into the trial. While this design could be highly pragmatic, the effects of non-uptake of the novel intervention are unclear. We are undertaking a simulation study to test the validity of the TwiCs design. The primary aim of this study is to identify the presence of bias in the estimated intervention effect induced in the TwiCs trial design. We will attempt to account for this bias through the use of instrumental variable (IV) analysis and other statistical methods. The secondary aim is to accommodate for the effect of competing risks (such as mortality) and secular trends, which may be present. The aims will be addressed in a variety of scenarios which will include varying degrees of correlation between a patient’s risk of an event and i) a patient’s probability of refusing the intervention treatment and ii) probability of a clinician refusing to give the novel intervention and iii) a patient’s risk of mortality/censoring. The TwiCs design will be tested both as a standard randomised trial and as a cluster trial (randomising clinics). In the cluster trial, the number and size of clusters will also be varied. The setting for the TwiCs trial will be usual clinical practice. The analysis will be a survival analysis on the...
time until a clinical event and so will be generalizable to all survival data. At the conclusion of this study, we aim to have identified: (i) a set of scenarios where the TwiCs are viable designs for point-of-care trials, (ii) a set of scenarios in which alternate statistical methods such as IV must be applied to accurately estimate the intervention effect (iii) a set of scenarios in which the TwiCs trial designs are not viable for use. We will present the results at the symposium.

Abstract 1639

RESEARCH ETHICS COMMITTEE DECISION MAKING IN RELATION TO NOVEL METHODOLOGIES FOR EFFICIENT TRIAL DESIGN: A COMPARATIVE STUDY

Authors: Chris Gale1, Matthew J Hyde1, Helen Robberts2, Mark Turner3, Amanda Forster4, Jon Dorling5, Zoe Chivers4, Tjeerd van Staa6, Neena Modi7

1 Imperial College London, 2 Parent of preterm twins, 3 University of Liverpool, 4 James Cook Hospital NHS Trust, 5 University of Nottingham, 6 Bliss: for babies born too soon, too small, too sick, 7 University of Manchester, Farr HeRC

Presenting Author: Chris Gale

Corresponding author email: christopher.gale@imperial.ac.uk

Background: Novel methodological approaches to reduce waste in research by facilitating efficient, large, simple trials include the use of routinely recorded Electronic Patient Record (EPR) or disease registry data in point-of-care trials, streamlined patient information sheets that include explicit mention of potential inclusion benefit from trial participation and opt-out consent for comparative effectiveness research. The acceptability of these approaches to UK Research Ethics Committees (REC) is unclear.

Aim: To determine the acceptability to UK REC of methodologies to facilitate efficient, large, simple clinical trials.

Methods: An identical application for a planned neonatal comparative effectiveness trial (WHEAT), proposing to compare two existing, widely used practices in relation to blood transfusion in preterm neonates, was submitted to 12 UK REC. The trial design was developed in close collaboration with parents of preterm babies. The Health Research Agency had informed all UK REC of this exercise. Included REC had agreed to take part, but were unaware of when this would be or the nature of the application. The planned WHEAT trial involved a point-of-care design (using existing EPR for patient identification, randomisation and data acquisition), a streamlined parent information sheet that mentioned the possibility of inclusion benefit through participation, and a consent process involving enrollment as the default unless a parent opted-out of participation for their infant. Researchers (CG, MJH) attended 6 REC meetings in person and were available by phone for a further six. Identical responses were provided to similar points of clarification or justification requested by a REC.

Results: A favourable opinion was granted by 9 REC all of which considered the provision of a statement about potential inclusion benefit in the parent information sheet to be acceptable. Three REC provided an unfavourable opinion because they considered opt-out consent invalid. No REC raised concerns about the proposed point-of-care design using EPR data. The concerns of one REC about a shortened parent information sheet were allayed after demonstrating parent involvement in the design.

Conclusions: There is variation in UK REC decision-making. While UK REC hold discrepant views on the validity of opt-out consent in randomised comparative effectiveness research, the majority considered this approach acceptable. Both point-of-care trial methodology using EPR data, and informing parents of the possibility of inclusion benefit through participation, are considered valid and acceptable. We thank Dr Hugh Davis and Dr Joan Kirkbride from the Health Research Authority for their support.
Abstract 1663  
**SUPPORTING CLINICAL TRIAL DATA CURATION AND INTEGRATION WITH TABLE MINING**

**Authors:**  
Nikola Milosevic¹, G Nenadic², C Gregson³, R Hernandez³  
¹University of Manchester, ²School of Computer Science, University of Manchester; HeRC, Farr Institute of Health Informatics Research, ³AstraZeneca

**Presenting Author:** Nikola Milosevic  
**Corresponding author email:** nikola.milosevic@manchester.ac.uk

PubMed contains nearly 800,000 clinical trial citations, which report detailed trial planning, execution and results, including descriptions of study arms, demographic data, inclusion/exclusion criteria, protocols that have been followed, specific outcomes etc. Although clinical trial reports are relatively structured and deliberately written in a typical semi-structured format in order to standardise and improve reporting, automated extraction and curation of structured information is challenging given that data is presented either in the form of unstructured natural language text or in semi-structured tables. For example, demographic characteristics of trial arms and inclusion/exclusion criteria, as well as interactions between substances, drug side effects, etc. are often presented in tables. So far, medical text mining has mostly focused on extracting information from the main body of text with some success. Processing of information from tables is often limited to textual captions, whereas data presented in tables are typically ignored in large-scale automated processing. Here we report on a methodology developed to support semi-automated data curation and integration from clinical trial reports that relies on processing both the main text and tables.

Tables are processed in four steps. (1) Structural processing aims to establish a table’s layout and relationships between cells. We differentiate between three types of table (one, two and multi-dimensional) and use a number of heuristics to link each cell to the corresponding header(s) and stub(s). (2) Syntactic processing identifies values and data types within the cells (e.g. numerical, ranges, text). (3) Based on the results of that step and the surrounding text, we use a machine-learning approach to establish the (main) reason of table use (e.g. settings, findings, support-data). (4) Finally, semantic mining aims to link the cell values to semantic resources (e.g. ontologies) and to represent the information as linked data using semantics technology standards.

In the initial evaluation of the methodology on a case study with the extraction of values of body mass index and/or weight of patients involved in clinical trials, we achieved a F-measure of 85% for body mass index extraction and 58% for participant weight extraction. We note that one of the main challenges is identification of trial participant groups (F-measure 71.3%), which requires the integration of the results of mining both the main body and tables within a clinical trial report.

Abstract 1514  
**TEXT MESSAGING REMINDERS FOR INFLUENZA VACCINE IN PRIMARY CARE (TXT4FLUJAB)**

**Authors:** Emily Herrett¹, T van Staa², T Chadborn³, E Williamson¹, L Smeeth¹  
¹London School of Hygiene and Tropical Medicine, ²University of Manchester, ³Public Health England

**Presenting Author:** Emily Herrett  
**Corresponding author email:** emily.herrett@lshtm.ac.uk

**Background:** Primary care practices in the UK fall short of government targets for influenza vaccination in patients aged under 65 with chronic conditions. A third of practices use text messaging to remind patients about vaccination, despite lack of evidence for its effect.

**Aims:** (1) to implement a cluster randomised trial testing the effectiveness of a text messaging reminder
to increase influenza vaccine uptake in patients aged 18–64 with chronic conditions, compared with standard care, and (2) to develop methodology for conducting cluster randomised trials using routine electronic health records.

**Methods:** This cluster randomised trial recruited and randomised 156 English primary care practices to either standard care or a text messaging influenza vaccine reminder to eligible patients. Flu vaccine uptake was ascertained using routinely collected, anonymised electronic patient records. An intention to treat analysis was performed, using a t-test, to compare uptake between the intervention and control groups. Feedback regarding the intervention was ascertained using a questionnaire to practice staff and a subgroup of patients in the intervention group.

**Results:** In an intention to treat analysis of provisional data from 146 practices, mean practice-level vaccine uptake was higher in the text messaging group (53.5%) than in the intervention group (50.9%), representing a 5.1% relative increase. However, a t-test showed no evidence for a difference between those in the intervention and standard care groups (P=0.75). Full vaccine uptake data were collected from 155/156 practices and more detailed results will be presented. Practice and patient feedback regarding the text messaging was positive: 86% of practice staff felt that text messaging was worthwhile and only 4% experienced difficulties sending the message. Fifty two percent of patients reported that the message encouraged them to receive vaccination and just 5% objected to the message.

**Conclusions:** This trial has demonstrated the feasibility of using electronic health records for outcome data collection in cluster randomised trials. Uptake was higher among practices that were randomised to the text messaging group than those randomised to standard care but the difference was small and could have been due to chance. We found no evidence that simple, single text message reminders produce a substantial increase in influenza vaccine uptake among those aged 18–64 in at-risk groups. Given the low cost of this intervention, future research to evaluate the effect of repeated and/or personalised messages is suggested.

---

**SESSION E4 – RESEARCH COHORTS.**

**CHAIR: PROF RONAN LYONS, SWANSEA UNIVERSITY & THE FARR INSTITUTE**

Abstract 1660  
**DIABETES CASE CONFIRMATION IN THE UK BIOBANK BY DATA LINKAGE**

**Authors:**  
Mark Atkinson¹, Sophie Eastwood², Sinead Brophy¹, Nishi Chaturvedi²  
¹CIPHER, College of Medicine, Swansea University, ²Institute of Cardiovascular Science, University College London

**Presenting Author:** Mark Atkinson

**Corresponding author email:** m.atkinson@swansea.ac.uk

**Introduction:** UK Biobank (UKB) is a major national health resource which aims to improve the prevention, diagnosis and treatment of a range of illnesses. Between 2006 and 2010, 500,000 people between the ages of 40 and 69 were recruited in England, Scotland and Wales. At baseline assessment, physical measurements were made and tissue samples taken for genotyping and biomarker measurement. Participants were questioned extensively about lifestyle and health.

To do effective genetic or other studies of particular diseases, confidence in the identification of cases is required. Studies of prevalent diabetes in this cohort need to distinguish between type 1 (T1) and type 2 (T2) diabetes. Participants with baseline diabetes answered questions about date of diagnosis, medications and complications. However, inconsistencies exist in the raw baseline self-report data.

We developed an algorithm for identifying and adjudicating case status and type of diabetes, using the self-report data alone. We validated this algorithm using linked primary care data.

**Methods:** Algorithms were developed based on the UKB data. The algorithm identified; Diabetes unlikely (having neither diagnoses nor medications), Possible T2 diabetes (current non-metformin orals,
or age over 30), Possible T1 diabetes (self-report of current insulin or insulin within a year of diagnosis). A second flowchart further classified all T1 cases by assigning probable T1 status to those with more evidence of medication. A third flowchart examined possible T2 cases and reassigned some as probable T2 and others as possible T1 and diabetes unlikely. UKB participants living in Wales were linked to the SAIL databank and GP records were examined for diabetes diagnostic and medication codes, and results compared to those from algorithms applied to self-report data.

**Results:** Of the 26809 Welsh UKB participants, 1431 (5.3%) had adjudicated diabetes found by the algorithm. GP data were available for 760 (53%) of these. Of the 760, 644 (85%) had diagnostic codes for diabetes in the GP data at UKB baseline. There were 60 people who had GP diagnostic codes for diabetes at baseline but who had no self-report of diabetes.

**Conclusion:** In the absence of primary care data linked to all or most UKB participants, the present algorithm for detecting diabetes using self-report data alone has a positive predictive value of 85%. The validation of the further classification of classification by the algorithm into T1 or T2 is currently being done.

---

**Abstract 1579**

DETERMINING THE PROGRESSION AND RISK FACTORS OF TYPE 2 DIABETES FOLLOWING GESTATIONAL DIABETES IN THE BIB COHORT: A LINKED DATABASE STUDY

**Authors:** Muhammad Faisal¹, Brian Kelly², Anne Graham³, Dan Mason², John Wright², Mohammed Amin Mohammed⁴

¹University of Bradford, ²Bradford Institute for Health Research, Bradford Teaching Hospitals NHS Foundation Trust, Bradford, UK, ³Bradford School of Medical Sciences, University of Bradford, Bradford, UK, ⁴Faculty of Health Studies, University of Bradford, Bradford, UK

**Presenting Author:** Muhammad Faisal

**Corresponding author email:** M.A.Mohammed5@bradford.ac.uk

**Background:** Diabetes is a major public health issue throughout the world but is a particular issue in the Bradford population as the risk of developing Type 2 diabetes (T2DM) is higher in individuals from Pakistani and Bangladeshi backgrounds. One pathway to diabetes for women is via gestational diabetes, which usually resolves but can convert into T2DM. The ‘conversion rate’ from gestational diabetes to T2DM is not well known. Data linkage between the Born in Bradford birth cohort data and electronic healthcare records provides an opportunity to investigate this pathway in different ethnic groups while controlling for potential confounders.

**Methodology:** Mothers recruited around 26-28 weeks gestation to the Born in Bradford birth cohort (N = 12450) self-reported ethnicity at baseline. An indicator of gestational diabetes was derived from glucose tolerance test (GTT) results in the maternity hospital record. Of the 10159 women with complete ethnicity and GTT data, 9909 were matched to their SystmOne primary care records on the basis of NHS number, date of birth, sex and surname. We used logistic regression to determine the association of ethnicity with frank diabetes following gestational diabetes.

**Results:** Of the 9909 BiB mothers, 811 (8.2%) had gestational diabetes and of these 112 (13.8%) had T2DM. The median time to T2DM was 2.41 years (IQR = 3.75). The risk (odds ratios (OR)) of T2DM following gestational diabetes by ethnicity, after controlling for age, body mass index (BMI) and smoking status, are as follows: Pakistani 7.11 (95% CI 3.54 - 14.30), Indian 9.96 (95% CI 3.65 - 27.21), and Bangladeshi 23.24 (95% CI 9.10 - 59.36) relative to white British mothers.

**Conclusion:** Data linkage between cohort research records and primary care records provided a novel and cost-efficient approach to identify possible risk factors of T2DM following gestational diabetes. This conversion rate varies significantly with ethnicity even after adjusting for, age, BMI and smoking during pregnancy.
Abstract 1756  

**BIRTH WEIGHT AND COGNITIVE FUNCTIONS OVER THE LIFE-COURSE IN THE GENERATION SCOTLAND (GS) SUBSAMPLE OF THE ABERDEEN CHILDREN OF THE 1950s STUDY (ACONF).**

**Authors:**  
Dorota Chapko¹, Corri Black¹, Roger T Staff², Alison D Murray¹  
¹University of Aberdeen, ²NHS Grampian, Aberdeen Royal Infirmary

**Presenting Author:** Dorota Chapko  
**Corresponding author**  
email: dorota@abdn.ac.uk

**Introduction:** There is a consensus that birthweight is associated with cognitive abilities in childhood but it is unclear whether this association persists into adulthood. In this study we examine the relationship between birthweight and cognitive functions from childhood to mid-life based on the repeated measures (at ages 7, 9, 11 and 51-61 years) available through the linkage performed between the Aberdeen Children of the 1950s study (ACONF) and the Generation Scotland: Scottish Family Health Study (GS:SFHS) (“ACONF-GS:SFHS”; n=558).

**Methods:** Childhood cognitive ability was tested within six months of the child’s 7th (n=494), 9th (n=491) and 11th birthdays (n=414) [Moray House Picture Intelligence; Schonell&Adams Essential Intelligence Test Form; Moray House verbal reasoning tests I/II, Moray House English and Moray House Arithmetic respectively]. All tests were standardized for age. Mid-life cognitive functions were assessed with Verbal Fluency (n=447), Mill Hill Vocabulary (n=445), and Logical Memory Delay (n=446). Birthweight in lbs (transformed to kg), father’s occupation at birth, and mother’s age at birth were abstracted from the Aberdeen Maternity and Neonatal Databank. Birth orders were self-reported. The statistical method was based on the study reported in Richards et al., 2001: birthweight was split into five categories [BW<2.49kg (“low”); 2.49kg≤BW<3.18kg (“normal-lower”); 3.18kg≤BW< 3.63kg (“normal-middle”); 3.63kg≤BW<4.08kg (“normal-higher”); BW≥4.08kg (“high”)]. The analytical subsample had complete information on birthweight and confounders (father’s occupation, mother’s age, birth order, mid-life age). The associations between birthweight grouped into five categories and cognitive functions at every age were examined with linear regression.

**Results:** Overall, 89% of ACONF-GS:SFHS participants (240 men, 318 women; mean age=57.4) had birthweight within normal range. About 57% had a father in skilled/semi-skilled manual occupation and 33% completed higher/professional education. In the analytical sample, “low” birthweight was negatively associated with childhood cognition at age 7 compared with “normal-middle” birthweight (β =-8.92, SE=3.64, p=0.015). “Normal-higher” birthweight was positively associated with childhood cognition at age 7 compared with “normal-middle” birthweight (β =3.78, SE=1.75, p=0.031). No other significant associations have been found. However, the expected pattern of associations was found only in mid-life (i.e. increasing mean cognitive function for the first four birthweight categories and decreasing for “high” birthweight).

**Conclusion:** Our preliminary analysis indicates that birthweight might not have long-term effects on cognitive functions as shown in Richards et al., 2001. Next, we will perform adjusted analyses and examine the association between birthweight adjusted for gestational age and cognition over the life-course.
Abstract 1507

USING LINKED ADMINISTRATIVE DATA TO REDUCE BIAS DUE TO MISSING DATA: A STUDY OF THE ASSOCIATION BETWEEN BREASTFEEDING AND IQ USING ALSPAC DATA AND SIMULATIONS

Authors: 
Rosie Cornish1, Kate Tilling2, Andy Boyd2, Amy Davies2, John Macleod2
1ALSPAC, University of Bristol, 2University of Bristol

Presenting Author: Rosie Cornish

Corresponding author email: rosie.cornish@bristol.ac.uk

Background: Most epidemiological studies have missing information, leading to reduced power and potential bias. Exposure-outcome associations will generally be biased if the outcome variable is missing not at random (MNAR). Linkage to administrative data containing a proxy for the outcome allows assessment of MNAR. We used data from the Avon Longitudinal Study of Parents and Children (ALSPAC) and simulations to examine bias in the association between infant breastfeeding and IQ at 15 years, using linked school attainment data as a proxy for IQ.

Methods:
ALSPAC: Subjects were those who enrolled in 1990-91 and were alive at one year (n=13,795), of whom 36% had IQ measured at 15. For those with missing IQ, 79% had data on attainment at age 16 obtained through linkage to the National Pupil Database. Breastfeeding information was collected via questionnaire at 1, 6 and 15 months. A number of potential confounders / factors predictive of non-response were collected during pregnancy. We estimated the association between duration of breastfeeding and IQ at 15 years using a complete case analysis, multiple imputation (MI), and MI including linked attainment data.

Simulations: In the simulations we changed the strength of association between the outcome and the linked proxy, the proportion of missing data, and the extent to which the outcome was MNAR.

Results: IQ measured at 15 in ALSPAC was MNAR – individuals with higher attainment were less likely to have missing IQ, even after adjusting for socio-demographic factors. The correlation between IQ and the main attainment variable was 0.59. Both complete case analysis and MI underestimated the association between breastfeeding and IQ compared to MI informed by linkage (mean difference in IQ comparing those breastfed for at least 6 months to those breastfed for less than one month was 4.2 (95% CI 3.4-5.0) using MI informed by linkage but 3.5 (2.5-4.4) in the complete case analysis). In simulations, including the linked proxy reduced bias and increased precision in all cases, although improvements were small when the correlation between the outcome and its proxy was low (r=0.5).

Conclusions: Linkage to administrative data containing a proxy for the outcome variable allows the MNAR assumption to be tested and more efficient analyses to be performed. Key limiting factors are the strength of association between the outcome and its proxy and coverage of the linked data; in our case, where the correlation was modest and linked data were not available for all individuals, some bias may remain.
Abstract 1593  

**COHORT & LONGITUDINAL STUDIES ENHANCEMENT RESOURCES (CLOSER): A PROGRESS REPORT ON DEVELOPING RECORD LINKAGE RESOURCES WITHIN COHORT STUDIES.**

Authors:  

Andy Boyd  
ALSPAC, University of Bristol

Presenting Author:  

Andy Boyd

Corresponding author email:  

a.w.boyd@bristol.ac.uk

Data collection via linkage to routine health and administrative records offers great potential to observational research. While cohort and longitudinal studies have recognised the value of adding linked records to their databanks, overcoming ethico-legal barriers has been challenging. Where data access has been established, the access arrangements require ongoing maintenance, evolutionary change as perceptions of good practice evolve, and sometimes seismic change in reaction to high profile events and incidents. Notwithstanding these difficulties, cohort studies are accessing a wide variety of health and administrative records and are developing the governance and technological methodologies needed to utilize them. However, it is also widely acknowledged that despite many successes, studies are still facing barriers to fully exploiting the potential of record linkage in a systematic and consistent manner. The Cohort & Longitudinal Studies Enhancement Resources (CLOSER) network was established to address challenges such as these. Through pooling the expertise and skills from nine of the UK’s longitudinal studies along with the British Library and the UK Data Archive, CLOSER aims to stimulate interdisciplinary research, develop shared resources and infrastructures to enable research, assist with training and development and to share expertise.

In this paper I will initially describe the progress cohort and longitudinal studies have made in developing record linkage as a means of collecting data on study participants. The evidence will be drawn from the experiences of UK cohort studies from the CLOSER network, in comparison with international cohort studies who participated in the European Child Cohort Network (EUCCONET). I will discuss some of the reasons behind the successes and challenges faced by studies in order to secure access to data and also emerging challenges once data have been acquired. The paper will also outline the relevant areas of CLOSER’s work programme, the relationship between CLOSER cohorts and the Farr Institute and Administrative Data Research Network, and the role that CLOSER plays as an advocate for longitudinal and cohort studies.
INCIDENCE OF TUBERCULOSIS IN MIGRANTS SCREENED PRE-ENTRY AFTER ARRIVAL IN THE UK: A RETROSPECTIVE COHORT STUDY USING DATA LINKAGE.

Rob Aldridge, Dominik Zenner, Elizabeth Williamson, Ibrahim Abubakar, Peter White, Andrew Hayward

1Farr Institute and University College London, 2Public Health England, 3Farr Institute and LSHTM, 4University College London, 5Public Health England and Imperial College London

Presenting Author: Rob Aldridge

Corresponding author email: r.aldridge@ucl.ac.uk

Background: UK tuberculosis surveillance data in their current configuration allow only a period prevalence of tuberculosis to be calculated in individuals not born in the UK. In 2005 a pre-entry screening programme for tuberculosis in migrants to the UK was started. Probabilistic matching can be used to identify incident cases of disease and missed prevalent cases, among individuals screened by the pre-entry programme. This would enable the estimation of burden of disease in the UK, among screened migrants abroad, and the risk factors associated with these cases.

Methods: This was a retrospective cohort study. The primary outcomes were incidence of bacteriologically confirmed pulmonary and extra-pulmonary tuberculosis. The database of migrants screened pre-entry between 1st January 2006 and 31st December 2012 was probabilistically linked to the UK Enhanced Tuberculosis Surveillance database for cases notified between 1st January 2006 and 31st December 2013. Individuals were followed up until the first of tuberculosis, death, or emigration. Poisson regression was used to estimate incidence rates, and a multivariable risk factor analysis was conducted. Multiple imputation was used to deal with missing data.

Results: After excluding duplicates and missed prevalent cases, 519,955 migrants entered the cohort. There were 622 cases of bacteriologically confirmed pulmonary tuberculosis with an estimated incidence rate of 65 cases per 100,000 person years at risk (95% CI: 60.7, 70) and 674 cases of extra-pulmonary tuberculosis with an incidence rate of 70 per 100,000 person years at risk (95% CI: 65.7, 76). After adjusting for age and sex, there was strong evidence that a history of contact with a case of tuberculosis before migration (IRR 4.9; 95% CI: 2.5, 9.4; p-value < 0.001) and a chest radiograph classified as consistent with tuberculosis at migration (IRR 4.4; 95% CI: 3.5, 5.5; < 0.001) were associated with an increased risk of bacteriologically confirmed pulmonary tuberculosis. Migrants screened pre-entry at clinics where culture testing was performed had a lower incidence of bacteriologically confirmed pulmonary tuberculosis post entry than those screened at sites where this was not conducted (IRR 0.6; 95% CI: 0.5, 0.7; p-value < 0.001).

Conclusion: Several novel findings are presented, including the first time direct estimates of the incidence of tuberculosis in a high-risk population that has been screened for active pulmonary disease prior to arrival. Incident cases can represent reactivation, re-infection and newly acquired tuberculosis after migration, and further work should examine this in more detail. These estimates and the risk factors identified can be used to inform health improvement and health protection policy in the UK.
SESSION E5 – STATISTICS & ANALYTICS.
CHAIR: DR MATTHEW SPERRIN, UNIVERSITY OF MANCHESTER

Abstract 1667  COLLIDER BIAS: DOES IT EXPLAIN THE OBESITY PARADOX?

Authors: Matthew Sperrin1,2, J Candlish1,2, E Badrick1,2, I Buchan1,2, A Renehan3
1Institute of Population Health, 2MRC Health eResearch Centre (HeRC North), 3Institute of Cancer Science, University of Manchester

Presenting Author: Matthew Sperrin
Corresponding author email: matthew.sperrin@manchester.ac.uk

The epidemiological literature is awash with findings that overweight and obese individuals have superior survival compared with individuals of a healthy weight. This is particularly common in patients with chronic disease (such as heart failure, coronary artery disease and diabetes). This is considered a paradox as it is well established that being overweight/obese should cause worse outcomes.

In observational studies we usually estimate association rather than causation: crudely, we can consider the paradox ‘resolved’ by the association-causation gap. However, it is of interest to understand why the association and causal effect may disagree in these particular scenarios. The ‘bias’ is the extent of the disagreement between the measured association and the true causal effect.

One statistical explanation that has gained popularity in the literature is that of collider stratification bias. Since being overweight is a strong risk factor for chronic disease, those in the diseased population who are not overweight must have other (possibly unmeasured) severe risk factors that caused that disease and in turn increase their risk in mortality.

In this talk we will describe, through mathematical results from causal analysis and simulation studies, the situations in which collider stratification bias can lead to associations being observed in the reverse direction of the causal effect. We will look at how severe the collider bias needs to be for this to happen, and whether this therefore represents a plausible ‘explanation’ for the obesity paradox. Naturally, the implications extend to any observational study on a selected population where we hope to estimate a causal effect.

Abstract 1474  THE IMPORTANCE OF UNDERSTANDING THE DETECTION TIME BIAS: EXAMPLES USING PATIENTS NEWLY DIAGNOSED WITH TYPE 2 DIABETES

Authors: Ellena Badrick1, Sperrin M2, I Buchan2, A Renehan3
1Institute of Population Health, 2MRC Health eResearch Centre (HeRC North), 3Institute of Cancer Science, University of Manchester

Presenting Author: Ellena Badrick
Corresponding author email: ellena.badrick@manchester.ac.uk

Background: The power of using routine data in epidemiology is large, however any biases must be understood and quantified before conclusions can be drawn from analysis. Detection time bias results from increased surveillance around the time of diagnosis of a condition, it’s more likely for a doctor to find another undiagnosed condition often with subclinical symptoms. This can lead to overestimates of risk between two conditions or even the reporting of a spurious association that has no biological plausibility. I will use a cohort of people newly diagnosed with Type 2 diabetes (T2D) to:

• Explore whether diagnosis of T2D coincides with cancer diagnosis
• Explore whether diagnosis of T2D coincides with hypothyroidism
• If detection bias is present explore how it can conflate any genuine increased risks

**Methods:** Patients with T2D are at increased risk of developing several common cancers, there is a general appreciation that research into cancer risk is needed but quantification of this risk requires consideration of time-varying factors. We used general practice data from the Salford Integrated Record (SIR, n= 248,917 of these 16,272 have indication of any diabetes diagnosis) linked with data from the North West Tumour Registry. Cancers were categorised by ICD10 codes, we defined hypothyroidism using readcodes or prescription of thyroxine. We split follow-up time and performed a split interval analysis, to generate Hazard Ratios (HR), for 0-6 months, 6-12 months, 1-2 years, 2-5 years and 5-10 years after diagnosis of T2D.

**Results:** For cancers, there were 4 to 10 fold increases in cancer diagnoses in first 12 months i.e a detection time bias. When any cancers detected in the first 2 years of follow-up were excluded: there was an increase in risk of obesity-related cancer (HR: 1.426, 95% CI: 1.116-1.821), but not for non-obesity related cancers (HR: 1.028, 95% CI: 0.912-1.361), compared with matched controls adjusted for smoking status and BMI at diagnosis. There were 1213 individuals with a diagnosis of hypothyroidism, they comprised of more women (70.8%) then non-cases. In split interval analysis we saw no evidence of detection time bias.

**Conclusion:** The detection time bias can occur in routinely collected data and to interpret and quantify cancer risk in patients with T2D, there is a need to account for this. However this pattern was not exhibited for hypothyroidism. It raises important methodological consideration when reporting cancer risk in people with T2D and clinical questions about the presentation and screening for cancer in primary care.

---

**Abstract 1420**

**THE DOUBLE BURDEN OF HEALTH INEQUALITY IN SCOTLAND: THE CASE FOR MEASURING LIFESPAN VARIATION BY SOCIOECONOMIC DEPRIVATION USING ROUTINE DATA.**

**Authors:** Rosie Seaman, A H Leyland, F Popham

**MRC/CSO Social and Public Health Sciences Unit, University of Glasgow**

**Presenting Author:** Rosie Seaman

**Corresponding author email:** r.seaman@sphsu.mrc.ac.uk

**Background:** Public health aims to improve average population health and reduce inequalities. Life expectancy measures improvements in average population health and increases when a death at any age is avoided. However this means increasing life expectancy can be achieved at the expense of widening the age distribution of death. Lifespan variation measures, in years, the level of uncertainty in the age distribution of death. It is important to measure by socioeconomic deprivation because deprived groups may experience a double burden of health inequality: lower life expectancies and higher lifespan variation. This may have worsened for Scotland since the 1980s because of rising premature mortality amongst the most deprived. Lifespan variation by deprivation in Scotland has not been measured and compared to life expectancy. Nor have the ages of death driving changes to each been identified.

**Methods:** Life expectancy and lifespan variation by area deprivation (Carstairs quintile) were calculated from lifetables generated using death and population data surrounding the 1981, 1991, and 2001 censuses (2011 pending). Lifespan variation was measured using Edagger. This shows the average life expectancy lost per death and is a weighted average of the remaining life expectancy at all ages. Stepwise decomposition calculates the contribution each age made to changes in (i) life expectancy and (ii) lifespan variation between 1981 and 2001.

**Results:** Life expectancy for men in Scotland was 69.0, 71.6 and 73.4 years and lifespan variation 11.8, 11.5 and 11.7 years in 1981, 1991 and 2001 respectively. Falls in mortality at ages 60+ led to increasing life expectancy across socioeconomic groups but rises in mortality for ages 20 to 40 meant gains for the most deprived group were smallest. Falls in mortality in ages 75+ led to increases in lifespan variation across socioeconomic groups but these were counterbalanced by falls in mortality at younger ages.
and a decreasing trend maintained. The most deprived group was an exception: rising mortality in ages 20 to 40 caused an increase in lifespan variation that was so large it had an impact on lifespan variation at the population level.

**Conclusions:** Measuring life expectancy and lifespan variation in Scotland shows the double burden of health inequality worsening over time. Measuring only life expectancy would have masked this growing disparity in age at death. Lifespan variation and decomposition methods can identify where mortality needs to be reduced to improve average population health and reduce inequalities.

**Abstract 1635**

**USING OBSERVATIONAL DATA FOR STRATIFIED MEDICINE: SOME CHALLENGES AND LIMITATIONS**

**Authors:**

Antonia Marsden¹, R Emsley², G Dunn², W Dixon³

¹Centre for Biostatistics, The University of Manchester; ²Centre for Biostatistics, Institute of Population Health, The University of Manchester; ³Arthritis Research UK Centre for Epidemiology, Centre for Musculoskeletal Research, The University of Manchester

**Presenting Author:**

Antonia Marsden

**Corresponding author email:** antonia.marsden@posgrad.manchester.ac.uk

A key goal of medicine is to treat the right patient with the right treatment at the right time. With an increasing number of therapeutic options available for an expanding and ageing population, to be delivered with a finite resource, it is preferable to avoid the ‘trial and error’ of current medical practice. The choice of the most appropriate treatment for specific subgroups of patients, rather than the population as a whole, is the science of stratified medicine.

Observational studies are increasingly being used to evaluate predictors of response to available treatments. Whilst the best quality evidence for stratified medicine comes from prospective randomised controlled trials, ethical considerations and sample size limitations, as well as time and cost, often make such clinical trials an unattractive and unfeasible option. As routine observational datasets can contain information on large numbers of patients, with numerous variables measured longitudinally over long periods of time, they offer an alternative environment in which to conduct the analysis.

We will discuss some of the statistical challenges that arise when attempting to identify predictive variables in observational data. We will highlight features of the analysis that have tended to be overlooked in the current literature and areas where the statistical methodology requires development. This includes 1) the distinction between predictors of treatment response (predictive variables) and predictors of outcome, regardless of treatment response (prognostic variables), and how this affects the analysis 2) the choice of appropriate measurement scale for interactions – effect modification on the additive scale measures differences in absolute treatment effects whilst effect modification on the multiplicative scale measures differences in relative treatment effects, 3) extending methodology for combining information from multiple predictors of treatment response currently limited to a clinical trial setting, to an observational setting, 4) the appropriate method of confounding adjustment, since treatment exposures in observational data are not randomly assigned.
EXTRACTING PATIENT JOURNEYS FROM HEALTHCARE NARRATIVES

Authors: A Dehghan1, M McCabe1, I Kamaly-Asl2, E J Estlin3, J A Keane1, G Nenadic1
1The University of Manchester, 2Central Manchester Hospitals NHS Foundation Trust, 3Blackpool Teaching Hospitals NHS Foundation Trust

Presenting Author: A Dehghan

Introduction: Automated extraction and representation of patient journeys from hospital records can aid a range of clinical quality improvement efforts. For example, this type of application can aid large-scale analyses of implemented care pathways and subsequently help compare, develop, and adjust clinical guidelines for both chronic diseases (where there is plenty of data) and rare conditions (where potentially there are no established guidelines).

Methodology: Clinical Text Mining (TM) methods have been developed to extract and chronologically order clinical events from unstructured longitudinal healthcare narratives, including internal clinical notes and patient correspondence. Specifically, data- and knowledge-driven Natural Language Processing (NLP) methods have been developed to address the following tasks:

1. Identification of mentions of clinical events, including medical problems (e.g., mentions of signs or symptoms, disease or syndrome), treatments (e.g., mentions of therapeutic or preventive procedures, medication), and tests (e.g., diagnostic procedure);
2. Identification of temporal expressions (i.e., mentions of time, date, duration, and frequency);
3. Temporal ordering of clinical events using temporal relations (e.g., Before, After, Overlap) between events and temporal entities. Subsequently, clinical events are clustered into temporally predefined buckets (e.g., 6-months temporal intervals, starting from the first consultation).

Finally, automated aggregation and visualisation techniques were engineered to generate patient journeys which represent the most common clinical concepts in the given bucket. Extracted patient journeys could then further be aggregated over multiple patients to generate pathways that indicate common practice in a given cohort.

The NLP methods were developed and validated using high quality datasets available in the NLP community as part of clinical text mining challenges (https://www.i2b2.org/NLP).

As a case study, we used a set of electronic health records within The Christie NHS Foundation Trust. The suite of TM methods was applied to chronologically order clinical events from unstructured longitudinal clinical narratives of young survivors of childhood medulloblastoma. Subsequently, concept clustering and visualisation techniques were used to represent individual and aggregated patient journeys.

Results: Both customary text mining and qualitative evaluation of the automatically generated patient journeys were conducted. Aggregated patient journeys achieved 97% in average precision and recall, while the individual patient journeys ranged between 80-94%. In addition, the qualitative evaluation showed expected patient journeys for the given cohort.

Conclusion: TM methods have shown promising result to reconstruct patient journeys from unstructured clinical narratives. In future work we aim to combine unstructured and structured information sources to reconstruct fine-grained patient journeys.
UTILITY OF SYNTHETIC MICRODATA FOR HEALTH RESEARCH

Authors: Joshua Snoke\textsuperscript{1}, Beata Nowok\textsuperscript{2}, Gillian M Raab\textsuperscript{2}, Chris Dibben\textsuperscript{2}, Aleksandra Slavković\textsuperscript{1} \\
\textsuperscript{1}Department of Statistics, Penn State University, \textsuperscript{2}School of Geosciences, University of Edinburgh

Presenting Author: Joshua Snoke

Corresponding author email: beata.nowok@ed.ac.uk

Wide access to microdata is crucial to the advancement in health research and evidence-based health policy but it is often constrained by confidentiality concerns. Synthetic data techniques developed to allow for the release of high-quality microdata without compromising confidentiality provide an attractive solution to this problem. In synthetic data risks of disclosures are minimized by replacing some or all of the data values with simulations from statistical models estimated from the original confidential data. The usefulness of the disseminated synthetic data depends, however, on the correct specification of these models which can be a difficult and complex task. The recently developed synthpop package for R simplifies considerably the process of generating synthetic data but utility assessment of the resulting synthesised version of the underlying confidential data is limited.

In this paper we develop a framework for evaluating the utility of synthetic microdata and then apply it to real datasets to explore different synthesising strategies in order to identify approaches that produce data of satisfactory quality. Note that the desired utility level will depend on whether synthetic data are going to be used for final analysis or for preliminary investigation only.

We use both analysis-specific and broad utility measures and implement them in the synthpop package. Researchers usually fit regression models to data and therefore analysis-specific measures focus on comparing regression estimates from the original and synthetic data. They are based on overlap of confidence intervals and statistically significant relations. Broad measures evaluate similarities between overall structures of the confidential data and their synthetic version. They rely on discrimination methods that are used to check the possibility to distinguish the two datasets and a measure based on propensity scores is the most promising one.

MODELLING OUTCOMES OF WHOLE GENOME SEQUENCING IN UNSELECTED POPULATIONS USING THE EXAMPLES OF BRCA1 AND BRCA2

Authors: Charlotte Warren-Gash\textsuperscript{1}, Mark Kroese\textsuperscript{2}, Hilary Burton\textsuperscript{2}, Paul Pharoah\textsuperscript{3} \\
\textsuperscript{1}UCL, \textsuperscript{2}PHG Foundation, \textsuperscript{3}University of Cambridge

Presenting Author: Charlotte Warren-Gash

Corresponding author email: c.warren-gash@ucl.ac.uk

Background: Population screening using whole genome sequencing (WGS) is controversial: while the American College of Medical Genetics and Genomics advocates actively screening clinically actionable 56 genes whenever WGS is undertaken, the European Society of Human Genetics recommends a more targeted testing and reporting strategy. In the 100,000 Genomes Project, Genomics England recommends actively seeking mutations in 16 genes including the BRCA1 and BRCA2 genes linked with hereditary breast and ovarian cancer syndrome. This approach will provide a model for future NHS practice, but implications for patients and services are unclear.
**Methods:** We modelled the test performance of WGS for identifying pathogenic BRCA1 and BRCA2 mutations in an unselected hypothetical population of 100,000 UK women, using published literature to derive model input parameters. We calculated analytical and clinical validity, described potential health outcomes and highlighted current areas of uncertainty.

**Results:** WGS is predicted to identify correctly 93 pathogenic BRCA1 mutations and 151 BRCA2 mutations in 120 and 200 women respectively, resulting in an analytic sensitivity of 75.5-77.5%. Based on population penetrance estimates of 59% for BRCA1 and 51% for BRCA2, we estimated that 132 women with identified mutations would develop breast cancer, compared to 41 women whose mutations were missed and 12,460 women without mutations. Uncertainties remain about the number of mutation carriers who would be identified across all the genes recommended for routine analysis on WGS, the penetrance of mutations in people without a family history of disease and the appropriate threshold of absolute disease risk for clinical action.

**Conclusions:** We use the example of BRCA1 and BRCA2 to demonstrate the type of process that should be undertaken to determine likely outcomes of WGS-based testing in unselected populations. Applying this process to other gene-disease combinations is likely to reveal further gaps in understanding of clinical validity and utility which should be considered before offering routine testing of genes in unselected populations.

---

**SESSION E6 – NATURAL EXPERIMENTS.**

**CHAIR: PROF SHEILA BIRD, MRC BIOSTATISTICS UNIT, CAMBRIDGE**

**Abstract 1399**  
**IMPACT OF OPIOID SUBSTITUTION THERAPY FOR SCOTLAND’S PRISONERS ON DRUGS-RELATED DEATHS SOON AFTER PRISONER-RELEASE**

**Authors:** Sheila Bird\(^1\), C M Fischbacher\(^2\), L Graham\(^2\), A Fraser\(^3\)

\(^1\)MRC Biostatistics Unit, Cambridge CB2 0SR. \(^2\)NHS National Services Scotland. \(^3\)NHS Health Scotland

**Presenting Author:** Colin Fischbacher

**Corresponding author email:** sheila.bird@mrc-bsu.cam.ac.uk

**Aim:** To assess whether the introduction of a prison-based opioid substitution therapy (OST) policy was associated with a reduction in drugs-related deaths (DRDs) within 14 days after prison-release.

**Design:** Linkage of Scotland’s prisoner database with death registrations to compare periods before (1996-2002) and after (2003-2007) prison-based OST was introduced. Variation on permissions for an extant record-linkage study on prisoners’ mortality was sought to meet our stated aim of policy-evaluation.

**Setting:** All Scottish prisons.

**Participants:** Persons released from prison between 1 January 1996 and 8 October 2007 following an imprisonment of at least 14 days [as in Bird & Hutchinson (2003)] and at least 14 weeks after the preceding qualifying release.

**Measurements:** Risk of DRD in the 12 weeks following release; percentage of these DRDs which occurred in the first 14 days. As designed, the before/after evaluation had at least 80% power to discern a reduction in the 14-day DRD-spike from 60% of 12-week DRDs down to 47% [the percentage that applied in New South Wales where prison-based OST was established much earlier than in Scotland, see Merrall et al. (2010)].

**Results:** Before prison-based OST (1996-2002), 305 DRDs occurred in the 12 weeks after 80,200 qualifying releases, 3.8 per 1,000 releases (95% CI: 3.4-4.2); of these 175 (57%) occurred in the first 14 days. After the introduction of prison-based OST (2003-2007), 154 DRDs occurred in the 12 weeks after 70,317 qualifying releases, a significantly reduced rate of 2.2 per 1,000 releases (95% CI: 1.8 to 2.5). However,
there was no change in the proportion which occurred in the first 14 days, either for all DRDs (87: 56\%) or for opioid-related DRDs.

**Conclusions:** The DRD-rate in the 12 weeks following release fell by two-fifths after prison-based OST was implemented. However, the proportion that occurred in the first 14 days did not change appreciably, suggesting that in-prison OST does not reduce the DRD-spike in the first 14 days post-release. Further targeted interventions are required to reduce DRDs soon after release.

**Reflection:** Despite three-quarters of Scotland’s DRDs being opioid-release, and that our policy evaluation was specifically powered to investigate the 14-day DRD-spike, which the international meta-analysis by Merrall et al. (2010) also focused on, referees considered our policy-evaluation to be 'fatally flawed' unless we could present results for the subset of DRDs that were opioid-related.

We are grateful that Scotland’s Privacy Access Committee allowed us to identify opioid-related DRDs, and satisfy referees.

### Abstract 1542

**IMPACT OF TREATMENT FOR OPIOID DEPENDENCE ON FATAL DRUG-RELATED POISONING: A NATIONAL COHORT STUDY IN ENGLAND**

**Authors:**

M Pierce, S M Bird, M Hickman, J Marsden, G Dunn, T Millar

1 University of Manchester, 2 MRC Biostatistics Unit, Cambridge, UK, 3 University of Bristol, 4 King’s College London

**Presenting Author:** S M Bird

**Corresponding author email:** matthias.pierce@manchester.ac.uk

**Background:** Observational studies suggest that opioid agonist pharmacotherapy for opioid dependence reduces the risk of fatal drug-related poisoning (DRP). To date, there have been few studies of DRP risk reduction for other intervention modalities for opioid dependence. This English national study examines whether pharmacotherapy, psychosocial and residential treatments have the same DRP risk-reduction, and whether differences in DRP risk-reduction are modified by client characteristics, treatment completion or referral from the criminal justice system.

**Methods:** For a cohort of adults in England who were treated for opioid dependence during April 2005 to March 2009 (151,983 individuals and 442,950 person-years of observation), death from DRP was established through data-linkage. Proportional hazards regression assessed DRP risk associated with being in treatment versus not in treatment and with setting and intervention modality, with adjustment for confounders.

**Findings:** There were 1,499 DRP deaths (3.4 per 1,000 person-years, 95\% CI 3.2–3.6). DRP risk was increased when members of the cohort were not in treatment (adjusted hazard ratio [aHR] 1.7, 95\% CI 1.6–1.9). DRP risk when enrolled in only psychosocial intervention was double that observed during pharmacotherapy (aHR 2.1, 1.7–2.5). There was insufficient evidence that the DRP risk for periods out of treatment was modified by whether or not the client had completed prior treatment (p = 0.11). The relationship between treatment status and DRP risk was stronger among males (aHR 1.9, 1.7–2.1), illicit drug injectors (aHR 2.3, 2.0–2.6) and those reporting an alcohol use disorder (aHR 2.4, CI 1.9–3.0); it was considerably weaker for those referred via the criminal justice system (aHR 1.1, 0.9–1.5).

**Interpretation:** In England, opioid agonist pharmacotherapy for opioid dependence is associated with a 40\% reduction in DRP risk. DRP risk is not reduced among those who receive psychosocial treatment only. There is a need to enhance the effectiveness of efforts to reduce behaviours and health disorders linked to fatal DRP risk among clients who receive psychosocial intervention.
**Abstract 1558**

THE EFFECT OF INCENTIVE PAYMENTS TO PRIMARY CARE PHYSICIANS FOR COMPLEX CARE: MORE EVIDENCE FROM BRITISH COLUMBIA, CANADA

**Authors:**
Kim McGrail¹, Ruth Lavergne¹, Michael Law¹, Jerry Hurley², Sandra Peterson¹, Scott Garrison³

¹University of British Columbia, ²McMaster University, ³University of Alberta

**Presenting Author:** Kim McGrail

**Corresponding author email:** kim.mcgrail@ubc.ca

**Context:** The Canadian province of British Columbia introduced incentive payments to primary care physicians for chronic disease management, with no simultaneous changes to service delivery model. Primary care physicians in BC still largely practice in individual or small group settings and are paid predominantly through fees-for-service. The existing literature is equivocal about the effect of incentive payments on patient care. The complex care fee item introduced in 2007/08 offers a bonus payment of $315 for each patient who has at least two of a select list of chronic diseases, and for whom the physician has provided continuity of care guided by a care plan.

**Objective:** To examine the effect of the chronic care incentives on the population of potentially eligible patients.

**Methods:** Population-based interrupted time series, using fee-for-service physician payment records from 2005/6 to 2011/2012, linked to patient registry, pharmacy, and hospitalization data. The study population was patients eligible for the incentive based on diagnosis codes in administrative data. Outcomes included: continuity of care; hospitalization rates; and total per-capita healthcare costs.

**Results:** The study population was 132,350 patients, of which about 70% had at least one incentive billed for their care. Those with incentives billed within the study period were older, had more chronic conditions, and were more likely to receive their qualifying diagnosis in the latter two years of the study period. The introduction of incentives produced a very small increase in visits with primary care physicians (less than half a visit per patient over 2 years) and no effect on continuity or acute hospitalization. Total costs increased by an amount that is roughly twice the cost of the incentive.

**Conclusion:** These findings are generally consistent with previous research that has found limited impact of incentives within primary care. Mounting evidence for at best limited effects of incentive schemes, and particularly for those using this style of “pay for participation”, should encourage policymakers to look to other levers to improve health care quality, equity and efficiency.

**Abstract 1590**

USING A NATURAL EXPERIMENTAL APPROACH TO EVALUATE THE IMPACT ON PRESCRIPTIONS AND HOSPITAL ADMISSIONS OF THE PRESCRIPTION FREE ABOLITION IN SCOTLAND

**Authors:**
Andrew James Williams¹, Ruth Jepson², Marion Bennie³, Chris Dibben², Anne Ludbrook⁴, Jill Pell⁵, John Frank²

¹Farr Institute @ Scotland, ²Farr Institute @ Scotland, University of Edinburgh, ³Farr Institute @ Scotland, University of Strathclyde, ⁴Farr Institute @ Scotland, University of Aberdeen, ⁵Farr Institute @ Scotland, University of Glasgow

**Presenting Author:** Andrew James Williams

**Corresponding author email:** a.j.williams@ed.ac.uk

**Introduction:** In April 2011 the Scottish Government abolished fees for prescriptions, after a 3 year period of stepped reductions. However prior to the abolition, exemptions from the fees were already in place for people of certain ages, with certain health conditions, or below certain means (e.g. in receipt of financial benefits). Consequently prior to the abolition around 80% of prescriptions had not...
incurred a fee. The aim of this study was to detect any impact of the policy change not only on prescriptions (number, cost and defined daily doses (DDDs)) but hospital admissions, savings from which might help balance the cost of the policy.

**Methods:** The Prescribing Information System (PIS) in Scotland holds data on community issued and dispensed prescriptions. Linking data from the PIS and Scottish Morbidity Record (SMR01) permitted a difference in differences, interrupted time series approach. The fee exemption criteria were used to define groups for whom prescriptions became or remained free as intervention and control (counterfactual) respectively. The identified intervention group comprised adults (ages 19-59 years) receiving inhaled corticosteroids presumably for asthma or chronic obstructive pulmonary disease (COPD). Two control groups were identified which had been entitled to free prescriptions before the policy change; (a) those receiving inhaled corticosteroids but in a different age group and (b) those receiving prescriptions for a different condition (diabetes). Those eligible for means-related fee exemption could not be differentiated in the SMR01, thus potentially obscuring any effect. PIS limitations meant that only practice-level data and secondary non-adherence could be assessed. Subsequently, the analysis was stratified by a number of practice characteristics, including quintiles of the Scottish Index of Multiple Deprivation.

**Results:** Following preparation and cleaning, the final dataset comprised data from 798 practices across Scotland for the period from July 2005 to December 2013. Across the reduction and abolition periods prescriptions increased in each of the intervention and control groups. No change in the trend of admissions for diabetes, concurrent with the policy was detected, whereas the trend in admissions for asthma or COPD levelled off during the reduction and abolition periods. However, this change occurred in both the intervention and age-based control groups.

**Conclusion:** The initial findings indicate that there were changes in admissions for asthma or COPD concurrent with the policy change. However, the complexity of the prior fee exemption system makes detecting whether these changes were attributable to the policy difficult.

**Abstract 1466**

**DOES INADEQUATE TREATMENT OF ESCHERICHIA COLI (E. COLI) URINARY TRACT INFECTIONS LEAD TO INCREASED RISK OF E. COLI BACTERAEMIA?**

**Authors:** A M Walters¹, M L Heginbothom², R A Howe², M R Evans², J Arnott², R A Lyons¹

¹The Farr Institute, Swansea University, ²Public Health Wales

**Presenting Author:** A M Walters

**Corresponding Author email:** a.m.walters@swansea.ac.uk

**Introduction:** Bacteraemia is the presence of bacteria in the bloodstream and is of public health importance due to the high morbidity and mortality associated with this condition. There has been a rise in the number of E. coli bacteraemia in Wales during the last decade. Public Health Wales have been requested to undertake an investigation into the rise of E. coli bacteraemia by the Chief Medical Officer for Wales. This project has several work streams; using the anonymised routinely collected administrative data stored in the Secure Anonymised Information Linkage (SAIL) databank to provide descriptive and risk factor analysis forms work package 2.

**Method:** Anonymised urine and blood microbiology culture data reported between 2005 to 2011 are included in the SAIL databank. E. coli urine cultures with antibiotic sensitivities and E. coli blood cultures were identified and linked along with GP antibiotic prescribing data to test the hypothesis that patients who receive inadequate treatment of E. coli urinary tract infections (treating the E. coli urinary tract infection with an antibiotic to which the organism has been reported as resistant) are more likely to develop E. coli bacteraemia. Logistic regression modelling was used to test this hypothesis with E. coli bacteraemia as the outcome variable and treatment adequacy and demographics as explanatory variables. Survival analysis was also carried out to find the risk of mortality at 30 days, 90 days and 1 year following E. coli bacteraemia.

**Results:** 11.5% of patients with E. coli urinary tract infection received inadequate treatment. Initial results
show that, after adjustment, the odds of a patient who received inadequate treatment getting *E. coli* bacteraemia are 5.24 times the odds of a patient who received adequate treatment getting *E. coli* bacteraemia. The odds of a patient who did not receive GP treatment getting *E. coli* bacteraemia are 12.99 times the odds of a patient who received adequate treatment getting *E. coli* bacteraemia. Survival analysis shows that the estimate of the risk of death 30 days, 90 days and 365 days following *E. coli* bacteraemia is 18.1%, 25.9% and 37.3% respectively.

**Abstract 1524**

**SOCIAL CARE AND HEALTH OF OLDER PEOPLE: WHAT CAN WE LEARN FROM LINKING ROUTINE DATA?**

**Authors:** Sian Morrison-Rees¹, Martin Bardsley², David Ford³, John Grenfell⁴, Alison Porter¹, Martin Heaven⁵

¹College of Medicine, Swansea University, ²Nuffield Trust, ³College of medicine, Swansea University, ⁴City and County of Swansea, ⁵Farr at Swansea University

**Presenting Author:** Martin Heaven

**Corresponding author email:** s.morrison-rees@swansea.ac.uk

**Background:** Appropriate social care delivered at the right time can potentially maintain health, and delay or avoid the need for health care interventions such as hospitalisation. Conversely, a spell in hospital – for example after a fall – can often mark the beginning of a need for social care.

By using linked anonymous data of people aged 65 and over in one local authority area, we aim to examine how social care interventions may help to reduce hospitalisation, and how demand on social care may change over time as a result of changes in health care input.

**Methods and data sources:** We worked with Social Services staff in a Welsh local authority to securely prepare and extract electronic service user records from their information system for a seven year period, 2006-2012. This data was then anonymously incorporated into the SAIL databank and linked to healthcare data relating to hospital admissions. We explored the feasibility of extending this data linkage model to all of Wales by analysing the data preparation tasks and by interviewing the other local authorities in Wales.

**Results:** We achieved a 91% match rate and have 64,295 anonymous linked records. 28,884 people had an emergency hospital admission during the study period; 7,592 were in receipt of local authority supported social care; and 5,576 experienced both types of care. Detailed analysis is underway to examine the data in terms of the nature and timing of social care in relation to hospital admissions and how this is relationship is affected by age and type of social care received. We will also illustrate our findings by presenting representative individuals’ journeys as they use both health and social care services.

**Conclusions:** With the anticipated greater integration of health and social care, and increasing demand on both sectors, it is important to understand how the two service areas interact with each other. We have demonstrated that analysis at an individual level can inform this process. We have identified opportunities and challenges for extending this model across Wales.
SESSION F1 – APPS AND INTERVENTIONAL INFORMATICS.
CHAIR: PROF IAIN BUCHAN, UNIVERSITY OF MANCHESTER & THE FARR INSTITUTE

Abstract 1597

Computable Synergy between Clinical Audit & Feedback and Decision Support

Authors: B Brown, N Peek, R Williams, I Buchan
Health e-Research Centre, Farr Institute for Health Informatics Research, Centre for Health Informatics, University of Manchester

Presenting Author: Iain Buchan

Corresponding author email: benjamin.brown@manchester.ac.uk

Introduction: The challenge of implementing evidence-based care into practice is called the ‘second translational gap’. Health Informatics interventions often attempt to close this gap by delivering feedback to clinicians during encounters with patients (“clinical decision support” [CDS]) or at the population-level (“audit and feedback” [A&F]). Systematic reviews suggest both CDS and A&F are moderately effective and highly variable at improving care for patients.1-3 We suggest CDS and A&F share common mechanisms, and that cross-fertilizing features and concepts between them may consequently improve their efficacy.

Aim and objectives: The aim is to cross-fertilise the utility of CDS and A&F. The objectives are to: 1) develop a model of computable synergy between CDS and A&F; 2) demonstrate how the model can enhance UK primary care.

Methods/Results: CDS and A&F use the same ‘substrates’ (patient data and pre-defined clinical standards), similar analytic methods (event-condition-action rules versus quality indicators), and analogous behaviour change methods (information feedback to clinicians). Opportunities for cross-fertilisation between CDS and A&F include the provision of actionable improvement suggestions at the individual patient, population and organisational levels. These principles will be demonstrated in prototype software that is currently being pilot tested in primary care settings in Greater Manchester, UK.

Discussion: Previous attempts to increase the effectiveness of CDS and A&F have been relatively unsuccessful and largely limited to bottom-up aggregation of empirical evidence concerning heterogeneous quality improvement interventions. Our approach advocates a principled fusion of mechanisms that underpin how these tools work, going beyond the notion of simple ‘multifaceted’ or ‘co-’ interventions. We demonstrate how this approach may succeed where others have failed.

References

A MULTI-AGENT PLATFORM FOR AUTOMATING THE COLLECTION OF PATIENT-PROVIDED CLINICAL FEEDBACK

Authors: Z Ibrahim¹, L Fernandez De La Cruz², A Stringaris¹, R Goodman¹, R Dobson¹
¹King’s College London, ²Karolinska Institutet Stockholm, Sweden

Presenting Author: Z Ibrahim

Corresponding author email: zina.ibrahim@gmail.com

Patient-provided Outcome Measures (PROMs) are well-established quantitative assessments of medical treatment and have been used as clinical tools to improve patient experiences and outcomes. However, recent studies suggest that timely and long-term clinical outcome measures are not usually found in UK patient records due to the reluctance of clinicians treating the patients, whose active schedules incline them to rely on impressions to record patients’ progress, resulting in low rates of standardised outcomes.

To ease the clinicians’ burden and improve the process of patient follow-up, we implemented a platform to automate the process of collecting PROMs without clinicians’ intervention. The system also acts as a mediator between patients and scientists, automatically detecting patients’ eligibility to approved clinical trials and acting to seek patient consent to incorporate anonymised versions of the collected data into the trials. Therefore, the system seeks to 1) Improve outcome measure collection by incorporating them into routine clinical work without relying on clinicians 2) Improve outcomes using timely personalised communications empowering patients and guiding them to self-care outside hospitals, increasing their commitment to treatment and 3) Incorporate clinical trials into routine care.

To address the complexity of the healthcare domain and achieve an easily-generalisable design, we designed our platform as a multi-agent system composed of a number of autonomous, intelligent, distributed and social software components (agents) that work together to accomplish complex, heterogeneous tasks.

We implemented a prototype targeting children diagnosed with ADHD and prescribed methylphenidate at the South London and Maudsley National Health Services Foundation Trust (SLaM). Methylphenidate is a highly-efficient drug. Nevertheless, clinicians are ambivalent about the medication as it is essentially a controlled drug (with a chemical structure similar to cocaine). However, a recent study investigating the outcomes of methylphenidate treatment in SLaM found that only 1% of the 8000 patients investigated have timely PROMs.

We implemented the prototype using JADE (Java Agent DEvelopment Framework). The system communicates with patients through a web application (myHealthe.co.uk). We used the JadeGateway API to bridge the agent framework with the web application. We used Protégé to create a cross-platform ontology which the agents use for communication.

In summary, we have developed an agent-based system to automate the collection of patient-provided feedback, implemented at the South London & Maudsley NHS Foundation Trust, which is enabling trials of drug effectiveness in routine clinical practice.
Abstract 1463

ACEMOBILE – AN APP BASED PLATFORM SUPPORTING RESEARCH TO IMPROVE DEMENTIA ASSESSMENT VIA THE ACQUISITION OF DATA FROM DEMENTIA ASSESSMENT CLINICS.

Authors: Craig GJ Newman¹, Rupert F Noad², John Hodges³, John Zajicek⁴

¹Plymouth University Peninsular Schools of Medicine and Dentistry, ²Plymouth Hospitals NHS Trust, ³Neurosciences Research Australia, ⁴St Andrews University

Presenting Author: John Zajicek

Corresponding author email: craig.newman@plymouth.ac.uk

A prominent tool used in the clinical assessment of dementia, is the Addenbrooke's Cognitive Examination (currently, the ACE-III), a brief fifteen-minute assessment battery designed to support the differential diagnosis of dementia. The ACE-III is used mostly in secondary specialist care centres as recommended by the Alzheimer's Society for his purpose (The Alzheimer's Society, 2013). Its success reflects the fact that it was both designed specifically for assessment of dementia and is made available at no cost to all users.

ACEmobile is an iPad version of the ACE-III which has been designed to support the clinician in all areas of administration. ACEmobile provides a computer-based tool built on the familiarity of the ACE-III, deviating only in administrative behaviours and retaining the ACE-III in its original form. For the patient the experience is unchanged, for the clinician there is no need to digest a rules / scoring manual, no need to translate responses into scores, no need to pick up a timer, no need to perform arithmetic scoring and no lengthy sub-test totalling to perform. ACEmobile aims to guide more of the clinician’s focus towards the process of assessment and hopefully reduces chances for error.

ACEmobile is provided for free, maintaining the persistent ethos of the ACE brand, for clinical use both within and outside of the NHS on a worldwide scale. The tool, and its sustainability, is built on the framework of a research endeavour. It brings with it the ability to effortlessly collect anonymised data relating to the assessment of dementia sub-types. Specifically, it provides a framework for analysing the efficacy of all components of the ACE-III in varying clinical contexts and in answering varying clinical questions. The team behind the tool hope to encourage a community of researchers interested in using this pool of data to further develop the tool towards increased sensitivity and specificity in earlier stages of disease progression.

The model is innovative and relatively new. Data is collected, which can rationalise further funding which, in turn, could improve the tool further. A circularity that promotes sustainability, growth and provision towards a solution to an increasing NHS problem. ACEmobile and its underlying model present a method for research that is implicitly motivated to benefit patients, clinicians and the NHS. Since its release is a period where this marriage is continuously tested, novel methodological data acquisition alongside evolving clinical practice. This presentation provides an update on the project.
AN EHEALTH APPROACH TO REPORTING ALLERGIC REACTIONS TO FOOD AND CLOSING THE KNOWLEDGE GAP

Authors: Chris Munro¹, Aida Semic-Jusufagic², Katarzyna Pyrz³, Philip Couch¹, Audrey Dunn-Galvin¹, Niels Peek¹, Marina Themis², Clare Mills², Iain Buchan¹, Jonathan Hourihane³, Angela Simpson²
¹University of Manchester, ²University of Manchester and University Hospital of South Manchester, ³University College Cork

Presenting Author: Chris Munro

Corresponding author email: Christopher.munro@manchester.ac.uk

There is an important knowledge gap in food allergy management, in understanding the factors that determine allergic reactions to food, in gathering objective reports of reactions in real time, and in accessing patients’ reaction histories during consultations. We investigate how eHealth methods can close this knowledge gap. We report experiences with an online tool for reporting allergic reactions that we have developed as a web application. This application has been successfully validated by participants from Ireland and the UK, and is currently being used in a pilot where participants report allergic reactions in near-real time.

Allergic disease is a growing health risk, while its management by clinicians and patients is challenging. Food allergy has reached epidemic proportions in developed parts of the world with up to 20 million European citizens suffering from food allergy and reports of increasing prevalence in developing countries. The reasons for such an increase are not well understood. The estimated worldwide prevalence of food allergy varies according to age with 3-8% reported prevalence among children and 1-3% among the adults. Unsurprisingly, food allergy is the leading cause of anaphylaxis seen in emergency departments across the USA and UK.

Our aim was to improve the capture of objective information on accidental allergic reactions to food to address the knowledge gap in food allergy.

Our primary objective was to develop a system that can be used by food allergy sufferers to report information about suspected allergic reactions. Our secondary objective was to develop this system as a prototype reporting tool that could be used by clinicians and patients to report and view previous allergic reactions in a clinical context.

This web based system: AlleRiC (Allergic reactions in the community) has been developed and now enables reporting of suspected allergic reactions to food in near-real time. The system has been used in a validation study with 39 adults from the UK and Ireland who have diagnosed food allergies — successfully reporting historical reactions. The system is now deployed in a pilot study in the UK and Ireland where participants are reporting reactions prospectively in near real time.

As more people spend more of their lives on-line there is an epidemiological opportunity to tackle recall bias in in ways that might otherwise increase sampling bias due to differences in technology access. We have used this e-epidemiology tipping-point to address a big gap in food allergy knowledge.
LINKING ELECTRONIC HEALTH RECORDS, PATIENT-REPORTED SYMPTOMS, AND ENVIRONMENTAL DATA TO CHARACTERISE COPD EXACERBATIONS: A MIXED-METHODS STUDY NESTED IN CPRD

Authors: Elizabeth Moore1, Jennifer Quint1, Antonis Kousoulis2, Kirin Sultana2, Lia Chatzidiakou3, Rod L Jones3, Liam Smeeth1, Sean Beevers4, Frank J Kelly4, Benjamin Barratt4
1London School of Hygiene & Tropical Medicine, 2Clinical Practice Research Datalink (CPRD), 3The University of Cambridge, 4King’s College London

Presenting Author: Elizabeth Moore
Corresponding author email: Elizabeth.Worsley@lshtm.ac.uk

Background: Changes in temperature, humidity and air pollutants such as ozone, PM10 and PM2.5 are related to chronic obstructive pulmonary disease (COPD) exacerbations, but effects are not well categorised at the individual patient level. The aim of this study is to recruit patients in London using the Clinical Practice Research Datalink (CPRD) to examine the relationship between environmental exposures and seasonal changes in COPD exacerbations utilising information from personal air quality monitors and Electronic Health Records (EHR).

Methods: Across Greater London GP practices participating in research in CPRD are being approached, and using validated algorithms1 eligible individuals with COPDs are being identified and invited to take part in the study, targeting 160 consented recruits. The COPD algorithm is based on prior medical history, with a deliberate bias towards those showing a relatively high frequency of exacerbations (≥2 in the preceding year), who are not housebound.

Participants are asked to carry a Personal Air Monitor (PAM) with them every day for six months to monitor temperature, humidity, carbon monoxide, nitrogen oxide, nitrogen dioxide, ozone, PM10 and PM2.5, background noise levels, acceleration and GPS location. They are also asked to complete daily diary cards noting any changes to their symptoms, treatment (e.g. medications), sleep disturbance, and measure their peak expiratory flow (PEF) using a peak flow meter. Additionally, they are consenting to access of their full EHR, allowing us to obtain detailed information on health care utilisation exacerbations and important co-morbidities.

Results: We estimate that over 60 GP practices in Greater London who are registered with CPRD provide potential access to at least 3,000 patients for screening for the study. We anticipate that around 10% of eligible people identified are likely to be interested and participate. The initial recruitment results will be available in the summer of 2015.

Discussion: Identifying participants for inclusion in studies via anonymised EHR, such as CPRD, provides a unique approach. This method of recruitment has the potential to capture a different group of patients that are naive to clinical studies. The ability to then link data from individual patients to their full EHR provides another depth to research.
**Abstract 1574**

**MY DIABETES MY WAY: SUPPORTING DIABETES SELF-MANAGEMENT**

**Authors:** S G Cunningham, L Wilson, M Brillante, B Allardice, D J Wake  
*University of Dundee*

**Presenting Author:** S G Cunningham

**Corresponding author email:** scott.cunningham@nhs.net

**Background:** My Diabetes My Way (MDMW) is the NHS Scotland diabetes website for people with diabetes and their carers. It consists of an interactive information website available to all, and an electronic personal health record (ePHR) available to 276,430 people with diabetes registered with a general practitioner in Scotland. We analysed usage and activity during 2014.

**Methods:** We analysed system audit trails to monitor page accesses on the information website and logins and activity within the ePHR. The ePHR contains data from SCI-Diabetes, NHS Scotland’s flagship diabetes record. This system sources data from primary care, secondary care, specialist screening services and laboratory systems; including diagnostic information, demographics, process outcomes, screening results, medication and clinical correspondence. These data provide a more complete overview of diabetes than would be available from any single data source.

**Results:** The MDMW information website received an average of 52,837 page accesses per month during 2014 (54.7% increase from 2013; n = 34,151). 3,696 patients (34.9% of registrants) had completed the enrolment process and logged in to access their records by the end of 2014 (92.2% increase since end 2013; n = 1,923). Levels of engagement remain high amongst active users. 1,570 (42.5%) logged in at least once during the final 3 months of 2014: 2,197 (59.4%) from July to December and 2,912 (78.8%) during the full calendar year. There were 18,497 total logins during 2014 (average = 6.4 / patient; median = 3). During December 2014, 682 people with diabetes accessed their records (100.6% increase from December 2013; n = 340). Audit trails show 218,921 total page views (75.2 / patient). The ‘test results’ screen was the most popular summary (34,743 accesses, 11.9 / patient). The most utilised history graph was HbA1c (8,965 accesses, 3.1 / patient). Feedback: “newly diagnosed and find mdmw very handy as it is near impossible to get through to the doctors these days to get results”; “What a fab resource, wish we had this in @NHSEngland”

**Conclusion:** MDMW is a useful aid to diabetes self-management in Scotland. It is unique in offering access to an entire national population, providing information from many diabetes-related sources. MDMW supports the diabetes improvement, self-management, healthcare quality and eHealth strategies of the Scottish Government. It has potential to be adapted to work with other clinical systems and conditions.

**Abstract 1735**

**USING SMARTPHONES TO EXAMINE THE ASSOCIATION BETWEEN WEATHER AND JOINT PAIN IN PATIENTS WITH RHEUMATOID ARTHRITIS: A FEASIBILITY STUDY**

**Authors:** Karen Spencer¹, John Ainsworth¹, John McBeth¹, Max Little², David Schultz¹, Caroline Sanders¹, William G Dixon¹  
¹University of Manchester, ²Aston University

**Presenting Author:** William G Dixon

**Corresponding author email:** karen.spencer@manchester.ac.uk

**Background:** Patients with rheumatic disease have long reported an association between weather and joint pain. Previous studies have found 85% of patients believe an association exists, whilst 54% believe they can predict the weather based on their symptoms. Nonetheless, previous research has failed to identify the association.

The increasing penetration of smartphones opens new opportunities for epidemiology, enabling the collection and transmission of frequent patient data using both self-report and passive movement and
position data.

**Aim:** This feasibility study is to demonstrate proof of concept that patients will use smartphones to support patient-led research in musculoskeletal disease, using the exemplar study of weather and joint pain in patients with rheumatoid arthritis (RA). In addition to the weather project, it aims to test whether passively collected position and movement data will enable development of a validated algorithm to define disease severity in RA.

**Methods:**
1. Qualitative methods using focus groups (n=2) and a patient advisory group to inform co-design of an appropriate app to capture data in collaboration with uMotif
2. Enrollment of 20 patients to pilot the app collecting daily data (pain symptoms, weather data and passive movement) over 8 weeks
3. Qualitative longitudinal interviews to explore barriers and enablers of ongoing engagement with data collection
4. Descriptive statistics outlining the extent of missing data and attrition to demonstrate the feasibility and likely success of conducting a larger-scale project
5. Exploratory analysis of the correlation between i) weather and joint pain, ii) movement data collected from the smartphone app and a wrist-worn accelerometer, and iii) movement data and self-reported disease severity

**Progress to date:**
* Smartphone app for self-reported data co-designed with patients
* Parallel ‘capture app’ developed to collect and transmit GPS and accelerometer data, and to collect weather variables from local weather station
* Baseline focus groups demonstrated participants were enthusiastic to be involved in providing data to test associations between weather/passive functioning and self-reported symptoms. The uMotif app was considered to be user-friendly
* 18 patients recruited so far
* Two month data collected due to complete June 2015

**Summary:** Smartphones offer new opportunities for patient-led data collection for population research. Early work suggests this is feasible, and patients value the opportunity for great involvement in population research.
INTERACTIVE VISUALISATIONS TO EVALUATE AND PREDICT HEALTH

Authors:
Kerry Bailey\textsuperscript{1}, Helen Thomas\textsuperscript{2}, Kelly Nock\textsuperscript{1}, Luke Player\textsuperscript{3}
\textsuperscript{1}We Predict, \textsuperscript{2}Abertawe University Health Board, \textsuperscript{3}Swansea University

Presenting Author: Kerry Bailey

Corresponding author email: Kbailey@wepredict.co.uk

Background: A challenge ‘Can we improve health through the better use of data?’ was posed by Abertawe Bro Morgannwg University Health Board and SAIL (Secure Anonymised Interlinked) Databank based at Swansea University. This was fully funded by Innovate UK through the SBRI (Small Business Research Initiative). The scheme works through a ‘competition’, six companies being funded to complete a feasibility phase then two successful through to the demonstration phase.

We Predict is a small Welsh start-up, who provides a predictive analytics service to the automotive sector and had recruited a doctor to explore whether their skills would be of benefit in the health sector.

Objectives:
- Use data linking and analysis methods on data held within the SAIL databank to describe, evaluate and predict health and healthcare in the ABMU Health Board area.
- Present this intelligence in a user friendly interactive interface for use by lay people, health managers and clinicians to inform planning and evaluation of services.
- Produce tables and views for SAIL for analysts to interrogate.
- Adapt and apply We Predict expertise in predictive analytics to the health sector.

Method: Swansea University project managed and provided system and governance support. We Predict data scientists cleaned, redefined and linked data at individual level from 4 datasets – GP, OPD, ONS & HES. A SQL database was built with new tables & views, then systematised so processes are repeated when data is updated. Complex algorithms were necessary, e.g. over 200 separate READ and ICD10 codes were used to define type 2 diabetes. All three organisations met regularly to decide on priority areas and develop the specification.

Results: Individual level data was available on 1 million individuals over 20 years and 360 million lines of data in the GP dataset alone. Quality and completeness varied across datasets and through time. In the initial priority areas of work, obesity and diabetes, 4.5 million values of obesity were used. Innovative descriptive and predictive visualisations were produced which would not otherwise have been available and can inform decision makers.

Conclusion: The private sector has relevant skills and approaches which can bring innovation to the visualisation of intelligence which in collaboration with NHS and academia can bring improvements to health intelligence. The SBRI is a funding opportunity to develop innovations. The total spend on this project was £800,000. Analyst teams cannot usually expect this level of funding for data analysis or finding innovative ways to display intelligence.
Abstract 1661  CLiMGuiDeS: AN ONLINE COLLABORATIVE ENVIRONMENT FOR THE DEVELOPMENT OF CLINICAL MANAGEMENT GUIDELINES FOR RARE DISEASES

Authors: Idoia Gomez-Paramio1, G. Nenadic2, J. Clayton-Smith3
1Faculty of Life Sciences, 2School of Computer Sciences, The University of Manchester, Manchester, United Kingdom; HeRC, Farr Institute of Health Informatics Research, Manchester, United Kingdom, 3Central Manchester University Hospitals NHS Foundation Trust, Manchester, United Kingdom

Presenting Author: Idoia Gomez-Paramio

Corresponding author email: idoia.gomez-paramio@manchester.ac.uk

Rare diseases are defined as life-threatening or chronic debilitating illnesses whose prevalence is lower than 1 in every 2,000 people. There are around 6,000 known rare diseases and it is estimated that around 7% of the population will be affected by them at some point in their life. Thus, despite each condition may be individually rare, they are important collectively.

Clinical management guidelines for rare diseases are essential because individual doctors, when faced with a patient with a rare disease will almost certainly be unfamiliar with it. Optimum management of rare diseases improves quality of life, and also has financial benefits for the individual, their family and society.

Guidelines need to be carefully drawn up, based on available evidence and multidisciplinary expert consensus opinion. However, because of their rarity the medical literature on many rare diseases is scarce and difficult to access and the evidence base for therapies and care is poor. Professionals and parents/families that have intimate knowledge of the natural history of the disorder need to be able to actively engage in this process.

The development of guidelines is a costly process in terms of time and human resources required. As an example, the Dyscercne project, produced guidelines for four rare diseases in three years with a team of three members working full time.

CLiMGuiDeS (a Clinical Management Guidelines Development Site) has been developed to support this development. It offers an all-in-one solution by providing a free, sustainable and user-friendly online platform to streamline the guidelines development process. This process is divided into several steps based on the successful, means-tested Dyscercne procedure, which adapted the SIGN (Scottish Intercollegiate Guidelines Network) methodology to rare conditions and limited evidence base.

CLiMGuiDeS models and provides information and tools to facilitate the completion of each step during this process. For example, the system integrates search, selection, access, organisation and evaluation of the medical literature that is used as part of the guidelines development. It also provides a collaborative environment for preparing draft guidelines.

The platform, content and tools can be used for free by any professional interested upon request. The system requires minimal IT knowledge and configuration, and relies on common tools such as Google Sites, MS Excel and citation management tools (e.g. Mendeley, Redcube). As an online platform, it does not require server, data storage, additional payable software or subscription, ensuring in this way, its long-term sustainability.
UTILISING STRATIFIED MEDICINE DATA FOR RESEARCH: A NATIONAL APPROACH TO COLLECTION AND INTEGRATION OF MOLECULAR AND NHS CLINICAL DATA

Authors: E Shaw1, A Tuff-Lacey1, S Harris2, M Jones3, Jim Davies2, C Lindsay1, R Cummings1, I Walker1, P Johnson1
1Cancer Research UK, 2Oxford University, 3NHS

Presenting Author: E Shaw

Background: The CRUK Stratified Medicine Programme aimed to address the challenges of implementing stratified medicine into the NHS. Working with 28 hospitals and 3 laboratories across the UK, the programme tested large volumes of samples for multiple markers and captured the associated molecular and clinical data for use by researchers.

Method: The clinical request and molecular information was transferred in XML (extensible markup language) format via a secure file transfer protocol. Upon receipt of the molecular results, 100 data items were added including demographic, treatment and outcome data using the Cancer Outcomes and Services dataset information standard. For the molecular data, Human Genome Variation Society approved nomenclature was agreed as standard. Sites chose a combination of automated and manual data collection approaches. Monthly data uploads were sent to the National Cancer Registration Service – Eastern Office.

Results: Over 8000 unique datasets were collected from the clinical records of consented patients. Data quality and completeness improved through regular analysis, reporting, feedback and resolution of issues. Although defined attributes were set, the data proved heterogeneous between individual patients and geographical sites, representing the use of different terminology and coding systems. For example, in the first data review of 4,204 records for TNM (tumour, nodes metastasis) staging, 336 different forms of staging data were recorded and of 13 possible attributes for Cancer Imaging Modality, 37 variations were observed. Analysis of the molecular data showed variation in how mutations were described complicating aggregated analysis and requiring manual curation. Despite their key role in multidisciplinary team discussion, stage of disease, performance status and co-morbidity scoring were challenging to obtain as they were not captured in existing systems or stored as blocks of texts. Integrated TNM staging had 40.6% completeness across the programme, 99.8% in a site with dedicated manual resources and 32.8% at a site with an automated approach.

Conclusion: Use of and adherence to information standards, application of validation rules and regular review of the data is crucial in providing high quality data amenable to aggregated analysis in a combined dataset. Close relationships with clinical teams is critical in identifying data sources and providing clarity on relevant data items for research. Fully integrated electronic patient records provide an effective, scalable and cost-effective mechanism of data capture but until this is achieved, dedicated data management staff are essential to source and collate data from different systems. Alternative approaches being explored include use of natural language processing.
BRISSKit offers a suite of integrated web applications to accelerate research by facilitating cohort discovery, manage samples, configure patient questionnaires and combine or query on all of these and link to external datasets and NHS data. This makes it easier for researchers to manage the identification, selection, engagement and recruitment of suitable subjects for research. It has developed a shared service solution which can be hosted in a local or cloud type data centre and is being considered for use in the new Jisc brokered Shared Data Centre.

BRISSkit has been developed at the University of Leicester since 2011 with funding from Jisc and users. It aims to:

- Deliver a suite of useful and relevant research applications, as a cloud-based service, accessible at any UK research institution, and meeting the needs of researchers across the fields of biomedicine and translational research.
- Bridge the divide between biomedical research and healthcare by providing secure, reliable and affordable open source database hosting solutions in-house or as a national cloud based service.
- Support the management and integration of tissue samples with clinical data and electronic patient records, having appropriate information governance and anonymisation where required.
- Accelerate translational research in the UK by providing a software service available on a low-cost pay-as-you-go pricing model which would otherwise costs many tens or hundreds of thousands of pounds to replicate within a research institution.
- Deliver state-of-the-art functionality with a robust and reliable service model, ensuring researchers meet their Information Governance, Data Protection and Confidentiality requirements without sacrificing research objectives.
- Facilitate the use of open standards, open source software and the sharing of research data and workflows, allowing collaboration by the widespread use of common tools and the appropriate use of data sharing and anonymisation technologies.

The service is unique in offering:

- fully UK customised versions of mature open source applications used more widely internationally including: CiviCRM, Onyx/RedCap, i2b2 and OpenSpecimen (formerly caTissue)
- seamless integration between applications through well defined use cases
- easy set up through browser access to cloud configuration, capable of being hosted in any compliant cloud provider
- API and Integration tools available for integration with clinical systems
- service tailoring to fit the needs of many research groups and projects, promising significant savings in research database and IT support costs.

We describe the advantages for researchers, research groups and institutions with examples of use of such integrated tools in research at University Hospitals Leicester NHS Trust.
Abstract 1771 SECURE ANONYMIZED DATA LINKAGE – AN ON-GOING COLLABORATION BETWEEN THE INSTITUTE FOR CLINICAL EVALUATIVE SCIENCES (ICES) AND THE ONTARIO BRAIN INSTITUTE (OBI)

Authors: Conrad Pow1, Luca Pisterzi1, Moyez Dharsee2, Khaled El-Emam3, Karey Iron1, Matthew Tucciarone3, Kenneth R. Evans2
1Institute for Clinical Evaluative Sciences. 2Indoc Research. 3Privacy Analytics

Presenting Author: Conrad Pow

Corresponding author email: conrad.pow@ices.on.ca

Purpose: The purpose of this study was to implement an approach to securely link anonymized clinical brain research data with anonymized health administrative data held at ICES.

Background: Securely linking clinical research data with administrative data can provide valuable evidence for clinical and policy decision making. In Ontario, ICES holds linkable health administrative data associated with each person eligible for universal health care since 1991 (13.5 million people in 2014 alone). The Ontario Brain Institute (OBI) is an organization mandated to promote brain research. OBI-funded project data are encrypted and stored on a tiered, open access informatics platform known as Brain-CODE. ICES and OBI are collaborating to implement a privacy preserving protocol to link two anonymized datasets to enhance brain research.

Methods: Privacy Analytics, a privacy software vendor in Ottawa, Ontario, was engaged to provide software that matches encrypted identifiers between data files, and to serve as a semi-trusted third-party required by the privacy preserving protocol. To test software integrity and performance, ICES created simulated data relating to one million individuals. Small (3,000) and medium (30,000) sub-cohorts were randomly selected and encrypted from the larger sample to measure processing time for both small and medium-sized linkages. Following software testing, a data flow and governance framework was developed by OBI and ICES: (1) Application, authorization and approvals for projects; (2) Dataset creation and application of encryption software; (3) Comparison of encrypted identifiers; (4) Transfer of encrypted comparison results to semi-trusted third party; (5) Matching of records and transfer of results to ICES and OBI; (6) Assembly of linked research dataset based on match results; (7) Risk of re-identification assessment and further de-identification of individual records; (8) Transfer of final dataset to secure workspace accessible by researchers.

Results: The software matching rate was 100% using simulated data. This excellent result was not surprising as ICES acted as all three parties in the process. The proposed data flow and governance structure will be tested using actual research data stored in Brain-CODE which is REB-approved for this purpose.

Conclusions: The successful completion of this project will provide a flexible and effective platform and process in Ontario for linking anonymous clinical research data to Ontario administrative data held at ICES. Future plans include the use of the platform for linking anonymous non-health data, such as education or correction data to health data to evaluate areas such as mental health and addictions services, for example.
Abstract 1477

A BRAIN IMAGING REPOSITORY OF NORMAL SUBJECTS ACROSS THE LIFE COURSE: BRAIN IMAGES OF NORMAL SUBJECTS (BRAINS)

Authors: Dominic Job, David Rodriguez, Susan Shenkin, Andrew Robson, Sammy Danso, Joanna Wardlaw

Brain Research Imaging Centre (BRIC), & Centre for Clinical Brain Sciences (CCBS), The University of Edinburgh.

Presenting Author: Dominic Job

Corresponding author email: dominic.job@ed.ac.uk

The Brain Images of Normal Subjects (BRAINS) Imagebank is an integrated repository project sponsored by the University of Edinburgh and the Scottish Imaging Network: A Platform for Scientific Excellence (SINAPSE) collaborators.

The purpose of BRAINS is to provide sharing and archiving of detailed normal human brain imaging and phenotypic data, to create better estimates of the range of normal brain size and integrity across the life-course.

The definition of ‘normal’ is not simple and therefore this Imagebank can be searched by a range of linked data, such as gestational age at birth, blood pressure, medications, other risk factors, and several MRI sequences, including T1, T2, T2*, FLAIR, and DTI.

The availability of clinically relevant MRI sequences from healthy volunteers across the life-course, linked with related phenotypical, demographic and cognitive measures, without diagnosed disease is an essential resource:

(i) As a reference atlas, for interpretation of brain images in clinical diagnosis, such as having access to healthy subject reference images and linked data closely matched to a patient’s scan, to improve diagnostic accuracy (Farrell, C. et al., 2009)

(ii) For the biomedical research community to develop and test new methods, e.g. machine learning, to detect brain pathology and associated clinical manifestations, e.g.,

1. Early markers of neurodevelopmental impairment or dementia.

2. Precise estimates of disease risk.


BRAINS is a living Imagebank where new data will be added. Initially BRAINS will contain existing data from n=763 healthy volunteer subjects (0-81 years of age) from projects in 3 centres. A further n=2119 subjects (prenatal to 90 years) from 15 other projects in Scotland are currently being added.
GEOGRAPHIC VARIATION OF INPATIENT CARE COSTS AT THE END OF LIFE

Authors: Claudia Geue¹, Olivia Wu¹, Jim Lewsey¹, Alastair Leyland², Terence Quinn¹
¹University of Glasgow, ²MRC/CSO Social and Public Health Sciences Unit, University of Glasgow

Presenting Author: Claudia Geue
Corresponding author email: Claudia.Geue@glasgow.ac.uk

Background: Costs incurred at the end of life are a main contributor to health care expenditure (HCE). Geographical inequalities in health outcomes, driven by distance to hospital have been demonstrated for various medical conditions. Issues around geographical patterning of the association between remaining time to death (TTD), age and HCE remain under-researched. It is not known whether these differences in outcomes translate into differences in costs at the end of life. In fact, two scenarios are possible: i) increased costs for inpatient care due to remoteness and related challenges (e.g. difficulties in discharge planning due to increased length of stay), or ii) reduced costs in the context of increased early mortality (e.g. patients not hospitalised in time to receive emergency care). Both situations are important but the health care and policy implications would be very different.

Methods: We used a large representative sample of the Scottish population obtained from death records and linked to acute inpatient care episodes. We performed retrospective analyses of costs and recorded the most frequent reasons for the last admission to hospital. Using a two-part model, we first estimated the probability of incurring any costs; i.e. utilising healthcare services. In a second step we estimated costs for those patients who incurred costs. The estimation was carried out separately for each of the identified admission reasons.

Results: After adjusting for age, sex, year of admission, social deprivation (SIMD) and remaining TTD the overall effects of urban – rural area classification on costs for the last admission to hospital were similar across disease areas. Overall, patients admitted from remote and very remote areas incurred significantly higher costs than patients from large, urban areas after controlling for age, sex, socio-economic status and TTD. We found costs for ‘acute myocardial infarction’ to be significantly higher in very remote, rural areas (£2,706) compared to large urban areas (£2,403). Similar results were also found for ‘fracture of femur’ (£2,479 in very remote, rural areas versus £1,939 in large, urban areas).

Conclusions: Our results provide further evidence of the additional costs associated with remote locations and interventions to reduce these costs might be called for. If inequalities are driven by the hospital admission, then for an end of life scenario, care delivered closer to home would seem intuitively attractive and potentially cost-saving.
Abstract 1612

THE MOLECULAR EPIDEMIOLOGY OF CLOSTRIDIUM DIFFICILE IN SCOTLAND

Authors:

Nicosha De Souza\textsuperscript{1}, Charis Marwick\textsuperscript{1}, Gillian Douce\textsuperscript{2}, Peter Donnan\textsuperscript{1}, Derek Brown\textsuperscript{3}, Cosmika Goswami\textsuperscript{2}

\textsuperscript{1}University of Dundee, \textsuperscript{2}University of Glasgow, \textsuperscript{3}NHS Greater Glasgow & Clyde

Presenting Author: Nicosha De Souza

Corresponding author email: n.desouza@dundee.ac.uk

Clostridium difficile infection (CDI) is the most common cause of antibiotic associated diarrhoea and healthcare associated infection in the developed world, linked to mortality and morbidity affecting mostly the elderly. In this study we are using novel data linkage to investigate the molecular epidemiology of \textit{C. difficile} in Scotland by studying the relationship between community and hospital associated strains, characterising the most common strain (ribotype 078) in Scottish hospitals and to develop a unique electronic resource to link patient healthcare and whole genome sequences.

The study sample includes 500 patients with \textit{C. difficile} strains in the Scottish Microbiology Reference Laboratory from four defined cohorts - a historical Tayside community-associated CDI study, a current Health Protection Scotland community-associated CDI surveillance programme (involving six health boards from across Scotland), a cohort with hospital-associated CDI from the same health boards as the current HPS community cohort, and a cohort with ribotype (078) CDI from across Scotland.

We are using anonymised record linkage between \textit{C. difficile} bacterial genotypic data, sequenced by the University of Glasgow, and health informatics data on the patients who had infection with the sequenced \textit{C. difficile} strains. These include data from Scottish Morbidity Records (SMR01, SMR02), General Registrar Office (GRO), demography (CHI) and prescribing (PIS) datasets in addition to bespoke study datasets. SMR data contains admission and discharge dates, hospital facility and up to six medical conditions relevant to that admission. GRO contains date and cause of death.

The data from multiple sources are being linked by the Health Informatics Centre (HIC) at the University of Dundee, using each patient’s unique community health index (CHI). All data are anonymised by HIC within a safe haven and made available to the authorised researchers for analyses. The HIC safe haven provides a secure environment for analyses and facilities for data sharing among approved researchers. This project has been approved by the National Caldicott Guardian, NSS Privacy Advisory Committee (PAC), Tayside Caldicott Guardian, and the CHI Advisory Group (CHIAG).

The linked data will allow us to study the interaction between the epidemiological characteristics of the patient and the genotypic characteristics of the bacteria by applying statistical techniques such as chi-squared test, temporo-spatial clustering, Cox proportional hazard and logistic regression models. The findings will inform on factors that may influence future laboratory testing, patient treatment and infection control, for example antibiotic resistance and virulence.
SPACE-TIME STATISTICAL ANALYSIS OF INFECTIONS WITH TOXOPLASMA GONDII IN PETS IN THE UK

Abstract 1652

Authors: Fernando Sánchez-Vizcaíno¹, S Dawson², R.m. Gaskell¹, P.h. Jones¹, T Menacere¹, P.-J.M. Noble², M Wardeh¹, A.d. Radford¹
¹University of Liverpool, Institute of Infection and Global Health, Leahurst Campus, Chester High Road, Neston, UK, ²University of Liverpool, School of Veterinary Science, Leahurst Campus, Chester High Road, Neston, UK

Presenting Author: Fernando Sánchez-Vizcaíno

Corresponding author email: fsvb@liv.ac.uk

Introduction: Toxoplasmosis is a zoonotic disease caused by the protozoan parasite Toxoplasma gondii with cats being the definitive host. Recently, efforts have been made to enhance the surveillance system for human toxoplasmosis in the UK. However, companion animals lack coordinated toxoplasmosis national surveillance. Here we use data from commercial diagnostic laboratories, gathered by SAVSNET – The Small Animal Veterinary Surveillance Network to estimate the proportion and spatial distribution of positive T. gondii samples in dogs and cats. We also investigate both space-time clusters of infections with T. gondii and samples in which a diagnosis was not reached.

Material and Methods: The study was performed using results of immunofluorescence assays conducted between January 2012 and May 2014. A retrospective space-time permutation scan statistic model which requires only case data was used to investigate clusters of T. gondii infections. Clusters of samples tested negative were investigated using a space-time Bernoulli (i.e. case/control study) scan statistic model. The maximum size of the spatial and temporal windows for the clusters was assumed to be 50% of the study area and study period, respectively. Models were run 999 times using a Monte Carlo sampling method using SaTScan software version 9.2. Results were mapped using Quantum GIS (version 1.8.0-Lisboa).

Results: In dogs and cats, the mean percentage of samples testing positive to T. gondii was 9.0% (95% CI = 8.1% – 9.9%) and 12% (95% CI = 10.3% – 13.9%) respectively. Samples for T. gondii testing were submitted to the labs from 30 (24.8%) UK postcode areas. Edinburgh was the area where the highest percentage of samples tested positive (mean = 15.5%, 95% CI = 11.4% – 20.7%). However, no significant (P<0.05) space-time infection clusters were identified. Significant space-time clusters of negative samples were located: in dogs in northern and southern regions of the UK during the first half of the year 2012, and in cats in the southeast of the country between April and October 2012.

Conclusions: This is the first time lab-based surveillance has been used to generate national maps of distribution for T. gondii infection in dogs and cats. In the future, any significant hotspot of infection in cats may be used to target human health interventions. In addition, identifying clusters of samples in which a diagnosis is not reached could help identify the emergence of new strains or new diseases over space and time.
A privacy-preserving technique to allow distance comparison of anonymised geospatial location information

This presentation outlines a geometric and computational technique which allows spatial information to be encoded in a form which avoids the use of explicit coordinates while still allowing two encodings to be compared to ascertain their separation distance without revealing the location of either. The encodings contain no explicit location information. The comparison is computationally inexpensive and may be tailored to an arbitrary level of accuracy by increasing the size of the encoding. The encoding is highly privacy-preserving as any individual encoded location is essentially a collection of random numbers which reveal nothing. Once the data is encoded the encoding parameters can be deleted, as no decryption is necessary to ascertain distance computation. The method trades off a configurable desired level of accuracy against the length of the encoding, i.e. any desired level of accuracy may be achieved provided one does not mind the length of the encoding.

Analysis shows that using modest sized encodings nearly all computed distances are within ±10% of the true result and 70% of distances lie within 5% of the true result. The mean absolute relative error can be made to approach 1%. The method is a new tool for privacy preserving data handling of spatial information and is highly configurable in terms of both the maximum separation distance computable and the desired size of encodings. The technique extends to other spatial feature computations such as ‘distance from a road’ or ‘area of overlap’ and extends to the inclusion of other geospatial identifiers and even arbitrary information for comparison. The approach is suitable for calculations on information, e.g. address information, where individual locations must not be readily identifiable for privacy reasons but where records may need to be compared to obtain their distance from one another or from other geospatial features and service locations, e.g. for computing distance of place of residence from the nearest hospital. The privacy implications of this method are discussed as well as some conclusions able to be drawn about the privacy implications of any method which provides comparison/similarity data at a record level. The proposed method is currently one of the two available privacy preserving technique in public use which offers high precision on small distances.
SESSION F4 – LINKED OBSERVATION FOR HEALTH POLICY.
CHAIR: DR CRAOLINE GAO, TURNING POINT EASTERN HEALTH

Abstract 1534

TURNING COST RE-IMBURSEMENT INTO LIFE SAVINGS: DERIVING EHRS FROM CHINA’S NATIONAL HEALTH INSURANCE SYSTEM TO CREATE OUTCOME DATA FOR THE CHINA KADOORIE BIOBANK

Authors: Sam Sansome
University of Oxford

Presenting Author: Sam Sansome

Corresponding author email: sam.sansome@ndph.ox.ac.uk

Background: The China Kadoorie Biobank Study (www.ckbiobank.org) provides a uniquely rich and powerful resource for investigating environmental and genetic determinants of chronic disease in the Chinese population. It is a long term prospective cohort study of 513,000 men and women aged 30-79 years recruited from ten geographically diverse urban and rural regions in China. Participants are actively followed-up through established morbidity and mortality registries, and by linkage to the national health insurance system to identify disease events.

In this session I share some of the challenges we have encountered as a large-scale epidemiological study in deriving Electronic Health Records (EHR) for participants, from the Chinese national health insurance system (designed to enable cost re-imbursement), and discuss solutions for overcoming them.

Methods: Heterogeneous national health insurance datasets from multiple agencies across the ten study regions are collated and sent to Oxford by colleagues in China. At Oxford we begin the complex challenge of transforming the data into hospital events through manual and automated processes. Original disease descriptions are supplied in Mandarin. These are standardised by our clinicians to ICD10. We have developed an internal lookup library to match original disease descriptions to descriptions already standardised to ICD10. Algorithms are designed to handle variability of supplied admission and discharge coding information including dates.

Linkage from agency data is by supplied National ID (unique) or health insurance number (non unique) to our internal Kadoorie Study id via gathered matching lists and personal information updates. We preserve, as far as possible, the true state of the data for the end user.

Results: To date we have imported over 73 million rows of data and derived 1.6 million ‘health’ events from the Chinese health insurance data and linked these to study participants. We have successfully standardised all Cancer, Stroke, Diabetes, IHD and COPD to ICD10 and plan to include many more diseases of interest such as CKD and Pulmonary heart disease.

Challenges that we have overcome include cultural, language, regional and process driven influences when developing a methodology to transform, standardise and link these data to provide a single dataset for the benefit of our research community.

Conclusion: Data from the different health insurance agencies provides an invaluable source of disease outcome data to enhance our study. However, in transforming these data we have to identify, document and resolve many notable challenges that each agency presents.
Abstract 1556

**HUMAN RESOURCE SYSTEMS IN HEALTHCARE: A SYSTEMATIC SCOPING REVIEW**

**Authors:**

Aizhan Tursunbayeva¹, Claudia Pagliari², Raluca Bunduchi³, Massimo Franco⁴

¹University of Molise/University of Edinburgh/The Farr Institute of Health Informatics Research, ²eHealth Interdisciplinary Research Group, Centre for Population Health Sciences, University of Edinburgh, ³Business School, University of Edinburgh, ⁴University of Molise

**Presenting Author:** Aizhan Tursunbayeva

**Corresponding author email:** aizhan.tursunbayeva@gmail.com

The public health sector is complex and while it must inevitably respond to treatment advances it can be slow to adopt other types of innovation¹ and has lagged behind other sectors in its use of Information and Communication Technologies (ICT)². Health organizations are characterized by a dual structure: one for clinical functions and the other for business and support functions, which is reflected in their ICT³. Despite the importance of business and support systems in health organizations very little research on their adoption or impacts exists, compared with other areas in health informatics and eHealth⁴. Human resource management information systems (HRIS) are vital for the effective running of health systems and address many of the information, communication, and training issues of health professionals as well as providing human resource managers with high-quality data⁵,⁶, however they have not been well studied⁷. This is despite the fact that ‘people costs’ normally constitute about 65-80% of the organization’s total cost and successful implementation of HRIS in HR departments has been linked to improvements in patient care⁸.

This systematic review will search academic and grey literature, in order to identify and classify the existing evidence on HRIS adoption, implementation and impact in health organizations worldwide. The topic lies at the intersection of Informatics, Management and Health, and for this reason the review will draw on a range of sources across the information systems, medical, and business literature. As well as documenting and classifying the research, it will seek to understand the approaches to, barriers for, outcomes and implications of HRIS implementation in health organizations, the theoretical frameworks used to study them, and to identify unexplored areas. The review also aims to support health leaders by summarizing the current evidence-base and informing recommendations on HRIS implementation and research.

Abstract 1573

**LINKING NATIONAL DATABASES TO STUDY VARIATION IN GENERAL PRACTICE MEDICINES COSTS FOLLOWING STANDARDISATION**

**Authors:**

Susan Dovey¹, AM Tomlin², MW Tilyard³

¹University of Otago, ²BPACNZ, ³University of Otago and BPACNZ

**Presenting Author:** Susan Dovey

**Corresponding author email:** susan.dovey@otago.ac.nz

**Background:** In New Zealand 80% of pharmaceutical costs are generated outside hospitals so government programmes to promote efficient drug use are focussed on general practices. Practices with high drug costs tend to be targeted above other practices, but there is poorly understanding of the reasons for practice variation in drug costs. We aimed to establish practice prescribing after standardisation to account for patient demographics and morbidity and to investigate reasons for high- and low-cost drug use in general practices.

**Methods:** We analysed data from two national data collections administered by the New Zealand Ministry of Health: the Pharmaceutical Collection, containing records of all publicly subsidised...
medicines dispensed in community pharmacies, and the Primary Health Organisation Enrolment Collection containing records of patients registered in general practices. We extracted dispensed medicines data for 1045 general practices in 2011, and 917 practices continuously existing 2008-2011. For each practice in each study year, using indirect standardisation we calculated a standardised prescribing cost ratio (SPR: the ratio of actual: expected prescription costs). Case studies of 5 outlier practices explored reasons for their status.

**Results:** SPRs ranged from 0.53 to 2.28 (median = 0.98). Of 469 practices with higher than expected costs (SPR>1.0) in 2011, 204 (43.5%) had a single medicine or therapeutic drug class accounting for more than 15% of total costs. Case studies showed that high-cost practices had overall pharmaceutical expenditure influenced strongly by a few patients needing high cost medicines, or more patients using medicines in one high cost therapeutic drug class (antiretrovirals), or high medicines use across all therapeutic drug classes. Practices with lower than expected costs were characterised by low prescription costs per patient across most therapeutic groups and usually a low prescription rate per patient.

**Conclusions:** Routine data collections can measure inter-practice variation in prescription costs, adjusted for differences in the demography and morbidity profile of each practice’s patients, allowing better targeting of interventions to modify prescribing behaviour. Small groups of patients using high cost medicines influence general practices’ expenditure on pharmaceuticals.

---

**Abstract 1675**

**ARE WAITING TIMES AND LENGTH OF STAY CONNECTED? EVIDENCE FROM THE ENGLISH NHS**

**Authors:**

Arthur Sinko¹, S Nikolova²

¹University of Manchester, ²University of Leeds

**Presenting Author:**

Arthur Sinko

**Corresponding author email:**

arthur.sinko@manchester.ac.uk

**Background:** The UK Government has implemented maximum waiting time reforms to tackle long waiting times. Recent research has found that hospitals changed patient ordering for treatment in response to these reforms. Queueing theory suggests that prioritisation based on hospital length of stay can significantly decrease average waiting times.

**Aim:** To examine the relationship between waiting times and hospital length of stay and how this changes over the 1998—2008 period of progressively increasing waiting time targets and their brief suspension in 2010.


**Data:** Hospital Episode Statistics for 1998-2011 (N=16,979,542).

**Results:** Preliminary OLS results show that there is a statistically significant relationship between length of stay and waiting time for all but two disease categories, however the direction of the effect differs. For cancer diseases we find that, prior to the reforms, patients who stayed longer in hospital waited less for treatment. This pattern was reversed by 2008. For cardiovascular patients we find no such change in pattern; however the positive effect is three times larger in 2008 compared to 1999. In 2011 cancer patients who stayed longer in hospital appear to have shorter waiting times again. For cardiovascular patients the positive coefficient decreased by a factor of 35. These results suggest that shifting the focus away from government enforced waiting time targets brought back the initial relationship between waiting time and length of stay variables.
HOW CAN RESEARCH AND CLINICAL DATA LINKAGE HELP INFORM HEALTH SERVICE PLANNING? INVESTIGATING VARIATION IN GP CONSULTATION RATES IN THE BORN IN BRADFORD COHORT.

Authors: B Kelly1, D Mason1, E S Petherick1, S Prady2, C Bates3, J Wright1, M A Mohammed4
1Bradford Institute for Health Research, 2Health Sciences, University of York, 3The Phoenix Partnership (TPP), 4School of Health Studies, University of Bradford

Abstract 1734

Presenting Author: B Kelly

Corresponding author email: brian.kelly@bthft.nhs.uk

Background: General practice services, particularly in deprived urban communities, are increasingly under strain. General practice consultation rates are known to be relatively high for women of child bearing age, young children and those with higher levels of material deprivation. There is mixed evidence regarding rates for different ethnic groups, and it is uncertain whether variations in general practice consultation rates for different socio-economic and ethnic groups persist after controlling for differences in underlying ill health, and other potential confounding variables.

Methods: The Born in Bradford cohort study recruited 12,450 pregnant women between March 2007 and December 2010, and the 13,857 children born to these women. Around a half of the cohort are of South Asian origin and many individuals have high levels of material disadvantage, providing data well suited to exploring variation in general practice consultation rates for these groups. A range of study data have been collected for cohort members. In addition primary care records were obtained from SystmOne, successfully matching 99.1% of cohort members on NHS number, date of birth, sex and surname. Each cohort member has around 95 percent of their time in the study that is linked to general practice records. Consultation rates per person year are derived, and regression models are employed to determine incidence rates for different groups after controlling for confounding variables.

Results: Preliminary results suggest that general practice consultation rates are higher in the years following birth for children, and around the time of pregnancy for mothers. Mothers and children from a South Asian origin were found to have consultation rates around 1.3 times higher than White British groups, after controlling for levels of material deprivation. Individual and neighbourhood deprivation were associated with increased consultation rates for mothers (around 1.4 times higher for the most deprived, compared to the least deprived), but this relationship was less apparent for children. Ongoing analysis is looking to determine whether such differences remain after controlling for other individual factors, including underlying levels of ill health. Full results from the analysis will be presented.

Conclusions: Linking detailed cohort information with primary care records provides rich contextual data, not available to studies using electronic healthcare records alone. This has facilitated an investigation of variation in the rate of general practice consultations, after controlling for potentially confounders. This provides valuable insight into the implications for patient demand and the provision of general practice services.
IDENTIFYING THE TRUE COST OF ASTHMA IN THE UK: SECONDARY ANALYSES OF NATIONAL STAND-ALONE AND LINKED DATABASES IN ENGLAND, NORTHERN IRELAND, SCOTLAND AND WALES

Authors:
M Mukherjee1, R Gupta2, A Farr3, B Nwaru4, A Stoddart4, M Heaven5, A Bandhopadhyay5, M Shields6, D Strachan2, G Davies7, R Lyons5, C Dibben8, B Mc1, A Sheikh9
1Edinburgh Health Services Research Unit, Usher Institute of Population Health Sciences and Informatics, The University of Edinburgh, Edinburgh, UK, 2Population Health Research Institute, St George’s, University of London, London, UK, 3Swansea Centre for Health Economics (SCHE), College of Human and Health Science, Swansea University, Swansea, UK, 4Usher Institute of Population Health Sciences and Informatics, The University of Edinburgh, Edinburgh, UK, 5CIPHER—Centre for the Improvement of Population Health through e-Records Research, Centre for Health Information, Research and Evaluation (CHIRAL), College of Medicine, Institute of Life Science 2 (ILS2), Swansea University, Swansea, UK, 6Centre for Infection and Immunity, School of Medicine, Dentistry and Biomedical Sciences, Queen’s University Belfast, Belfast, UK, 7Asthma & Allergy Group, Institute of Life Science, College of Medicine, Swansea University, Swansea, UK, 8Department of Geography & Sustainable Development, School of Geography & Geosciences, The University of Edinburgh, Edinburgh, UK, 9Centre of Medical Informatics, Usher Institute of Population Health Sciences and Informatics, The University of Edinburgh, Edinburgh, UK

Presenting Author: M Mukherjee
Corresponding author email: mome.mukherjee@ed.ac.uk

Introduction: Asthma is now one of the most common long-term conditions in the UK. It is therefore important to develop a comprehensive appreciation of the healthcare and societal costs in order to inform decisions on care provision and planning. Commissioned by Asthma UK, we built on our earlier estimates of national prevalence and costs from asthma by filling the data gaps previously identified in relation to healthcare and broadening the field of enquiry to include societal costs. This work will provide the first UK-wide estimates of the costs of asthma. In the context of asthma for the UK and its member countries (ie, England, Northern Ireland, Scotland and Wales), we sought to: (1) produce a detailed overview of estimates of incidence, prevalence and healthcare utilisation; (2) estimate health and societal costs; (3) identify any remaining information gaps and explore the feasibility of filling these and (4) provide insights into future research that has the potential to inform changes in policy leading to the provision of more cost-effective care.

Methods and analysis: Secondary analyses of both stand-alone and linked datasets from national health surveys, general practice, prescribing, emergency care, out of hours, ambulance service, hospital, mortality and administrative data sources, were undertaken across the four countries to estimate prevalence, healthcare utilisation and outcomes from asthma between 2001-12 to 2011-12. Data linkages and economic modelling were undertaken in an attempt to populate data gaps and estimate costs. Separate prevalence and cost estimates were calculated for each of the UK-member countries and these were then aggregated to generate UK-wide estimates. The study is in the final stage and the presentation will be on the results of the study on

1) Epidemiology - incidence; lifetime and annual prevalence of patient reported or GP reported asthma
2) Health care utilisation in primary care - GP and nurse consultations; prescriptions, out-of-hours
3) Healthcare utilisation in secondary care - outpatient clinics; ambulance service; accident and emergency attendances; inpatient and daycases in hospital, intensive care
4) Wider societal impact - school absenteeism; work absenteeism; disability living allowance; premature retirement; mortality
5) Data gaps
6) Financial cost of asthma in 2011-12 for England, Northern Ireland, Scotland and Wales and for whole UK. We have also produced interactive maps. An UK Asthma Observatory is getting created using these. Our exercise can lead to important bearings for mapping the epidemiology, disease burden and costs of other long-term conditions.

Abstract 1780

PREDICTORS OF FREQUENT EMERGENCY DEPARTMENT AND AOD TREATMENT SERVICE UTILIZATION AMONG TREATMENT SEEKING POPULATION

Authors: Caroline Gao, Sharon Matthews, Belinda Lloyd
Turning Point, Eastern Health and Monash University, Melbourne, Australia

Presenting Author: Caroline Gao

Corresponding author email: carolineg@turningpoint.org.au

Objective: Clinical studies suggested that alcohol and other drugs (AOD) using populations are at higher rates of health care service utilization due to their drug using behaviours and unmet needs for treatment. Data was linked to examine patterns of emergency department (ED) visits and AOD treatment episodes among the treatment seeking population and to explore possible reasons for frequent re-admission.

Method: A retrospective cohort of clients with AOD treatment records in Victoria Australia between July 2009 and Jun 2012 were extracted and linked with ED data. The cohort was divided into four subgroups: G1 – non-frequent users; G2 – non-frequent ED and frequent treatment users; G3 – frequent ED and non-frequent treatment users and G4 – frequent users. Frequent service users were defined as those with service use higher than 95% of the cohort population. Reasons for ED visits were investigated for each group and multinomial logistic regression was used for identifying predictors of different service using patterns.

Results: The mean number of AOD treatment episodes was 2.6 (range 1-71) in this three year period. 44% of clients visited ED (mean: 1.7; range: 0-377). Leading causes of ED visits among G1 clients and G2 clients were similar, although with a higher proportion of AOD intoxication among G2 clients (9.6% in G2 vs 4.4% in G1). Proportions of ED visits due to mental health disorders among frequent ED users were substantially higher (10.2% in G3 and 10.8% in G4 vs 5.8% in G1 and 6.7% in G2). Proportion of ED visits due to AOD intoxication was highest among G4 clients (25%). Regression analysis suggests that frequent use of either or both of these services is associated with being female, not being employed, being homeless, recent injecting, and higher proportion of ED visits for AOD intoxication. Being treated for alcohol alone or multiple types of drugs is a significant predictor for frequent use of AOD treatment services, and higher proportion of ED visits for mental health disorders is a significant predictor for frequent ED use.

Conclusions: The study suggested that increasing health care service use among the treatment seeking population is associated with severity of drug using problems, social disadvantage as well as mental health status. This highlights the need for service providers to change from traditional discrete and activity-based service model to the model that promote service integration to improve treatment outcomes and reduce avoidable burden on acute health care services.
SESSION F5 – NATURAL EXPERIMENTS.
CHAIR: DR CHRISTINE ROBERTS, UNIVERSITY OF SYDNEY

Abstract 1722 USING LINKED HEALTH AND SOCIAL CARE DATA TO VALIDATE THE INDICATOR OF RELATIVE NEED (IoRN) TOOL

Authors: Anne Canny¹, Peter Knight², Adam Redpath², Miles D Witham¹
¹University of Dundee, ²Joint Improvement Team, Scottish Government

Presenting Author: Anne Canny
Corresponding author email: a.canny@dundee.ac.uk

Background: The Indicator of Relative Need (IoRN) tool is widely used by health and social care services in Scotland to record dependency of clients. More recently, services have expressed interest in using the IoRN to predict events such as death, hospitalisation and care home admission, as well as future care needs. Such uses require external validation of the tool for these purposes, for which the IoRN was not specifically designed. Using linked health and social care data provides a novel method of externally validating the IoRN for these uses.

Methods: Clients aged 65 and over who underwent IoRN assessment by Dundee Social Work department over a 5 year period (2008-2012) were included in this analysis. Routinely collected health and social care data from NHS Tayside and Dundee Social work department were probabilistically linked via the Health Informatics Centre at the University of Dundee. Cox regression analysis was used to test the association between categories of dependency on the IoRN score and risk of death, hospitalisation and care home admission. Analyses were adjusted for age, sex, number and length of stays in hospital during last year to test whether IoRN provided additional predictive ability over and above these variables. Analyses were also performed to test the association between dependency category and the hours of home care per week allocated to clients six months after assessment.

Results: 1732 individuals were included in the analysis; mean age 81 years. 1214 (70%) were female and 144 (8%) died during a mean follow up period of 2.5 years. The adjusted hazard ratio for death in the most dependent category compared to the least dependent category was 5.9 (95%CI, 2.0 to 17.0); for care home admission the hazard ratio was 7.2 (95%CI, 4.4 to 12.0) and for hospital admission the hazard ratio was 1.1 (95%CI, 0.5 to 2.6). The mean number of allocated hours of care 6 months after assessment was higher in the most dependent group compared to the least dependent group (5.6 hrs vs 1.4 hrs, p=0.005)

Conclusion: IoRN category is associated with risk of death, care home admission and future need for care, but does not provide additional information to predict future hospital admission.
ONE CONDUIT OR TWO? ASSOCIATION OF SINGLE AND BILATERAL MAMMARY ARTERY GRAFTS WITH MORTALITY OUTCOMES IN ENGLAND AND WALES, USING ROUTINELY COLLECTED DATA.

Authors: K S L McAllister¹, Graeme L Hickey², Stuart W Grant³, Ben Bridgewater⁴, Iain E Buchan¹
¹Health e-Research Centre, University of Manchester, ²Institute of Infection and Global Health, University of Liverpool, ³Institute for Cardiovascular Sciences, University of Manchester, ⁴University Hospital of South Manchester

Presenting Author: K S L McAllister

Corresponding author email: kate_mcallister@hotmail.co.uk

Aim: To examine the comparative effectiveness of single (SIMA) and bilateral (BIMA) internal mammary artery grafts for coronary revascularisation, using routinely collected data from the UK national audit.

Background: The internal mammary arteries are frequently used during coronary artery bypass grafting (CABG). While grafts from a single mammary artery are common, there is clinical debate as to the relative safety and efficacy of using bilateral mammary artery grafts. We used data from the Society for Cardio-thoracic Surgery (SCTS) audit to examine real-world performance of these procedures. All-cause mortality follow-up data were obtained by linkage to Office for National Statistics mortality records.

Methods and results: Between January 2000 and March 2011 there were 422,293 cardiothoracic procedures recorded in the SCTS database across the UK. Propensity scores were used to create a 1:3 case-mix matched cohort of 6,036 BIMA and 14,999 SIMA cases in England and Wales throughout the period. Multiple imputation was used to address missing data in candidate variables. Logistic regression in the matched cohort showed that BIMA was associated with a higher risk of in-hospital mortality relative to SIMA (adjusted OR: 1.90, 95% CI 1.37 to 2.64, P<0.001). Conversely, a piecewise Cox proportional hazards model of survival indicated that BIMA was associated with better survival rates after 30 days had passed (before 30 days: adjusted HR: 1.55, 95% CI 1.16 to 2.06, P=0.003; after 30 days: adjusted HR: 0.56, 95% CI 0.41 to 0.76, P<0.001). A sensitivity analysis exploring the influence of surgeon experience of BIMA (a binary variable indicating whether the surgeon carried out at least 10% of CABG procedures as BIMA) indicated this was associated with an in-hospital mortality benefit (adjusted OR: 0.42, 95% CI 0.23 to 0.78, P=0.006) but no gain in survival was detected in the long term (P=0.456)

Conclusions: SIMA is associated with a lower risk of in-hospital mortality, but does not offer as great a survival benefit as BIMA. Where BIMA is carried out by surgeons with greater experience of this procedure, the odds of in-hospital mortality may lessen. Routinely collected data can provide evidence on real-world effectiveness of treatments, especially where trial data are lacking. Larger-scale data linkage and model sharing would improve clinical outcome surveillance in examples such as this, particularly where case-mix adjustment is needed.
Abstract 1711

LINKAGE OF A UK NATIONAL CARDIAC REGISTER TO ROUTINE DATA TO DEVELOP EVIDENCE FOR DECISION-MAKING BY NICE – CALON (CARDIAC ABLATION: LINKING OUTCOMES FOR NICE)

Authors:

Ruth Louise Poole¹, Megan Dale¹, Christopher Gale², Antony Wilkes³, Arfon Rees⁴, Ting Wang⁴, Luca Ruschetti⁴, Daniel Thayer⁴, Mauro Lencioni⁵, Hannah Patrick⁶, Grace Carolan-Rees¹

¹Cedar, Cardiff & Vale University Health Board, ²National Institute for Cardiovascular Outcomes Research (NICOR), University College London, ³Cedar Consultant, ⁴Secure Anonymised Information Linkage (SAIL) Databank, College of Medicine, Swansea University, ⁵Queen Elizabeth Hospital, Birmingham, ⁶National Institute for Health and Care Excellence (NICE)

Presenting Author: Ruth Louise Poole

Corresponding author email: Ruth.Poole@wales.nhs.uk

Background: The Interventional Procedures programme at the UK’s National Institute for Health and Care Excellence (NICE) evaluates new and established procedures where there are questions about their efficacy or safety. NICE was interested in whether data linkage might help to address the evidence gaps, with a view to potentially reviewing current guidance.

Objectives:

- To investigate the feasibility and usefulness of linking a register to routinely held patient data.
- To gather evidence on efficacy and safety of cardiac ablation procedures for treatment of arrhythmias.

Methods: Data from the National Institute for Cardiovascular Outcomes Research (NICOR) Heart Rhythm Audit ablation dataset were linked at individual patient level to routine Welsh hospital, primary care, and mortality records within the Secure Anonymised Information Linkage (SAIL) Databank. Proxy measures were used to compare the general wellbeing of patients from two years before, and up to five years after, ablation procedures. Outcomes included number of encounters with primary care, hospital outpatient and inpatient services, as well as drugs prescribed. Mortality rates and incidence of major safety outcomes were calculated. Covariates were age, sex and Charlson comorbidity score. Commonly recorded post-procedural diagnosis codes were examined.

Results: 1057 records from NICOR were matched at a 99.7% success rate. An additional 1163 ablation procedures were identified from hospital OPCS codes, giving a total of 2220 patients. Patients attended 27% fewer outpatient appointments (p<0.001) and were admitted for 22% fewer days (p<0.001) after ablation. There was no significant difference in the number of GP events. Prescribing of antiarrhythmic drugs decreased by 65% (p<0.001) but the number of antidepressants/anxiolytics increased by 31% (p=0.016). The estimated survival rate at 5 years was 91%. Incidence of stroke/silent cerebral embolism, transient ischaemic attack, myocardial infarction and cardiac tamponade at 5 years were 1.7%, 1.3%, 2.1% and 1.8% respectively.

Due to limitations of available data and the study design, it was not possible to differentiate ablation procedures according to the categories used by the NICE IP programme.

Conclusions: A reduction in secondary care service use and anti-arrhythmic drugs prescriptions suggests that patients feel better after undergoing cardiac ablation, though increased prescribing of antidepressants warrants further investigation.

Data linkage can be used to obtain long-term outcomes of interventional procedures using routine data. This study highlighted the benefits of using established repositories of linked data, and the need for appropriate data collection in order to gain maximum value from specialist registers. Linked data has potential to strengthen the evidence base for clinical practice guidelines in the UK.

Abstract 1538

IDENTIFYING COMPLICATIONS FROM TENSION-FREE TRANS-VAGINAL TAPES AND TRANS-OBTURATOR TAPES FOR STRESS URINARY INCONTINENCE; A STUDY OF 8
YEARS OF HES DATA

Authors: Kim Keltie, Helen Cole, Ash Monga, Sohier El Neil, Andrew Sims
1Newcastle upon Tyne Hospitals NHS Foundation Trust, 2University Hospital Southampton NHS Foundation Trust, 3University College London Hospitals NHS Foundation Trust

Presenting Author: Kim Keltie

Corresponding author email: kim.keltie@nuth.nhs.uk

Aim: To identify peri-procedural complications arising from the introduction of trans-obturator tapes (TOT) and tension-free trans-vaginal tapes (TVT) for stress urinary incontinence in a cohort of women, using eight years of routine data extracted from the Hospital Episodes Statistics (HES) data repository.

Method: The HES in-patients "universe" was searched to identify all hospital admissions in England for introductions of TVT, TOT and trans-vaginal slings (OPCS codes M53.3, M53.6 and M52.1) during financial years 2007/8 to 2013/14. Episodes were analysed using the statistical computing language, R. After data cleaning, peri-procedural complications for the index spells, at which TOT or TVT was introduced, were identified from ICD-10 main and supplementary codes.

Results: After cleaning, there were 91,953 episodes involving the introduction of a TVT or TOT for stress urinary incontinence (SUI). Of these, 2538 episodes also involved the introduction of mesh for vaginal wall, vaginal vault prolapse or uterine prolapse and were excluded, leaving 89,415 episodes from 87,591 patients for analysis. In total TVT was introduced in 53,918 patients, TOT in 33,541 and slings in 875. 53,167 women had a TVT and no other tape or mesh procedure during their index admission. Of these, 15,163 received additional therapeutic procedures at their index admission and were excluded from analysis of peri-procedural complications to avoid confounding, leaving 38,004 TVT introduction procedures for analysis, from 173 healthcare providers. The median age was 50 (IQR 44 to 60) and 99.9% of the procedures were elective. A total of 905 complications were identified in 882 episodes (2.3% complication rate).

33,220 women had a TOT and no other tape or mesh procedure during their index admission. Of these, 9650 received additional therapeutic procedures at their index admission and were excluded from the analysis, leaving 23,570 TOT introduction procedures for analysis, from 166 healthcare providers. The median age was 51 (IQR 44 to 61) and 99.9% of the procedures were elective. A total of 250 complications were identified from 241 episodes (1.0% complication rate).

Discussion: Using the ICD-10 coding scheme, which provides a framework for recording procedural complications, lower limits for the rates of peri-procedural complication have been estimated for all tape procedures for stress urinary incontinence in England between 2007 and 2014 as 2.3% for TVT and 1.0% for TOT. Estimation of long-term follow up of these patients using HES records is in progress.
Abstract 1728  

CHILDHOOD HEALTH PROBLEMS FOLLOWING PLANNED CAESAREAN DELIVERY AT TERM: A POPULATION-BASED RETROSPECTIVE COHORT STUDY OF SCOTTISH DATA

Authors:  
Mairead Black1, D McLernon2, J E Norman3, S Bhattacharya4  
1University of Aberdeen, 2Division of Applied Health Sciences, University of Aberdeen, 3Tommy’s Centre for Maternal and Fetal Health, University of Edinburgh, 4Division of Applied Health Sciences, University of Aberdeen

Presenting Author: Mairead Black
Corresponding author email: mairead.black@abdn.ac.uk

Background: Caesarean section (CS) may increase the risk of offspring immune-mediated illness due to lack of exposure to maternal bowel flora during birth. However, planned CS avoids birth-related brain injury, so cerebral palsy and learning disability may be reduced.

Objective: To explore the relationship between planned cesarean section (CS) and offspring health problems by comparing health outcomes (salbutamol inhaler use aged ten years, severe asthma, obesity aged five years, inflammatory bowel disease, type 1 diabetes, cancer, death, learning disability and cerebral palsy) between offspring delivered by planned CS with those delivered following 1) an apparent plan for vaginal birth (actual vaginal birth and emergency CS combined), 2) vaginal birth only or 3) emergency CS only.

Methods: A population-based record linkage study followed up 321,287 first-born term singleton infants delivered in NHS Scotland between 1993 and 2007. Data was obtained from the Scottish Morbidity Records 01 and 02; Child Health Surveillance System; Support Needs System; Prescription Information System; Scottish Care Information Diabetes Collaboration; and National Records for Scotland. Survival analysis and logistic regression models were used to compare outcomes following planned CS with each comparison group. Confounders considered included maternal demographic, clinical and breastfeeding characteristics. Missing values on social class, smoking status and maternal BMI were imputed using multiple imputation.

Results: Compared with the group apparently planning vaginal birth, salbutamol inhaler use aged 10 years, severe asthma, obesity aged five years and learning disability were more likely in offspring delivered by planned CS {adjOR 1.15 (1.03-1.28), adjHR 1.17 (1.07-1.29), adjOR 1.16 (1.01-1.32), adjHR 1.34 (95% CI 1.07-1.68)}, but there was no difference in risk of inflammatory bowel disease, type 1 diabetes, cancer, death or cerebral palsy {adjHR 0.88 (0.51-1.51), adjHR 1.23 (0.97-1.55), adjHR 1.02 (0.70-1.50), adjHR 1.21 (0.90-1.62), adjHR 1.04 (0.40-3.07)}. These risks of planned CS were also evident when compared with actual vaginal birth, with the addition of increased risk of death {adjHR 1.41 (1.05-1.90)}.

Conclusions: Offspring delivered by planned CS are at increased risk of childhood asthma-related illness, obesity and learning disability when compared to those delivered following an apparent plan for vaginal birth. This is consistent with published data on risk associated with any CS delivery compared to vaginal birth. Potential long-term implications of CS delivery should be considered when balancing the risks and benefits of planned pre-labour CS. Future research should explore potential for residual confounding and mechanisms which explain the associations identified.
Abstract 1724

BIRTH AFTER CAESAREAN SECTION AND CHILDHOOD HEALTH PROBLEMS: A POPULATION-BASED RETROSPECTIVE COHORT STUDY OF SCOTTISH DATA

Authors:

Mairead Black¹, D J McLernon², J E Norman³, S Bhattacharya²
¹University of Aberdeen, ²Division of Applied Health Sciences, University of Aberdeen, ³Tommy’s Centre for Maternal and Fetal Health

Presenting Author: Mairead Black
Corresponding author email: mairead.black@abdn.ac.uk

Background: Vaginal birth after cesarean section (VBAC) carries a risk of scar rupture in labour leading to offspring cerebral palsy or death, but planned repeat caesarean section (CS) may compromise offspring immunity due to lack of exposure to maternal bowel flora.

Objective: To explore the relationship between planned repeat CS and offspring risk of: salbutamol inhaler use; severe asthma; obesity aged five years; inflammatory bowel disease; type 1 diabetes; cancer; death; learning disability and cerebral palsy compared with risk following VBAC.

Methods: A population-based record linkage study followed up 40,145 second-born term singleton infants delivered in Scotland between 1993 and 2007 to mothers who had previously delivered by CS. Data was obtained from the following Scottish datasets: Scottish Morbidity Records 01 and 02; Child Health Surveillance System; Support Needs System; Prescription Information System, Scottish Care Information Diabetis Collaboration and National Records Scotland. Outcomes following planned repeat CS were compared with a group delivered following VBAC. Logistic regression models were used where time at risk was fixed, while survival analysis models were used where time at risk varied between cases. Confounders included social, clinical and breastfeeding characteristics. Multiple imputation was used to deal with missing values on social class, smoking status and maternal BMI.

Results: Women who underwent planned CS were older {mean age 31.0yrs (SD5.0) vs 29.6 (5.0)}, less likely to smoke (17.9% vs 21.3%) and more likely to have type 1 diabetes (1.8% vs 0.5%) than those who delivered by VBAC. Offspring delivered by planned CS were delivered earlier {mean gestation 38.8 weeks (SD1.02) vs 39.7 (1.1)}, were heavier {mean birthweight 3504gm (SD507) vs 3438 (469)} and were less likely to be breastfed (32.3% vs 39.0%) than those delivered following VBAC. When compared with offspring born by VBAC, those born by planned repeat CS were at increased risk of salbutamol inhaler use aged five years and asthma requiring hospital admission {adjusted odds ratio 1.19 (95% confidence interval 1.01-1.40), adjusted hazard ratio 1.22 (95% CI 1.00-1.49)}, but were not at significantly different risk of obesity aged five years, inflammatory bowel disease, cancer, death, learning disability or cerebral palsy {adjOR 1.18 (0.97-1.44), HR 0.82 (0.43-1.56), adjHR 0.60 (0.35-1.01), adjHR 1.01 (0.66-1.53), adjHR 1.17 (0.82-1.65), HR 1.17 (0.26-5.23)}.

Conclusions: Offspring delivered following planned repeat CS are at increased risk of asthma-related illness compared to those delivered by VBAC, with no apparent difference in risk of other immunity-related illness or neurological problems.
TRENDS PREGNANCY HYPERTENSION AND THE IMPACT OF INCREASING OBSTETRIC INTERVENTIONS

Authors: Christine L Roberts, Algert Charles, Jillian Patterson, Jonathan Morris, Jane B Ford
Clinical and Population Perinatal Health Research, Kolling Institute, University of Sydney, Australia

Presenting Author: Christine L Roberts

Corresponding author email: clroberts@med.usyd.edu.au

Background: Since the 1990s, pregnancy hypertension rates have declined in some countries (including Scotland, Australia and Sweden) despite increasing rates of established risk factors such as obesity, nulliparity and maternal age. As most pregnancy hypertension occurs from 39 weeks gestation (~60%), the aim of this study was to determine whether increases in early planned deliveries could explain at least part of the decline in pregnancy hypertension.

Method: Data were obtained from linked birth and hospital records for >1.08M deliveries in Australia 2001-2012. Pregnancy hypertension included gestational hypertension, preeclampsia and eclampsia, and planned delivery comprised labour induction and prelabour caesarean section. Trends in pregnancy hypertension and established risk factors (eg nulliparity, maternal age ≥35 years, obesity, multi-fetal pregnancy, chronic diseases) were determined from longitudinally linked records. Multivariable logistic regression predictive models for pregnancy hypertension were developed using the first 2 years of data (2001-2) and the results applied to data for other years to produce predicted trends based on actual changes in established risk factors. Annual gestational age-specific pregnancy hypertension rates were determined among pregnancies-at-risk. Trends in the distribution of gestational age for women with and without pregnancy hypertension were compared. We also examined the interaction between a history of pregnancy hypertension and planned birth on the risk of pregnancy hypertension.

Results: The overall pregnancy hypertension rate declined from 10.3% to 7.7% (P<0.001). Based on changes in the prevalence of maternal risk factors, the pregnancy hypertension rate was predicted to increase to 11.8%. Contemporaneously, increasing planned deliveries led to a decline in deliveries ≥39 weeks (73.7% to 67.9%) and an increase at 37-38 weeks (19.8% to 25.2%). Gestational age-specific pregnancy hypertension rates declined from 38 weeks onwards, most steeply at ≥41 weeks. Births complicated by pregnancy hypertension basically disappeared from the later gestations over the study period, thereby changing the shape of the gestational age distribution. In contrast, for women without pregnancy hypertension the distribution was shifted to the left (to lower gestations) due to planned deliveries. Women who had pregnancy hypertension in a prior pregnancy and a planned birth had lower rates of pregnancy hypertension than women with prior pregnancy hypertension who continued to spontaneous labour.

Conclusions: The rate of pregnancy hypertension has decreased in Australia. An overall increase in early planned delivery can explain much of the observed decline in pregnancy hypertension rates. Women with risk factors for hypertension are being selected for earlier planned delivery.
OVERCOMING THE MOTIVATIONAL BARRIERS TO SHARING HEALTH AND MEDICAL RESEARCH DATA

Authors: Louisa Jorm
Centre for Big Data Research in Health, UNSW Australia

Introduction: Benefits of data sharing include speeding the pace of discovery, ensuring the reproducibility of research, and cost-efficiency. Major international research funders have signed a ‘Joint Statement on Data Sharing of Public Health Research’1 and are enacting policies to mandate or at least encourage data sharing. However, a recent systematic review2 identified a variety of barriers to sharing public health data, including some based on personal or institutional motivations and beliefs. Discussions with Australian health and medical researchers indicate that these ‘motivational’ barriers, rather than technical, legal or ethical ones, are the major deterrents to data sharing by researchers.

Methods: Review of motivational barriers to sharing health and medical research data and of ways that these can be addressed through policy, training and creating new incentives. Identification of case studies that could serve as models of ‘good practice’ in sharing health and medical research data.

Results: Motivational barriers to sharing health and medical research data include: lack of incentives in current academic reward systems; sense of ‘ownership’ of data and intellectual property and related fears that competing researchers will ‘exploit’ the data and gain most of the credit; concerns that the data will be ‘misused’ by researchers who don’t understand its provenance; and unwillingness to face scrutiny and possible criticism.

Ways that these barriers can be addressed include: explicit consideration of track record of data sharing in grant funding, recruitment and promotion processes; policies that require documented datasets to be deposited in public repositories within a specified period of the end of research grants; policies that specify how the original creators of data are to be involved and acknowledged in subsequent research using the data; mechanisms to make data, metadata and code citable (for example in the form of peer-reviewed ‘data papers’); and training researchers in good data management and programming practices so that their research is robust to scrutiny and reproducible. Case studies illustrating each of these approaches will be presented.

Conclusions/implications: Health and medical researchers are lagging behind other disciplines in sharing their data, when they should be at the forefront of this endeavour. Models of good practice exist, but addressing motivational barriers to sharing health and medical research data will require concerted efforts not only by research funders, but by the institutions that employ and train researchers.

References
1 Walport M, Brest P. Lancet 2011;377:537-539.
FIRST NATIONS HEALTH DATA LINKAGE: A COLLABORATIVE INDIGENOUS RESEARCH AND GOVERNANCE APPROACH

Authors: Saba Khan¹, Carmen Jones², Jennifer Walker¹, Tracy Antone², David Henry¹
¹ICES, ²Chiefs of Ontario

Presenting Author: Jennifer Walker

Corresponding author email: saba.khan@ices.on.ca

The history of colonization and cultural assimilation of Indigenous peoples in Canada has negatively influenced the health status of the First Nations population, who report significantly higher rates of chronic disease, infectious disease, and mortality compared to the general Canadian population. Conducting research to improve the health of First Nations requires innovative and collaborative approaches to community involvement and data governance when developing research projects and co-producing health-related knowledge. The First Nations-developed principles of OCAP™ (Ownership, Control, Access and Possession) describe the means by which First Nations control data collection processes in their communities and guide the ways that First Nations use and share those data, while ultimately retaining data ownership. However, there remain questions about how best to utilize existing health data sources to generate a comprehensive and representative picture of First Nations health outcomes.

In Ontario, Canada’s most populous province, the Chiefs of Ontario and the Institute for Clinical Evaluative Sciences (ICES) entered into a Data Governance Agreement in 2012. The Data Governance Agreement seeks to facilitate First Nations research, utilizing a rich array of health-related administrative data sources, in a manner that protects the interests of First Nations communities via the OCAP™ principles. The Data Governance Agreement has enabled the linkage of the federal Indian Registry System data file to ICES health administrative databases, allowing for the matching of First Nations individuals to their health and demographic records. This data linkage initiative has resulted in the creation of the largest First Nations research study cohort in Canada (n = 176,266, linkage rate of 93%) and will be used for disease surveillance and evaluation of health care, under the tenets of the Data Governance Agreement.

This presentation will describe the development of the ICES-Chiefs of Ontario partnership within the historical context of First Nations health and health research in Canada. The data governance and data sharing arrangements that protect the interests of First Nations communities will be reviewed, along with plans to use the linked data to support efforts in improving the health of First Nations peoples in Ontario. Finally, there will be a discussion around next steps in First Nations control of health data and the move towards establishing an independent First Nations Data Centre. This presentation will provide an illustrative example to researchers seeking to work in a mutually beneficial and collaborative manner with Indigenous populations to enable timely and relevant health research.

MANAGING LONG TERM STUDY DATA FOR LONGITUDINAL AND CHRONIC DISEASE RESEARCH

Authors: Donald Scobbie
University of Dundee

Presenting Author: Donald Scobbie

Corresponding author email: d.scobbie@dundee.ac.uk

Data management and data curation of long-term study and research databases are time consuming and complex activities that demand the attention of experts with very specific skills. Some of the most costly and complex data management activities emerge from consideration of two common scenarios. The first considers a single cohort used in a longitudinal study accruing data in distinct phases.
where the new data must be reconciled and merged with the existing data sets. The second scenario
occurs when distinct cohorts from different studies of the same disease are merged to create greater
scale in the research data. Again the data must be merged and reconciled in order to create an
aggregate data set that is valid in its totality.

Existing data management approaches are focused on the initial generation and preparation of
project research data and on preservation techniques that promote reuse of the data at the end of
individual research projects. These approaches do not consider longer term studies and research
programmes and fail to account for the key data merge, transformation and enrichment processes
that are applied over life-time study lengths and that shape the data to support analysis and results.
Failure to capture the project level transformation processes represents a major loss for long lived
research data sets, as data improvements identified by individual studies and cohorts are not fed back
into larger aggregated data sets to extend the data and improve the data quality.

Continuing dissatisfaction within the academic community with the lack of transparency in research
data management and the inability to reproduce study results and understand the provenance of
study data calls for further revision and extension of the research data management techniques. This
presentation examines the major data management issues associated with long term study data
management and presents a distinct life cycle for research data merge management. It argues that a
focus on transformation processes used within research projects brings transparency and reproducibility
benefits through process mining. It also argues that accommodating variation in the data and allowing
multiple simultaneous versions and potentially conflicting views to exist through the application of
competing transformation processes is possible through disciplined application of processes and
techniques from the software development life cycle.

Abstract 1444

PATIENT STRATIFICATION THROUGH A DATA SAFE HAVEN INFRASTRUCTURE:
EXPERIENCES FROM A MAJOR UK NHS TRUST AND PATHWAY MODELLING.

Authors:
Zisis Kozlakidis¹, Anil Gunesh²
¹Farr Institute of Health Informatics Research London/UCL, ²Farr Institute of Health
Informatics Research London

Presenting Author: Zisis Kozlakidis

Corresponding author email: z.kozlakidis@ucl.ac.uk

The need for an increasingly accurate patient stratification generates an ever growing pressure to
provide rich datasets that can be generated either through data linkage of existing sources or through
a wider information collection regime. However this ever richer collection of information per each
individual patient effectively counterbalances with any potential anonymisation strength, as patients
can become re-identified through the use of very detailed yet anonymised information. To overcome
this issue UCL has created a Data Safe Haven infrastructure using a walled garden approach, where
information can be stored, handled and analysed securely. This is achieved through a number of
features including a dual factor authentication and a file transfer mechanism that allows information to
be transferred simply and securely. Here we present our experiences at the Farr Institute in London on
the utilisation of this Data Safe Haven infrastructure for patient stratification purposes on a routine basis,
using actual patient data from one of the major UK NHS hospital trusts. We also highlight the challenges
relating to the flexibility and scalability of this approach as they relate to effective routine patient
stratification and provide an operational pathway model for the implementation of similar
infrastructure within data-sensitive healthcare environments.
SESSION G1 – DATA SCIENCE.
CHAIR: DR ELIZABETH FORD, BRIGHTON AND SUSSEX MEDICAL SCHOOL

Abstract 1662
INFORMATION ON RHEUMATOID ARTHRITIS DIAGNOSIS IN THE FREE TEXT OF GENERAL PRACTICE PATIENT RECORDS – WHAT IMPLICATIONS DOES THIS HAVE FOR DATA QUALITY?

Authors: Elizabeth Ford¹, John Carroll², Helen Smith¹, Kevin Davies¹, Rob Koeling², Jackie Cassell¹
¹Brighton and Sussex Medical School, ²University of Sussex

Presenting Author: Elizabeth Ford

Corresponding author email: e.m.ford@bsms.ac.uk

Introduction: Studies now report more sophisticated ways of finding cases in e-health datasets, but mainly use coded data only. Issues surrounding rates of false negatives and misleading diagnosis dates are relatively neglected, and it may be possible to address these using text from clinical notes. We used general practice (GP) patient records to explore diagnostic information about Rheumatoid Arthritis (RA) with the aim of describing how much information was found in free text and where. We sought evidence of delayed diagnostic coding of RA, by using occurrences in text of disease modifying anti-rheumatic drugs (DMARDs) as a proxy for disease codes.

Methods: Patients with incident diagnosis of RA between 2005-2008 were selected at random from the Clinical Practice Research Datalink. Patient records containing Read codes and free text were accessed from a period of 1-2 years before diagnostic RA code, until 14 days after. All codes indicative of RA-related activity prior to diagnosis were categorised in dummy variables. All text strings in the records were triple-annotated and adjudicated by domain experts for RA-related information. Pseudo-codes were assigned to text information and data were aggregated by patient.

Results: The sample comprised 209 women and 85 men, with 34,738 events recorded in the study period, of which 4,340 (12.5%) had associated text containing information related to RA. Non-specific marker tests, pain and DMARDs were the most common categories of disease information in text. For each category, about 10-15% of patients had information only in text, not in code; this increased to 27% for both joint signs and symptoms and other arthritis diagnoses. The top 20 codes associated with RA-related text included correspondence (26% of all text), general consultation (12%), and arthritis codes (11%). Delayed recording of the diagnosis was evident in 64 patients (22%) who had DMARDs mentioned in text >14 days prior to RA code. Evidence for RA activity and diagnosis prior to code in this group was high, with more and earlier referrals to rheumatology, Rh Factor tests, and DMARD prescriptions.

Conclusions: A proportion of disease and diagnostic information in RA is only recorded in the text of GP patient records, and evidence of a delay in diagnostic recording was present in 22% of patients. This has ramifications for the quality of e-health research; patients may lack a code for their condition and be missed from studies. Appropriate clinical management of patients will also be hampered by inadequate recording.
Abstract 1647 DEVELOPING MENTAL HEALTH RESEARCH USING ROUTINE HEALTH DATA AND NATURAL LANGUAGE PROCESSING

Authors: Angus Roberts¹, Robert Stewart²
¹University of Sheffield, ²King’s College London, NIHR Biomedical Research Centre

Presenting Author: Angus Roberts

Corresponding author email: angus.roberts@sheffield.ac.uk

Background: The South London and Maudsley NHS Foundation Trust (SLAM) is a large provider of secondary mental healthcare serving a geographic catchment of 1.2m people in South London, and has been using fully-electronic patient records (EPRs) across all services since 2006. The NIHR-funded Clinical Record Interactive Search (CRIS) application was developed and implemented at SLAM in 2007-8 to allow research-accessible datasets to be defined and derived from SLAM’s EPR. All clinical data, including free text, are available through CRIS for analysis in the form of de-identified datasets. CRIS currently accesses data on over 250,000 mental health service users, and has recently been installed at four other NHS Mental Health Trusts.

Methods: In providing CRIS data to researchers, we have had to address the fact that the EPR contains both structured and unstructured data (free text). Several studies using CRIS data have found that the number of data points available to researchers can be increased up to ten-fold by including information from the textual portion, greatly increasing data quantity. This includes information such as prescribing and test results that are often assumed to be well represented in the structured record. For other research, such as that involving confounding factors or symptomatology, the necessary information is not routinely included in the structured record, and so studies are impossible without recourse to the textual portion of the EPR.

Results: We have successfully applied Natural Language Processing (NLP), the computerised processing of human language, to tackle these difficulties. The CRIS NLP suite, implemented using the widely used GATE toolkit, routinely runs on a 90 node cluster to extract over 60 separate items of information from over 15 million full text documents. Work is ongoing to increase the number of information items available to researchers, to include more temporal information, and to provide linkage from the text to Linked Open Data resources. Successfully developed applications to date include those ascertaining clinical test and investigation data, pharmacotherapy and psychotherapy interventions, symptom profiles (e.g. 45 individual psychotic symptoms), illicit drug use and diagnostic statements.

Conclusion: We have been able to demonstrate the applicability of routinely collected health data to epidemiological and clinical research questions in mental healthcare, and of NLP to augment transformatively the structured information available from the EPR, having taken into account the tools available and their cost.
Abstract 1670

USING TEXT MINING TO EXTRACT DATA FROM CLINICAL NOTES AND MONITOR LONG-TERM CANCER SURVIVAL

Authors: Tom Liptrot¹, George Karystianis², Goran Nenadic², John Keane², Jac Livsey¹, Matthew Barker-Hewitt¹, Catherine O'Hara¹
¹The Christie NHS Foundation Trust, ²The University of Manchester

Presenting Author: Tom Liptrot

Corresponding author email: tom.liptrot@christie.nhs.uk

Background: As treatments improve it is becoming increasingly relevant to be able to offer cancer patients predictions of their long-term survival expectations beyond 3 or 5 years. In addition retrospective survival trend analyses allow the impact of changes in clinical practice over time to be evaluated. However, obtaining reliable data is challenging; several pieces of information are needed for many patients over many years. Some of these (e.g. death date and age) are available via the NHS Summary Care Record. However, in order to accurately model changes in survival, we need to consider changes in patient mix using data on major prognostic factors including, for example, disease site, histology, cancer stage and performance status.

The Christie Hospital patient data are now captured in a structured form. However, historic data exist only as free-text clinical notes, representing 10 years’ worth of unstructured electronic patient records for more than 60,000 patients. To manually extract the necessary data from notes would take around 40 person years and is, thus, unfeasible.

We have developed an approach that uses text mining to automatically extract patient data from clinical notes. By linking these data with the NHS Summary Care Record data, we can robustly measure long-term survival and trends, controlling for changes in patient mix over time.

Method: Two text mining approaches were developed to extract the data:

1. a rule-based approach to identify common lexical and syntactical patterns in text;
2. a machine learning approach using structured, consultant-entered data to automatically learn to classify free-text e.g. for diagnosis or staging.

These approaches have been applied to large numbers of patient records and the results have been combined and evaluated against ‘gold standard’ consultant-entered data.

Survival has been measured using Cox proportional hazards models.

Results: The rule-based approach focused on extracting disease site, histology and stage – achieving an overall accuracy of 98%. The machine learning approach achieved an F-score of 95% on known classes, covering 80% of patients. Applying these techniques to our historic patient records has produced a large dataset spanning around 10 years which has been used to measure long-term survival and trends over time.

Conclusions: We have demonstrated text mining methods can be used to reliably extract data from clinical notes to yield valuable information on long-term cancer outcomes.
ACCURACY OF DETERMINISTIC AND PROBABILISTIC RECORD LINKAGE: THE DE-DUPLICATION OF THE TB NOTIFICATIONS DATABASE

Authors: G P Oliveira¹, A L Bierrenbach², K R C Camargo Jr³, R S Pinheiro¹, C M Coeli¹
¹Federal University of Rio de Janeiro, ²Federal University of Goiás, ³State University of Rio de Janeiro

Presenting Author: R S Pinheiro

Corresponding author email: giselepoliveira@gmail.com

Introduction: Probabilistic and deterministic record linkage are techniques that allow the integration of distinct databases and the de-duplication of a database when there is no unique identifier for the patients. A major problem compromising the quality of the tuberculosis (TB) notifications database is the presence of duplicate records.

Objective: This paper aims to analyze the sensitivity and specificity of deterministic and probabilistic approaches to identify TB duplicated records, as well as the characteristics of the classification disagreement between the two techniques.

Methods: We analyzed all TB notifications (n=43,825) of the Rio de Janeiro State in the period 2009 to 2011. We developed a deterministic algorithm based on more than 70 rules using Stata 12.0®. Those rules were based on comparisons of three or more variables: patient’s name, mother’s name, date of birth and address. The deterministic approach used the frequency of the names in some rules. The probabilistic record linkage process was performed with OpenRecLink. We used the first name and sex of the patient to block the database, and a score of 18.3 to be the cut point for the de-duplication process. This cut point showed a 90.2% sensitivity and 99.9% specificity in a process done before. We used patient’s name, mother’s name, and date of birth for record comparison in the probabilistic approach. We built a gold standard by manual reviewing.

Results: The sensitivity and specificity of the deterministic and probabilistic approaches were 95.3% vs. 87.2% and 99.9% vs. 99.8% respectively. The deterministic approach identified that 21.3% of the records were duplicated, whereas the probabilistic approach identified that 19.5% were duplicated. Both approaches had a high agreement of the classifications (96%). However, deterministic linkage recovered 1285 additional pairs, compared to the probabilistic approach, and the probabilistic linkage identified 527 links that the deterministic linkage had not identified. The presence of missing values for the date of birth and the low similarity of the patient’s names were responsible for the failure to identify the records of the same individual by the probabilistic approach. For deterministic linkage, missing values for the mother’s name and low similarity in the date of birth were the main reasons for non-matching.

Conclusion: This study helps the improvement of tools currently used to identify duplicate records in secondary databases and the choice of the approach according the user’s needs for tuberculosis control.
SESSION G2 – JOINED-UP INFORMATION GOVERNANCE
CHAIR: DR WILLIAM DIXON, UNIVERSITY OF MANCHESTER

Abstract 1382

STUDYING THE ASSOCIATION BETWEEN GLUCOCORTICOID USE AND ADRENAL INSUFFICIENCY BY SUPPLEMENTING ELECTRONIC MEDICAL RECORDS WITH DIARIES AND BIOLOGICAL SAMPLES

Authors:
Rebecca M Joseph1, Kirin Sultana2, David W Ray3, Munir Pirmohamed4, Benjamin Brown5, William G Dixon6
1NIHR Manchester Musculoskeletal BRU, The University of Manchester, Manchester Academic Health Science Centre, UK, 2Clinical Practice Research Datalink, Medicines & Healthcare products Regulatory Agency, London, UK, 3Institute of Human Development, The University of Manchester, Manchester Academic Health Science Centre, UK, 4Wolfson Centre for Personalised Medicine, University of Liverpool, UK, 5Health e-Research Centre, Farr Institute for Health Informatics Research, The University of Manchester, Manchester Academic Health Science Centre, UK, 6Arthritis Research UK Centre for Epidemiology, The University of Manchester, Manchester Academic Health Science Centre, UK

Presenting Author: Rebecca M Joseph
Corresponding author email: rebecca.joseph@manchester.ac.uk

Adrenal insufficiency (AI) is a potentially serious condition in which cortisol production is impaired. It can be caused by use of glucocorticoids (GCs) yet little is known about how often this occurs (estimates 0-100%), who is at risk, how the dose and duration of GC use affects the risk, or the clinical impact. The aim of this research is to address these issues using a novel study design nested within the Clinical Practice Research Datalink (CPRD).

Primary care electronic medical record databases, such as the CPRD, have been used extensively in observational studies as they contain many person-years of clinical data, all prescriptions from primary care and a vast amount of disease and demographic information. However, misclassification of GC exposure in primary care data is probable due to unmeasured self-medicating and non-adherence. Furthermore, AI may be underdiagnosed in primary care. In this study, CPRD data will be supplemented with patient-reported medication use (to investigate misclassification) and saliva samples, collected to determine morning cortisol levels (low cortisol is a marker for AI) and for genetic tests.

Because of the importance of preserving the anonymity of patients and practices within the CPRD, a dataflow has been developed to allow the research team to coordinate the study whilst ensuring, outside the general practice, nobody can link identifiable patient information to pseudonymised medical records. Feasibility is conducted by CPRD and passed to GP practices who send invitations to eligible patients. Willing participants then provide their consent and contact details to the research team who post diaries and sample kits. Following data collection, the research team will deidentify the study data. A dedicated team will deidentify the CPRD data before the research team are provided with CPRD datasets.

In study set-up, the project team have faced challenges of multiple approvals and optimising protection of CPRD practice and patient anonymity. Approvals to date include the National Research Ethics Committee (by proportionate review), CPRD’s Independent Scientific Advisory Committee and NIHR CRN Portfolio adoption. A previous study (1) that supplemented CPRD data with biological samples highlighted the length of time needed to obtain research management approval for all participating general practices as a major challenge. In our study, practices will be acting as Participant Identification Centres (PICs), simplifying the approval process. Research activities will instead be coordinated by the research team. Recruitment is commencing in 2015.


The Farr Institute International Conference 2015: Data Intensive Health Research and Care
26-28 August 2015, St Andrews
THE ‘SCHOOL TRAVEL AND CHILD SAFETY SURVEY’ PROJECT: LINKING RESULTS FROM A SCHOOL BASED SURVEY TO OTHER ROUTINELY COLLECTED DATA

Authors: Samantha Turner¹, Ronan Lyons², Sarah Rodgers², Richard Fry²
¹Swansea University, ²The Farr Institute of Health Informatics Research, Swansea University

Presenting Author: Samantha Turner
Corresponding author email: s.turner@swansea.ac.uk

Background: The School Travel and Child Safety Survey (STCSS - www.childsafetysurvey.net) is a short online survey designed to be undertaken by children aged 10 – 13 years in school. Survey questions focus on children’s travel and safety behaviours as well their exposure to safety interventions and potential hazards in the local environment.

Methods: The STCSS was originally developed as part of the European project TACTICS (http://www.childsafetyeurope.org/tactics/), with the aim to improve the quality and comparability of local child safety data. As part of this work, the STCSS was piloted in several schools across Europe, including several schools in Wales.

Following this project, Swansea University worked Neath Port Talbot (NPT) Council to develop the survey as a Council-led tool. NPT readily adopted the STCSS as it offered a cost effective solution to their paper-based surveys, and improved the quality and breadth of data available to the Council. Pupil login screens were also added to the survey in NPT, to enable the results to be incorporated anonymously into the Secure Anonymised Information Linkage (SAIL) databank at Swansea University; an approach made possible by SAIL’s split file privacy protection methodology.

Incorporating the survey results into SAIL, will enable children’s survey responses to be anonymously linked to other routinely collected data (e.g. socioeconomic status, household demographics, health and educational data). This development will allow associations between children’s behaviours/exposures and their health/educational outcomes to be explored at a scale never previously possible.

Results: The TACTICS project demonstrated the STCSS to be a feasible and reliable survey, with the potential to support local governments with the planning and evaluation of local child safety interventions (Full Report - http://www.childsafetyeurope.org/tactics/info/child-safety-survey.pdf).

Currently, the STCSS is being developed in NPT Council, and is scheduled to be rolled out to all schools in June 2015. We plan to present the survey results, along with the success of the SAIL data linkage process, at the conference this August.

Conclusions: The ability to link the results from the STCSS to other routinely collected data in the SAIL database will provide much needed evidence on the medium and long-term effectiveness of child safety interventions and policies; as well as provide local governments with much needed data to support the planning and evaluation of local child safety interventions.

If the STCSS is successfully implemented in NPT, opportunities for its implementation in other regions across Wales/Europe will be explored.
Abstract 1423

USING ELECTRONIC HEALTH RECORDS TO PREDICT COSTS AND OUTCOMES IN CHRONIC DISEASE USING THE EXAMPLE OF STABLE CORONARY ARTERY DISEASE

Authors:
Miqdad Asaria¹, Simon Walker¹, Stephen Palmer¹, Christopher Gale², Mark Sculpher¹, Harry Hemingway³
¹Centre for Health Economics, University of York, ²University of Leeds, ³UCL

Presenting Author: Miqdad Asaria
Corresponding author email: miqdad.asaria@york.ac.uk

Objectives – To use electronic health records (EHR) to predict the lifetime costs and health outcomes of patients with stable coronary artery disease (SCAD) stratified by their risk of experiencing future cardiovascular events and to evaluate the cost-effectiveness of treatments targeted at these populations.

Design – Cohort study

Setting – One country (England) with one health system (the National Health Service).

Participants – 94,966 patients with SCAD in England between January 2001 and March 2010, identified in the CALIBER data set that comprises four prospectively collected, linked electronic health record sources: Clinical Practice Research Datalink (primary care data), Hospital Episode Statistics (hospital admissions), the disease registry MINAP (Myocardial Ischaemia National Audit Project), and the Office for National Statistics mortality register (cause specific mortality data).

Main outcome measures – lifetime costs and quality adjusted life expectancy for patients with stable coronary artery disease stratified by baseline risk.

Results – in the lowest risk tenth of SCAD patients (mean five year CVD risk of 3.4%), remaining mean lifetime healthcare costs per patient were £116,888 (95% CI £64,743 to 168,032) and remaining mean lifetime quality adjusted life years (QALYs) per patient were 19.1 (95% CI 18.1 to 19.9) years. In the highest risk tenth of the population (mean five year CVD risk of 43.7%), remaining mean lifetime healthcare costs per patient were £43,020 (95% CI £37,731 to £48,842) and remaining mean lifetime QALYs per patient were 3.3 (95% CI 3.0 to 3.7) years. A new treatment with a hazard reduction of 20% for acute myocardial infarction, stroke or CVD death and no side effects would be cost effective if priced below £72 per year in the lowest risk group and £646 per year in the highest risk group.

Conclusions – We provide a framework for using EHR to estimate the lifetime healthcare costs and health outcomes of patients with chronic disease. This model lends itself to informing decisions on commissioning, price and reimbursement and the ongoing evaluation of new interventions. In the case of SCAD we find that at current prices, established cardiovascular treatments and future therapies will require stratification by patient risk in order to be cost-effective.

Abstract 1559

DATA LINKAGE IN SOCIAL CARE: A PILOT PROJECT

Authors:
Alison Orrell¹, M Heaven², C Robinson¹, D Seddon¹
¹Bangor University, ²Swansea University

Presenting Author: Alison Orrell
Corresponding author email: a.orrell@bangor.ac.uk

Existing data linkage projects in Wales that utilise routinely collected administrative data have predominantly focused solely on health datasets. To build the complete picture of service provision there is a need to broaden this to include social service provision by Local Authorities and provision of support by third sector organisations. The research study presented here is a pilot project to test the feasibility of linking datasets from a local authority, the NHS and third sector organisations. The focus of this work is on individual level data from adults who are referred to social services in order to avoid admission to hospital or to facilitate their discharge from hospital. This presentation will include a
description and reflective commentary on the processes of drawing a range of organisations together, technical and organisational barriers to linking datasets, critical analysis of data quality and robustness and preliminary outcomes of linked data analyses.

SESSION G3 – LINKED OBSERVATION FOR HEALTH POLICY.
CHAIR: PROF JANE FORD, UNIVERSITY OF SYDNEY

Abstract 1359

EMERGENCY ADMISSIONS ACROSS THE TRANSITION FROM PAEDIATRIC TO ADULT CARE: RETROSPECTIVE ANALYSIS OF ENGLISH HOSPITAL DATA

Authors: Linda Wijlaars1, P Hardelid1, J Woodman1, R Cheung2, J Allister3, R Gilbert1
1UCL Institute of Child Health, 2Evelina Children's Hospital, 3Royal College of General Physicians

Presenting Author: Pia Hardelid

Corresponding author email: linda.wijlaars@ucl.ac.uk

Background: Transition from paediatric to adult health services is often sub-optimally achieved for children with long-term conditions (LTCs). We investigated trends of emergency admissions in England across the transition from paediatric to adult care.

Methods: We analysed 1,186,392 emergency admissions between 2009-2011 in children and young people (CYP) aged 10–24 years using Hospital Episodes Statistics data. We calculated emergency admission rates for CYP pre- (10-15 years), during (16-18 years), and post-transition (19-24 years), using mid-year population estimates as denominators to calculate rates. We stratified analyses by underlying LTCs, which we defined by validated International Classification of Disease version 10 (ICD-10) code clusters. We excluded injury- and maternity-related admissions. We used a multivariable random-coefficient Poisson regression to determine factors contributing to the rise in emergency admissions across transition, adjusting for sex, age, underlying LTCs, and deprivation quintile.

Results: Prevalence rates for emergency admissions increased 64% from 29/1000 pre-transition to 48/1000 post-transition for girls, and from 26/1000 pre- to 31/1000 post-transition for boys (17% increase). This increase overwhelmingly occurred during transition (16-18 years), where rates increased 4.0/1000 per year for girls and 2.3/1000 for boys.

At the same time, average length of hospital stay increased from 2.1 to 2.9 days for girls, and 1.8 to 5.0 days for boys (increases of 38% and 186% for girls and boys, respectively). Admissions were also more likely to be via the emergency department (increase from 55% to 69%).

Disparities in admission rates increased during transition, with CYP from the most deprived quintile more likely to have an emergency admission compared to the least deprived quintile, and more so than before transition.

The multivariable analysis showed the increase in admission rates during transition is driven almost entirely by CYP with underlying LTCs: overall, rates increased 14% (IRR: 1.14, 95%CI: 1.13-1.14), but for CYP with LTCs, rates increased 46% (IRR: 1.46, 95%CI: 1.46-1.47). Rates increased most rapidly for mental health problems, metabolic/endocrine disorders, CYP with multiple long-term conditions (both genders), and respiratory disorders (girls only). Deprivation had a modest effect on emergency admission rates when adjusting for LTCs, with the most affluent quintile 9% less likely to be admitted compared to the most deprived quintile (IRR: 0.91, 95%CI: 0.91–0.92).

Conclusion: Our results support the widely recognised difficulties young people face when they transition to adult health services. Efforts to improve services could start by focussing on mental health problems, metabolic/endocrine disorders, respiratory disorders, and multiple LTCs.
Abstract 1346  

TEN-YEAR RISKS OF HARM IN ADOLESCENTS HOSPITALISED WITH VIOLENT, DRUG/ALCOHOL-RELATED, OR SELF-INFLECTED INJURY: ANALYSIS OF LINKED HOSPITAL AND MORTALITY DATA

Authors:  

Annie Herbert¹, Leah Li¹, Arturo Gonález-Izquierdo², Alexandra Pitman³, Ruth Gilbert⁴

¹Institute of Child Health, University College London, ²Farr Institute of Health Informatics Research, University College London, ³Department of Psychiatry, University College London, ⁴Institute of Child Health and Farr Institute of Health Informatics Research, University College London

Presenting Author: Ruth Gilbert

Corresponding author email: annie.herbert.12@ucl.ac.uk

Background & Objective: The hospitalisation of an adolescent for violent, drug/alcohol-related, self-inflicted injury provides opportunity to identify vulnerable individuals and intervene to reduce harm. However, there is a lack of evidence on the long-term risk of harm in this population. We used administrative hospital data linked to death records to determine long-term risks of death and emergency re-admission in these adolescents, and compared these risks to those following accident-related injury.

Methods: We identified adolescents (10-19 year olds) in longitudinally-linked admissions data linked to death registrations for England (Hospital Episode Statistics, April 1997-March 2012), whose index emergency admission for injury was related to adversity (violence, drug/alcohol use, or self-harm) or an accident, defined using ICD-10 codes. We calculated Kaplan-Meier estimates of cumulative risks of death and emergency re-admission after discharge from the index admission and within ten years, by sex and age-groups (10-14y, 15-17y, 18-19y). We estimated hazard ratios (HRs) of death and emergency re-admission following adversity-related injury (relative to following accident-related injury), adjusted for age-groups. We also estimated age-adjusted HRs of a second, third, fourth, and fifth emergency re-admission (<5% had more than five).

Results: Among 333,009 adolescents admitted with adversity-related injury, the risk of death by ten years was 7.3/1,000 (1 in 137) for girls and 15.6/1,000 (1 in 64) for boys; risks of emergency re-admission were 54.2% for girls and 40.5% for boys. All ten-year risks were higher in 18-19 year olds than in younger age-groups. Risks of death were increased after adversity-related injury when compared with those after accident-related injury (girls age-adjusted HR: 1.61; 95% confidence interval [CI]: 1.43 to 1.82; boys 2.13; 1.98 to 2.29), as were risks of emergency re-admission (girls age-adjusted HR: 1.76; 95% CI: 1.74 to 1.79; boys 1.41; 1.39 to 1.43). These risks were increased after all combinations of adversity-related injury, but particularly after any drug/alcohol-related or self-inflicted injury (i.e., with or without violent injury). Risks of repeated readmissions (i.e. of a second-fifth emergency re-admission) were also increased following adversity-related injury (girls age-adjusted HR of a fifth emergency re-admission: 2.38; 95% CI: 2.29 to 2.46; boys 2.94; 2.81 to 3.07).

Conclusions: Adolescents admitted to hospital for any violent, drug/alcohol-related, or self-inflicted injury are at increased risks of subsequent harm. Clinical guidelines and interventions for adolescents presenting with such injuries need to be developed to reduce these risks.
Abstract 1393  

**DRUGS-RELATED DEATHS SOON AFTER HOSPITAL-DISCHARGE OF EVER-INJECTORS**

Authors:  
Sheila Bird¹, S M Bird², E L C Merrall³, S J Hutchinson⁴  
¹MRC Biostatistics Unit, Cambridge CB2 0SR, ²MRC Biostatistics Unit, CAMBRIDGE CB2 0SR, ³Novartis Vaccines & Diagnostics, ⁴Health Protection Scotland

Presenting Author:  
S M Bird

Corresponding author email: sheila.bird@mrc-bsu.cam.ac.uk

We validate that the 28 days after hospital-discharge are high-risk for drugs-related death (DRD) among drug users in Scotland and investigate key risk-factors for DRDs soon after hospital-discharge.

Data are from an anonymous linkage of hospitalisation and death records to the Scottish Drugs Misuse Database (SDMD), which include over 98,000 individuals registered for drug treatment during 1 April 1996 to 31 March 2010 with 705,538 person-years of follow-up, 173,107 hospital-stays, and 2,523 DRDs.

Time-at-risk of DRD was categorised as in Merrall et al. (2013) as: during hospitalization, within 28 days, 29–90 days, 91 days–1 year, >1 year since most recent hospital discharge versus ‘never admitted’.

Factors of interest were: having ever injected, misuse of alcohol, length of hospital-stay (0–1 versus 2+ days), and main discharge-diagnosis.


DRD-rate in the 28 days after hospital-discharge did not vary by length of hospital-stay but was significantly higher for clients who had ever-injected versus otherwise. Three leading discharge-diagnoses accounted for only 150/290 DRDs in the 28 days after hospital-discharge, but ever-injectors for 222/290. Hospital-discharge remains a period of increased DRD-vulnerability in 2006–2010, as in 1996-2006, especially for those with a history of injecting. For those with a history of injecting, we estimated that the DRD-rate in the 28 days after hospital-discharge was 2.4 DRDs per 1000 hospital-discharges, about half their rate in the 28 days after prison-release.

---

Abstract 1544  

**DUAL DIAGNOSIS OF MENTAL HEALTH PROBLEMS AND SUBSTANCE ABUSE IN PRIMARY CARE IN SOUTH WALES**

Authors:  
Sarah Rees¹, Ting Wang¹, Daniel S Thayer¹, Mathilda Castagnet¹, Ann John²  
¹SAIL Databank, College of Medicine, Swansea University, ²College of Medicine, Swansea University

Presenting Author:  
Daniel S Thayer

Corresponding author email: s.e.r.rees@swansea.ac.uk

Background: Literature suggests that mental health and substance misuse requires a dual model of care which is not currently offered in South Wales, due to the lack of evidence on service demand in the region.

Objectives:

- To provide the first insight into the magnitude of adults with dual diagnosis (DD) of substance misuse and mental health problems in primary care for Abertawe Bro Morgannwg University Health Board (ABMU) in Swansea, Wales.
- To provide early evidence of demand in support of potential service restructuring.

Methods: Prevalence: point prevalence was calculated for adults with both substance misuse and mental health in their GP record in the five years prior to snapshot date. Adults registered to the practice at the snapshot date were used as the practice population.
Incidence: first lifetime incidence was calculated for adults who had progressed from single to dual diagnosis for the first time or had a first event with a specified DD Read code. A six-month clearance period was used to ensure as far as possible that only first-time diagnoses of DD were included as incident cases.

Results: Mean results of prevalence/incidence rates (57.9/15.9, respectively) for DD were presented per 10,000 person-years in 2012.

Discussion: The analysis of prevalence and incidence is affected by coding behaviour, as only patients with relevant codes in their GP record were detected. The analysis provides a crude, non-standardised estimate of the prevalence and incidence of DD without taking into account differences between practices based characteristics such as age, gender and deprivation. Due to the assumed chronic nature of the conditions included in the analysis, rates may be under-estimated, as they exclude patients who have recovered and relapsed during the study period.

The five year look back in prevalence analysis may exclude patients who drop out of the system regularly. Conversely, the assumption made regarding the chronic nature of dual diagnosis may lead to over-estimated prevalence (e.g. patients prevalent at some point within the five year period who have recovered).

Conclusions: The benefit of this work is the ability to offer care providers evidence to inform practice in an area where evidence is lacking. The richness of the primary care data takes into account medication and administrative records, in addition to diagnosis records, to provide a better estimate than diagnosis alone, but there are challenges of utilizing this type of dataset.

Abstract 1380

LINKING OBSTETRIC PATIENT RECORDS TO BLOOD PRODUCT DATA: CHALLENGES AND OUTCOMES

Authors: Jane B Ford, Jillian A Patterson, Christine L Roberts

Kolling Institute, University of Sydney, Australia

Presenting Author: Jane B Ford

Corresponding author email: jane.ford@sydney.edu.au

In Australia there is no routine system for tracking blood products from donation through to patient receipt and outcomes. With rising rates of obstetric haemorrhage and obstetric transfusions, there is interest in the volume of red cells being transfused and age of blood transfused in maternity patients.

Women receiving 1 or more packs of red blood cells during a hospital birth admission in New South Wales (NSW), Australia, between July 2006 and December 2010 were identified from routinely collected hospital and birth data and probabilistically linked with patient blood issue information from Blood Watch (public hospital blood bank records for each blood pack issued) to ascertain number of transfusions given. Using blood pack numbers, these data were deterministically linked to blood donor data from the Red Cross Blood Service to ascertain date of blood donation.

Births at the 67 hospitals reporting blood use to Blood Watch (n=231,845) represented 74% of all NSW public hospital births over the study period. There were 4207 obstetric transfusions (1.9% of Blood Watch hospital births). We found that, among a sub-population of women receiving between 1 and 4 packs of blood, two-thirds (66%) of women received 1-2 packs of blood. There were no differences in age of blood transfused a) between women with and without severe morbidity (21 days), and b) in women readmitted or not (22 days).

The following 3 challenges were identified using these data. Using administrative data it remains difficult to ascertain the relative timing of events and to completely untangle whether morbidity is a consequence or a reason for transfusion. Identifying the hospitals and reporting periods for inclusion in analyses is difficult when hospitals submitting blood pack issue data did so on a monthly basis and lack of returned data in a given month may represent no transfusions or non-reporting. Blood issued does not necessarily equate to blood given to women. Strategies used to address each of these challenges will be presented.
Abstract 1440

EFFECTS OF AN AIR POLLUTION PERSONAL ALERT SYSTEM ON HEALTH SERVICE USAGE IN A HIGH-RISK GENERAL POPULATION: A QUASI-EXPERIMENTAL STUDY USING LINKED DATA

Authors:

R Lyons¹, R Bailey¹, H Brunt², S Rodgers¹, H Snooks¹, S Thomas²

¹Farr Institute CIPHER, ²Public Health Wales NHS Trust

Presenting Author: R Lyons

Corresponding author email: r.a.lyons@swansea.ac.uk

Introduction: Outdoor air pollution is detrimental to health, particularly for those with respiratory or cardiac problems. The Committee on the Medical Aspects of Air Pollution (COMEAP) advises warning those most at risk, when air pollution levels rise, to take additional measures to reduce exposure and manage their conditions. A number of forecasting systems are in operation but have yet to be fully evaluated. An expected benefit of such systems is that those most at risk will seek to minimise their exposure and avoid acute episodes of ill health. In south Wales, a multiagency partnership implemented a novel personalised real-time air pollution alert system, airAware, in 2012. The performance of the system and its impact on participant health service utilisation was evaluated using anonymised linked routine data.

Method: 1395 patients at four local GP practices with diagnoses of asthma, COPD or coronary heart disease were invited by their GP to sign up to the system. We assessed the accuracy of airAware system alerts against measurements taken at six local air quality monitors. We used a quasi experimental design to compare health service utilisation in general practice, outpatients, emergency department and admissions between those signing up to the intervention or not, adjusting for differences in prior service utilisation and deprivation. The Secure Anonymised Information Linkage (SAIL) system was used to create the cohort. Participants were followed up for two years following the service launch. Incidence rate ratios were derived from negative binomial regression analyses.

Results: 180/1395 people accepted the invitation to participate in the airAware system. System sensitivity was 83.9% and specificity 96.6%. There were baseline differences, with those signing up using significantly more general practice services and having fewer emergency hospital admissions. Following adjustment for baseline differences the intervention was associated with increased all-cause emergency department attendances (IRR 1.89, 95% CI 1.34, 2.6) and emergency respiratory admissions (IRR 3.97, 95% CI 1.59, 9.93).

Conclusions: This study demonstrates the value of real-world evaluation of interventions using linked data. Whilst the alert system performed well, uptake was relatively poor. Results were in the opposite direction to expected, with significant increases in health service utilisation in patients using the system. Results from this single centre non-randomised study may not be generalisable, particularly given local variation in pollution exposures, but evidence from this study does not support continuation or widespread implementation of airAware.
LINKED POPULATION HEALTH DATA FOR EVALUATING THE EFFECT OF MATERNAL INFLUENZA VACCINATION AND BIRTH OUTCOMES IN WESTERN AUSTRALIA

Authors: Annette Regan¹, Hc Moore², G Shellam¹, N de Klerk², Db Mak³, Pv Effler³
¹School of Pathology and Laboratory Medicine, University of Western Australia, ²Telethon Kids Institute, ³Communicable Disease Control Directorate, Western Australia Department of Health

Abstract 1500

Presenting Author: Annette Regan

Background: Pregnant women are the highest priority for seasonal influenza vaccination, yet 66% of women in Australia remain unvaccinated. Previous research suggests maternal immunisation not only reduces respiratory infections in mothers and infants, but also prevents adverse birth outcomes in vaccinated mothers. Although the vaccine has been available to pregnant women since 2003, no population-based study has yet evaluated its impact on neonatal health in Australia.

Methods: The Maternal Influenza Vaccination in Western Australia (MIV-West) cohort was established by the Data Linkage Branch of the Western Australia Department of Health in 2013, using probabilistic linkage of state-held maternal vaccination records, birth registrations, midwives notifications, disease notifications, hospital records, and death registrations. Between 2012 and 2013, a total of 57,404 births were included in the cohort. We calculated the incidence of adverse birth outcomes, including stillbirth, very low birth weight (<1500 grams), and severe preterm birth (<32 weeks gestation). The incidence of stillbirth was compared by maternal vaccination status using Cox proportional hazards models, adjusting for propensity of vaccination, maternal smoking, and Aboriginal status. Other birth outcomes were compared using adjusted logistic regression models which included similar confounders.

Results: A total of 5,057 (8.8%) mothers were immunised against influenza during their pregnancy. Vaccination was more common among mothers who resided in highly accessible areas, were of higher socioeconomic status, had chronic medical conditions, or had risk factors for stillbirth, including plural birth, primiparity, pre-eclampsia or gestational diabetes. Propensity scores were calculated based on these factors. The incidence of stillbirth was 51% lower in vaccinated women compared to unvaccinated. There were 5.1 stillbirths per 100,000 pregnancy days in unvaccinated women, compared to 3.1 stillbirths per 100,000 pregnancy days in vaccinated women (HR: 0.49, 95% CI: 0.29-0.83). Infants born to vaccinated mothers were also 25% less likely to be severely preterm (aOR: 0.75, 95% CI: 0.56-1.00) and 31% less likely to be very low birth weight compared to unvaccinated mothers (aOR: 0.69, 95% CI: 0.50-0.96).

Conclusions: Our findings support international literature indicating there are health benefits afforded to infants by maternal influenza immunisation. Our results indicate influenza vaccination is not only safe during pregnancy, but also has perinatal health benefits. Additional analyses could help identify the mechanism underlying these health benefits. Considering >90% of women who choose to be immunised do so to protect their infant, these results could also be used to promote this important public health intervention to pregnant women in Australia and internationally.
THE EFFECTS OF AN AREA BASED SOCIAL HOUSING REGENERATION PROGRAMME ON THE PHYSICAL AND MENTAL HEALTH OF TENANTS.

Authors: Rowena Bailey¹, S Rodgers¹, R Johnson¹, F Dunstan², W Poortinga², R Lyons¹
¹Swansea University, ²Cardiff University

Presenting Author: Rowena Bailey

Corresponding author email: r.bailey@swansea.ac.uk

Introduction: Poor quality housing can have negative consequences for the health of occupants. Insufficient heating is associated with respiratory and cardiovascular conditions. Poor internal floor plans, particularly in bathrooms and kitchens, can affect the likelihood of falling or being burnt. The Welsh Government has introduced the Welsh Housing Quality Standards (WHQS) which Carmarthenshire County Council has been implementing through a regeneration program - the Carmarthenshire Housing Standard (CHS). Work began in 2007 and is due to be completed in December 2015. The research being undertaken seeks to evaluate the effect of implementing WHQS policy on the physical and mental health of the residents.

Methods: 9500 homes have received housing improvements, carried out through multiple elements of work and captured in data supplied by the council. A longitudinal design was used to evaluate this natural experiment. Health service utilisation in emergency admissions and general practice attendance of a dynamic cohort within the regeneration area were evaluated in relation to two comparator groups. Records relating to respiratory and cardiovascular conditions as well as falls and burns were obtained retrospectively for each individual, from January 2005 up until March 2015, with measurements captured in multiple time intervals. This was a repeated measures approach using multi-level modelling. Negative Binomial regression was used to evaluate the effect of change in WHQS compliance status on the health outcomes.

Results: There were over 20,000 individuals included in the intervention cohort, almost 50,000 in the social housing comparator group and nearly 200,000 in the same local authority. Over 75% of residents in the intervention cohort remained in the same home. Health utilisation measures were counted monthly for each individual, for 123 months resulting in over 4 million individual records. Trends in hospital admissions show increases over the 10 year period. We have compared socio-economic characteristics and the proportion of people with an admission between study groups. Once the intervention is complete we will compare the intervention and comparator cohorts and adjust the model for available confounders.

Discussion: This study demonstrates the benefit of linking routine data to evaluate a natural experiment. Our results will show if improved housing conditions effects health service utilisation of the occupants. We will identify any differences between residents in the study area and the comparator groups to evaluate the benefit of a targeted housing improvement program to implement the WHQS in relation to ‘business as usual’ practices of housing stock maintenance.
Abstract 1715  

**A POLICY MODEL OF ALCOHOL-RELATED HARM S FOR PREDICTING LIFE YEARS AND QUALITY-ADJUSTED LIFE YEARS**

**Authors:** Pattara Leelahavarong, Dr. James Lewsey, Prof. Andrew Briggs  
*Health Economics and Health Technology Assessment (HEHTA), University of Glasgow*

**Presenting Author:** Pattara Leelahavarong

**Corresponding author**
Email: p.leelahavarong.1@research.gla.ac.uk

**Objectives**
To develop an alcohol-related harms' policy model that predicts life years and quality-adjusted (QA) life years (LYs) of drinking patterns measured by the Alcohol Use Disorder Identification Test or AUDIT (range 0=non-drinker to 40=high-risk drinker).

**Methods**
The Scottish Health Surveys cohorts (SHeS) 1995-2012 linked to the Scottish Morbidity Record (SMR) and National Records of Scotland (NRS or GROS) from 1981 to 2013, which are the hospitalisation and death records, were analysed using survival analysis to estimate the cause-specific hazards of plausible alcohol-related harms classified by gender as follows: 1) hospitalisation and death due to wholly alcohol-attributable conditions; 2) hospitalisation and death due to partly alcohol-attributable conditions (i.e. chronic and acute conditions); 3) hospitalisation and death due to other causes. The relevant factors for adjusting models were age, drinking patterns (i.e. binge drinking (Yes/No) and AUDIT score), smoking status (i.e. never/ex-/current smokers), general health condition (GHQ score), having cardiovascular condition (Yes/No), prior non-alcohol related hospitalisations, marital status, and SIMD quintile. Multiple imputation approach was used to deal with SHeS missing data. Health utility scores of drinkers were also generated from SF-6D (SF-12) index measured in SHeS 2003. A state transition model was developed to estimate LYs and QALYs based on the alcohol-related harms survival analyses and utility scores.

**Results:** The total number of no prior alcohol-related hospitalised cohort was 46,230 (male=20,729 and female=25,501). The adjusted model undertaken multiple imputation showed the each of increasing AUDIT score, and hazard ratios (HRs) of wholly alcohol-attributable hospitalisation were 1.06 (95% CI: 1.01, 1.11) for male (p=0.021) and 1.10 (95% CI: 1.01, 1.21) for female (p=0.035). The HRs for partly alcohol-attributable condition was 1.04 (95% CI: 1.01, 1.08) for male (p=0.023), but female was not statistically significant (p=0.362). LYs and QALYs of each AUDIT score would be also estimated.

**Conclusions:** The policy model could be used in the evaluation of cost-effectiveness of alcohol consumption control programmes using estimated LYs and QALYs, which are recommended for economic evaluation.

---

Abstract 1718  

**ADMISSION TO HOSPITAL AND EDUCATIONAL ACHIEVEMENT IN CHILDREN AGED SEVEN YEARS: A RECORD-LINKED ELECTRONIC BIRTH COHORT ANALYSIS**

**Authors:** Annette Evans¹, Frank D Dunstan¹, Shantini Paranjothy¹, Melanie Healy², Joanne Demmler², Ronan Lyons², David Fone¹  
¹Cardiff University, ²Swansea University

**Presenting Author:** Annette Evans

**Corresponding author**
Email: EvansA50@cardiff.ac.uk

**Background:** Poor educational outcomes are associated with demographic, socio-economic and pregnancy-related factors but it is not known whether admission to hospital before the age of 7 years, as a measure of childhood morbidity, is also a risk factor for educational under-achievement.

**Aim:** To investigate the influence of admission to hospital in early childhood on educational achievement at seven years of age.
**Study design:** A total population electronic birth cohort with anonymised record-linkage of multiple health, education and administrative datasets – the Wales Electronic Cohort for Children (WECC).

**Methods:** Multilevel logistic regression analysis of 64,934 babies born in Wales between 1/1/1999 and 31/8/2001. The outcome measure was the expected performance (level 2+) at the Key Stage 1 assessment (KS1) in children aged 6 to 7 years with no Special Educational Need provision. Data from the core WECC dataset were linked to hospital admissions, the congenital anomaly register, the National Pupil Database and the Pupil Level Annual School Census (PLASC) educational dataset.

Odds ratios for not achieving KS1 were estimated for any emergency admission to hospital before KS1 assessment, and separately for any respiratory admission and external causes (e.g. injury, burns, poisoning). Children were modelled nested within schools within Local Education Authorities to estimate the effects of academic season of birth, free school meal eligibility, school moves, and school-level effects of the school percentage of free school meals and school size. The model was adjusted for sex, gestation at birth, small for gestational age (<10th centile), maternal age, parity, congenital anomaly, delivery by Caesarean section, Apgar score at five minutes, neonatal admission, maternal smoking and breast feeding during pregnancy.

**Results:** 48% of children had an all-cause emergency hospital admission prior to KS1, 19.7% with any respiratory admission, and 8.6% with any external cause. Poor educational achievement was associated with any emergency hospital admission, adjusted odds ratio (OR) 1.12 (95% CI: 1.05, 1.20). The OR for any respiratory emergency admission was 1.10 (95% CI: 1.02-1.20) and for any emergency external cause admission (e.g. injury, burns, poisoning) the OR was 1.19 (95% CI: 1.07-1.32). In addition to the known associations with pregnancy and birth circumstances, we also found important effects of late academic season of birth, free school meals eligibility, the number of school moves, and school-level catchment area free school meals.

**Conclusion:** Childhood morbidity leading to hospital admissions in the early years is associated with poor educational achievement at age 7 years.

---

**SESSION G5 – NATURALISTIC COHORT STUDIES.**

**CHAIR: DR SINEAD BROPHY, SWANSEA UNIVERSITY**

**Abstract 1345**

**OBESITY IN PREGNANCY: A RETROSPECTIVE STUDY ON HEALTH SERVICE UTILISATION AND COSTS ON THE NHS**

**Authors:** Sinead Brophy¹, Kelly Morgan², Muhammad Rahman³, Mark Atkinson³  
¹FARR Institute/Swansea University, ²FARR Institute CIPHER, ³FARR Institute Swansea

**Presenting Author:** Sinead Brophy

**Corresponding author email:** s.brophy@swansea.ac.uk

**Objective:** Maternal obesity is a rising public health problem, as is it is associated with heavier infants, need for more interventions for the birth, poorer health of the baby and greater likelihood that the infant will grow up to be overweight/obese. This study examines the direct healthcare cost of being obese at the pregnancy booking visit in order to inform the amount that could be spent on public health prevention and still be cost saving.

**Setting and participants:** Linked anonymised electronic datasets (GP records, hospital admissions) linked to questionnaire, data extracted from the maternal pregnancy record (notes) and the child health record (red book) on a cohort of women & infants participating the Growing up in Wales cohort study. Maternal booking weight was objectively recorded in clinic. Women were stratified in groups of normal body mass, overweight or obese.

**Primary outcome:** Total health service utilisation (GP visits, medications, inpatient and outpatient visits)
and direct health care costs. Costs were calculated separately for the mother and the infant and related to health service use throughout pregnancy (in the case of the mother) and in the 12 months from birth (in the case of the infant).

**Results:** There was a strong association between healthcare usage cost and BMI \( (p<0.001) \). Adjusting for maternal age, parity, ethnicity and co-morbidity mean total costs were higher both among obese women \( (RR: 1.39, 95\%CI: 1.38 \text{ to } 1.39) \) and their infants \( (1.74, 95\%CI:1.42 \text{ to } 2.1) \) compared to those mothers of normal weight. This translates into £1172 extra cost of the mother and £1138 extra cost of the infant \( (£2310) \) per pregnancy for obese women compared to normal weight women.

**Conclusion:** Obesity in pregnancy results in longer hospital stays, c-section deliveries and poor health for both mother and baby. The costs in the analysis are conservative as they do not take into account maternal costs post pregnancy or health care visits (which are not captured using routine data). One in five women are obese in pregnancy and there were 698,512 births in England and Wales in 2013. This suggests that up to £322,712,544 can be spent reducing obesity in young women, and this could still be costs saving to the NHS. This method of linking to health service use can be used to evaluate the economic impact of new interventions for the NHS.

Abstract 1389

**UTILIZATION OF EMERGENCY DEPARTMENTS (ED) IN CANCER PATIENTS IN MANITOBA, CANADA**

**Authors:**

Mark Smith\(^1\), Lisa M Lix\(^2\), Harvey Quon\(^3\), Heather Prior\(^4\), Say Hong\(^4\), Ina Koseva\(^4\)

\(^1\)Manitoba Center for Health Policy, University of Manitoba, \(^2\)University of Manitoba, \(^3\)Cancer Care Manitoba, \(^4\)Manitoba Center for Health Policy

**Presenting Author:** Mark Smith

**Corresponding author email:** mark_smith@cpe.umanitoba.ca

**Objectives:** The journey of a cancer patient through the healthcare system can be complex. The purpose of this research was to investigate emergency department (ED) utilization among individuals with a cancer diagnosis to test for differences in ED use before and after diagnosis and to predict time to death.

**Approach:** Data were from the Manitoba Centre for Health Policy and included cancer registry, hospital discharge abstracts, physician billing claims, ED visits and vital statistics death records. The study cohort included adults (18+) with selected cancer diagnoses (breast, colorectal, lung and prostate) made between 2007 and 2011. Rates of ED utilization 1 year before and up to 2 years after diagnosis were compared between cancer patients and cancer-free individuals matched 1:1 on age, sex and Charlson comorbidity score using generalized estimating equations. The impact of ED use on time to death was tested using a multivariable Cox proportional hazards regression model.

**Results:** When comparing ED utilization between breast \((n=1549)\), colorectal \((n=1295)\), lung \((n=1383)\), and prostate \((n=1194)\) cancer cases and their matches there were no significant differences for the year prior to diagnosis but elevated use for all sites in the one-month period before and after diagnosis, with relative risks \((RR)\) from 1.44 \((\text{breast})\) to 4.12 \((\text{lung})\), and up to two years following diagnosis, \((RRs\) of 1.18 \([\text{prostate}]\) to 1.86 \([\text{lung}]\)). ED use in the year prior to diagnosis was a significant predictor of time to death for colorectal \((\text{hazard ratio [HR]} 1.06, 95\% CI 1.01-1.12)\) and prostate \((HR 1.15, 95\% CI 1.05-1.27)\). Following diagnosis, ED use was significantly associated with time to death for breast \((HR 1.20, 95\% CI 1.10-1.31)\) and lung \((HR 1.07, 95\% CI 1.03-1.11)\).

**Conclusion:** The pattern of ED utilization varies with the duration of time from diagnosis and the type of cancer. All cancer sites exhibited increased ED use around the time of diagnosis. Cancer patients may benefit from interventions to ensure ready access to emergency services following diagnosis.
Abstract 1560  

CONGENITAL HEART DEFECT HOSPITALIZATIONS IN ARKANSAS OVER THE ADULT LIFESPAN, 2004 THROUGH 2012

Authors:  
James Robbins, RT Collins, A Goudie  
University of Arkansas for Medical Sciences

Presenting Author: James Robbins

Corresponding author email: robbinsjamesm@uams.edu

**Background:** Advances in pediatric cardiac surgery techniques and neonatal cardiovascular intensive care have resulted in many infants born with congenital heart defects (CHD) now surviving into adulthood who would have died 30 or more years ago. As a consequence it is now estimated that there are more adults living with CHDs than children with CHDs.

In absence of a population-based registry of adult cases with CHD, linking existing data sources for unique patients is required to understand the consequences of the condition as patients enter adulthood. Admissions to the hospital provide one view of the potential health difficulties experienced by adults with CHD. In this report we use hospital discharge data linked across nine years to assemble a cohort of adults diagnosed with CHD during a hospitalization and follow their subsequent hospitalizations. We determine how likely patients are to be re-hospitalized during this period, the primary diagnoses associated with their hospitalizations, and the overall charges for CHD and non-CHD hospitalizations.

**Methods:** Data come from the Arkansas State Inpatient Database (SID) for the years 2004 though 2012. The Arkansas SID includes information on all discharges from all Arkansas hospitals. A unique personal identifier allows individuals to be linked across hospitalizations. All patients 18 years of age or older in 2004 with any discharge diagnosis of congenital heart defect (ICD-9 745 – 747) during any hospitalization from 2004 through 2012 were identified.

**Results:** A total of 4,391 patients were hospitalized in the state with a CHD diagnosis during the 9-year study period. Most (85%) had only a single hospitalization where a CHD was noted, 11% had two CHD hospitalizations, and 5% had 3 or more CHD hospitalizations. A critical congenital heart defect was noted among 13% of the cohort and 24% underwent at least one cardiac procedure. Over three quarters (79%) had one or more hospitalization for conditions other than a CHD. Public health insurance covered the care of 57%. The most common comorbid conditions identified during hospitalizations were hypertension (71%), fluid and electrolyte disorders (44%), deficiency anemias (32%), valvular disease (32%), chronic pulmonary disease (31%) and congestive heart failure (31%). Median hospital charges were $16,662.

**Discussion:** By linking hospital discharge data for the same patient over time within a statewide system of care, this study contributes to an understanding of the health consequences of individuals with congenital heart defects as they age. Cardiovascular conditions are frequent reasons for hospitalizations.
Abstract 1607  

HOSPITALISATION FOR RESPIRATORY INFECTIONS IN INFANCY PREDICT THE OCCURRENCE OF LATER RESPIRATORY MORBIDITY IN EARLY CHILDHOOD

Authors:  
Nicholas de Klerk¹, Hannah C Moore², Graham L Hall³  
¹Centre for Biostatistics, Telethon Kids Institute, The University of Western Australia, Perth, Australia, ²Wesfarmers Centre of Vaccines and Infectious Diseases Telethon Kids Institute, The University of Western Australia, Perth, Australia, ³Paediatric Respiratory Physiology, Telethon Kids Institute, The University of Western Australia, Perth, Australia

Presenting Author: Nicholas de Klerk

Corresponding author email: Nick.deKlerk@telethonkids.org.au

Introduction: Hospital admissions for acute respiratory infections (ARI) in infancy are common and cause significant morbidity. Through following a birth cohort to age 10, we aimed to quantify the relationship between infant ARI and subsequent development of respiratory morbidity in early childhood.

Methods: Population-based longitudinal routinely collected hospitalisation data were linked to perinatal, birth and death records for 145,580 Western Australian children from 1996-2005. We conducted Cox regression analyses using various combinations of respiratory disease diagnoses, to assess the risk that recurrent ARI in infancy (0-12 months of age) has for later respiratory hospitalisation from age 3 to age 10.

Results: Admissions for ARI in infancy were significantly related to later respiratory morbidity before (Hazard Ratio, HR 3.5; 95% confidence interval, CI 3.1-3.8) and after (HR 3.0; 95%CI 2.6-3.4) adjusting for known risk factors including maternal smoking during pregnancy, season of birth, delivery mode and gestational age. There was an increasing dose-response with the number and length of infant ARI hospitalisations. Although there was increasing risk with decreasing gestational age, there was no change in the effects of infant admissions with gestational age. Results were similar when later respiratory morbidity was restricted to asthma hospitalisations only.

Discussion: Recurrent hospitalisation for ARI in infancy significantly increased the risk of respiratory morbidity and asthma requiring hospitalisation after age 3 years in a dose-dependent fashion. The increase in relative risk was not modified by gestational age. Efforts to reduce the occurrence of infant ARI are likely to lead to significant public health benefits.
Abstract 1406  A MOBILE APP TO ENABLE INTERNATIONAL DATA ACQUISITION FOR THE UK REGISTER FOR OUTCOMES OF PATIENTS WITH EPILEPSY AND INTELLECTUAL DISABILITY ON ANTI-EPILEPTICS

Authors: Rohit Shankar¹, Craig Newman², John Zajicek³
¹Consultant Neuropsychiatrist. ²Research Fellow. ³Professor

Presenting Author: John Zajicek

Corresponding author email: J245@st-andrews.ac.uk

Around 50 million people worldwide and about 600,000 in UK have epilepsy. Nearly 80% of the people with epilepsy are found in developing regions. 25 to 30% of patients with Epilepsy (PWE) have Intellectual Disability (ID). Individuals with ID tend to have various mental health and physical health co-morbidities and as a result can respond differently to medications when compared with the general population. There are very few studies which looked systematically at the influence of AEDs in PWE and ID/PDD. In the last 15 years there is no major research looking at the influence of the newer AEDs in this vulnerable population. In addition to the lack of research, no structured method exists for safely and effectively examining the safety of new AEDs in populations with an ID/PDD.

A register is set up to ascertain the safety of newly licensed anti-epileptic drugs (AEDs) in patients with epilepsy (PWE) and intellectual disability (ID). The Register has received Ethics approval from the National Research Ethics Service UK to run for 10 years with permission to use the approved methodology on all AEDs licensed in the UK post 2000. It is a Phase IV Multicentre observational study. The register looks to compare longitudinally outcomes of PWE with ID and PWE without ID on safety, tolerability, efficacy, seizure frequency and intensity.

A major problem in any study is recruiting sufficient numbers representative as best as possible of the diversity of human race in order to have scientific vigour and confidence in postulating findings.

A mobile application could be developed in order to harness data internationally for the study. Smartphone ownership is prolific across the world especially in developing countries, with rates of uptake forecast to expand exponentially in the next decade. It was estimated that phone users in India and China will together have bought more than 500 million smartphones in 2014, comprising half of the total that was sold in 47 key countries, according to a new forecast. They would have added 400 million new users of internet-enabled phones to the global network.

A smartphone based application that supports direct data transfer into an international registry is a relatively effortless and very portable solution. Approaches of this type are easily implemented and made immediately accessible, worldwide. This project aims to consider the feasibility of such an approach, as the foundation of an extension of an existing UK based register.
Abstract 1555  
NHS SCOTLAND INFECTION INTELLIGENCE PLATFORM (IIP): IMPROVING PATIENT OUTCOMES THROUGH IMPROVED INFORMATICS

Authors: Charis A Marwick¹, M Bennie², D Nathwani³, W Malcolm²

Presenting Author: Nicosha de Souza

Corresponding author email: c.z.marwick@dundee.ac.uk

Aim: Community and healthcare associated infections are a significant health burden and are a priority for the National Health Service (NHS) in Scotland. NHS National Services Scotland (NSS) currently hosts a wide range of infection related data on behalf of NHS Scotland. The NHS Scotland Infection Intelligence Platform (IIP) is being developed to improve patient outcomes and reduce harm from infection through innovative data integration to support clinicians across the NHS in Scotland.

Methods: The IIP will support clinicians by providing integrated information on infection (for example by combining risk factors, demographics, healthcare activity, medicines use and clinical data) by linking together the many strands of national and local infection information collected across NHS Scotland. NSS has been tasked with the development and implementation of IIP and is supported by a broad coalition of clinicians and national stakeholders.

What this will mean for clinicians and patients: Better use of our national and local data through a “collect once and use often” approach will enable better, faster and more efficient information provision to clinicians to improve care, reduce harm and reduce variation across NHS Scotland. See attached diagram for case vignette on the type of questions the IIP will be able to answer to support clinicians to provide safe, effective and patient centred care.

What this will mean for the service and Scotland: The creation of the first Infection Intelligence Platform in the UK is an exciting innovation in the prevention and treatment of infection and confirms Scotland’s position as a world leader in healthcare informatics.

Abstract 1685  
ASSESSING THE IMPACT OF DATA MODELLING IN REAL-TIME TRAUMATIC BRAIN INJURY MONITORING TO IMPROVE PATIENT OUTCOMES

Authors: Laura Moss¹, Annie O’Donnell²
¹NHS Greater Glasgow & Clyde, University of Glasgow, IDEAS Research Group, ²Aridhia, Data Science Team

Presenting Author: Laura Moss

Corresponding author email: Laura.Moss@glasgow.ac.uk

Background: A pilot programme involving NHS Greater Glasgow and Clyde and University of Glasgow’s IDEAS collaboration (led by Dr Laura Moss) and Aridhia investigated real-time data in Traumatic Brain Injury (TBI) patients to improve detection and prediction of specific clinical events, and inform evidence-based decision making.

Challenge: Patients in Neuro Intensive Care Units have their vital signs monitored on a 24 hour basis. Physiological monitoring equipment captures data in waveform quality and can collect and output data at high sampling rates (e.g. 500Hz) across multiple vital channels. Linking this data to lower frequency clinical data captured in parallel allows clinical events to be modelled in the waveform component, the analysis of which can identify features and patterns of transients that could be attributed to specific clinical events. To enable these models to be used in clinical practice, the rate of sampled raw physiological patient data has to be reduced (down-sampled). However, due to computer equipment limitations, down-sampling algorithms are prohibitively slow.

Requirement: IDEAS wanted to swiftly translate physiological and clinical data modelling into clinical
practice, and so developed a down-sampling algorithm using the R statistical package that performs peak-to-peak waveform signal processing. However, running the algorithm on a single patient’s data collected over 12 days took 16 hours to process; clearly an unrealistic timescale for researchers who need to tune an algorithm, develop and compare different models, or provide timely evidence to inform treatment decisions.

**Method**: Aridhia and IDEAS collaborated to re-engineer the down-sampling algorithm and explore methods to utilise parallel compute-and-data-storage in a small cluster. A data sample, collected across 7 channels and spanning 470 million rows, and the existing algorithm were deployed into Aridhia’s AnalytiXagility data platform. Approaching the problem in 3 stages, the team reviewed and optimised, first, the algorithm for a single processor, the data storage model and, lastly, parallel computation of the algorithm.

**Results**: By employing processing techniques designed to determine the optimal section size from the data sample, reduce the number of algorithm iterations, and exploiting AnalytiXagility’s multiple processor cores by running the algorithm in parallel, the time taken to process the original sample dropped from 16 hours to 48 minutes. The pilot’s success led to the award of further funding by Innovate UK. This will facilitate development of an app which will enable Neuro-ICU clinical teams to swiftly analyse high-frequency patient data at the bedside to improve outcomes and reduce healthcare delivery costs.

### STATISTICS AND ANALYTICS

**Abstract 1433**

**VISUALISING LINKED HEALTH DATA TO EXPLORE SERVICE USE AROUND PREVENTABLE HOSPITALISATIONS**

**Authors:** Michael Falster¹, L Jorm¹, A Leyland²

¹UNSW Australia, ²University of Glasgow

**Presenting Author:** Louise Jorm

**Corresponding author email:** m.falster@unsw.edu.au

**Background:** Data visualisations have the potential to enhance understanding of complex patterns of information. However, few tools are available for visually exploring longitudinal health data. Those that are available are primarily used for clinical decision support and quality assurance, and despite the increasing availability of longitudinal and linked electronic health records, such visualisations remain underutilised within health services research.

This presentation demonstrates the utility of a simple new data visualisation, by exploring patterns of health service use around the time of ‘preventable’ hospitalisations. These are used internationally as an indicator of accessibility of primary care services, but the actual use of services leading up to hospitalisation is largely unknown.

**Methods:** Linked data on hospital admissions, emergency department (ED) presentations, deaths, and Medicare claims for visits to general practitioners (GP) for 266,950 participants in the 45 and Up Study,* NSW Australia, were used to plot ‘Trajectories of Individual Patient Service use’ (TIPS). For each study participant, each type of service was plotted against time, with different coloured dots and lines representing hospital stays, GP consultations, and deaths. To provide structure to the TIPS, participants were sorted on factors related to their use or need of health services, such as number of hospitalisations, time of first hospitalisation, length of stay and self-rated health. TIPS were produced for both the whole cohort and sub-cohorts of admitted patients, and presented both over calendar time and centred on the period surrounding preventable hospitalisation.

**Results:** The use of TIPS revealed common usage patterns across health services, with a clustering of GP visits in the lead up to, and following, a preventable hospitalisation. People with more preventable hospitalisations, or a longer length of hospital stay, also had higher numbers of GP visits, ED presentations and other types of hospitalisation, and were more likely to be in the last year of life. The
The visual interpretation of the TIPS was consistent with comparable results from descriptive analyses. Policymaker and practitioner audiences engaged with TIPS visualisations and found them more accessible than traditional epidemiological metrics such as rates and risk ratios.

**Conclusions:** The TIPS visualisations provided a powerful tool for exploring and displaying patterns of health service use. They can easily be produced using standard statistical software, and have the potential to be broadly utilised across linked data and health services research.

*Medicare data was provided by the Department of Human Services. Data Linkage was performed by the NSW CHeReL.

---

**Abstract 1511**

**Predicting Initial Response to First-Line Glucose-Lowering Therapies in UKPDS Patients with New-Onset Type 2 Diabetes Mellitus – A Precision Medicine Approach.**

**Authors:**

O F Agbaje¹, R L Coleman², R R Holman³

¹University of Oxford, ²University of Oxford, Diabetes Trials Unit

**Presenting Author:** O F Agbaje

**Corresponding author email:** orlunsola.agbaje@dtu.ox.ac.uk

**Background:** Current guidelines base glucose-lowering treatment choices for type 2 diabetes mellitus (T2DM) on the mean glycaemic responses observed in clinical trials. In contrast, the precision medicine approach aims to identify the therapy which will provide the best response for a given individual. Precision medicine has become well-established in oncology where specific biomarkers and genetic profiling of cancer tissues can now be used to determine the most appropriate treatment on a patient-by-patient basis.

For some less frequent types of diabetes, e.g. HNF1A-MODY, genetic profiling has been highly successful, but for the majority of patients with T2DM there are no robust indicators of therapy response. We have investigated possible predictors of initial glycaemic response to the initiation of first-line glucose-lowering therapies in T2DM.

**Methods:** We included 2339 of 5102 UKPDS patients who had been assigned to, and remained on, a single monotherapy for one year from diagnosis of T2DM, and who had the requisite data. Of these, 453 (19%) were randomised to chlorpropamide, 391 (17%) to glibenclamide, 1047 (45%) to insulin and 448 (19%) to metformin. The primary outcome examined was the absolute change in HbA₁c at one year.

We developed internally validated clinical prediction models to determine the phenotypic characteristics and biomarker indicators that best predicted a change in HbA₁c at one year. These were identified from the totality of variables available by initial univariate selection of items of interest, and then by performing multivariate regression models to identify those covariates that remained independently and significantly associated with the absolute change in HbA₁c. All models were internally validated using a bootstrap procedure.

**Results:** The strongest predictor of HbA₁c change was baseline HbA₁c (p ≤ 0.001 for all therapies). Additional predictors were: Chlorpropamide, interactions between HbA₁c, age, ethnicity, HDL-cholesterol and HOMA-derived insulin sensitivity; Glibenclamide, interactions between HbA₁c, age, weighted-dose, and HDL-cholesterol and triglyceride; Insulin, interactions between HbA₁c, age, weighted-dose and drug adherence; Metformin, HOMA-derived beta-cell function and interactions between age, cholesterol, triglyceride and BMI. The degree to which variance was explained (R²), after adjusting for over-fitting, ranged from 42-45%.

**Conclusions:** Response to first-line glucose-lowering therapies can be predicted to a modest degree according to assigned modality of treatment, with baseline HbA₁c being the strongest indicator.
**DISTINGUISHING DISEASE ENDOTYPES USING LONGITUDINAL PROBABILISTIC MODELLING**

**Authors:**
Rebecca Howard\(^1\), Mattia Prosperi\(^2\), Adnan Custovic\(^3\), Magnus Rattray\(^4\)

\(^1\)Health e-Research Centre, University of Manchester, UK, \(^2\)University of Florida, Gainesville, USA, \(^3\)Centre for Respiratory Medicine and Allergy, Institute of Inflammation and Repair, University of Manchester & University of Hospital of South Manchester, Manchester, UK, \(^4\)Faculty of Life Sciences, University of Manchester, UK

**Presenting Author:** Rebecca Howard

**Corresponding author email:** r.howard@postgrad.manchester.ac.uk

“Asthma” and “atopy” are increasingly recognised as being umbrella terms for sets of discrete disease entities (referred to as “endotypes”). Data-driven approaches have been used in an effort to identify endotypes of complex diseases. Howard et al. (2015) reviewed the application of one such approach – latent class analysis – to asthma and atopy. The reviewed analyses were limited by computational tractability which typically led to two types of analysis: firstly cross-sectional analysis at single time points in order to reduce dimensionality in many variables; and secondly longitudinal analysis with a reduced, limited number of variables. However, without taking all of the variables measured across each time point into account, the ability to truly identify disease endotypes will be restricted.

Our study population are participants in a population-based birth cohort (Manchester Asthma and Allergy Study-MAAS). Children provided blood samples which underwent allergy diagnostics using a multiplex Immuno Solid-phase Allergen Chip (ISAC). This in vitro diagnostic tool for allergy is based exclusively on allergen components, and includes 112 species-specific and cross-reactive components from multiple sources including foods (egg, milk, chicken, fish, shrimp, peanut, tree nuts, seeds, pulses, fruits) aeroallergens (grasses, trees, weeds, animals, mites, moulds, insects), hymenoptera venoms, latex and parasites. We hypothesise that there is an underlying complexity and heterogeneity in the longitudinal patterns of the development of IgE responses to distinct allergen components, and that different longitudinal patterns may be associated with different allergic diseases.

Using the ISAC data, our group completed both cross-sectional and longitudinal analyses. The first of these analyses was cross-sectional, performed on all 112 allergen components at age 11 years (Simpson et al, JACI 2015). Latent variable modelling of this data identified three patterns of IgE responses, each including different protein families; different patterns of response to components of multiple allergens had differing associations with clinical disease. The second analysis was performed longitudinally across three of the time-points (ages 5, 8 and 11), to investigate the evolution of IgE responses to 7 components of timothy grass and 8 components of dust mite (Custovic et al, JACI 2015). This work demonstrated that the nature of developmental longitudinal trajectories of IgE responses differed between grass and mite allergen components.

The limitations experienced with computational tractability are being addressed, and we aim to build upon our previous work to longitudinally model the ISAC data for all 112 components.
GUIDELINES FOR THE REPORTING OF STUDIES CONDUCTED USING OBSERVATIONAL ROUTinely-COLLECTED HEALTH DATA (RECORD): AN EXTENSION OF THE STROBE REPORTING GUIDELINES

Authors: Sinead Langan¹, A Guttmann², K Harron³, D Moher⁴, I Petersen⁵, Ht Sorensen⁶, L Smeeth⁷, S Nicholls⁸, E von Elm⁹, El Benchimol¹⁰

¹London School of Hygiene and Tropical Medicine, ²Institute for Clinical Evaluative Sciences, Department of Paediatrics, University of Toronto, Toronto, Canada, ³London School of Hygiene and Tropical Medicine, London, UK, ⁴Ottawa Hospital Research Institute, Department of Epidemiology and Community Medicine, University of Ottawa, Ottawa, Canada, ⁵Department of Primary Care and Population Health, University College London (UCL), London, UK, ⁶Department of Clinical Epidemiology, Aarhus University, Aarhus, Denmark, ⁷Department of Non-Communicable Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, UK, ⁸Department of Epidemiology and Community Medicine, University of Ottawa, Ottawa, Canada, ⁹Cochrane Switzerland, Institute of Social and Preventive Medicine, University of Lausanne, Lausanne, Switzerland, ¹⁰Institute for Clinical Evaluative Sciences, Department of Pediatrics and School of Epidemiology, Public Health and Preventative Medicine, University of Ottawa, Ottawa, Canada

Presenting Author: Sinead Langan

Corresponding author email: sinead.langan@lshtm.ac.uk

Objectives: Routinely-collected health data, obtained for administrative and clinical purposes without specific a priori research goals, are increasingly used for research purposes. The rapid evolution and availability of these data have revealed issues not addressed by existing reporting guidelines, such as STROBE. The statement for Reporting of studies Conducted using Observational Routinely-collected health Data (RECORD) was created to extend STROBE and fill these gaps.

Methods: A large, international group of stakeholders (http://record-statement.org/group.php) was recruited to reflect the diversity of relevant researchers and consumers of research. Stakeholders participated in two consecutive modified electronic Delphi surveys. The first survey identified themes deemed important for the RECORD statement, and was analyzed using qualitative methods. The second survey requested quantitative prioritization of the themes for each manuscript heading (abstract, introduction, methods, results and discussion). The surveys were followed by a face-to-face meeting of the RECORD working committee, and re-engagement with stakeholders via an online commentary process. The drafting of the final RECORD reporting guidelines was informed by this multi-stage multi-disciplinary participatory process.

Results: The qualitative survey (76 responses from 123 surveys sent) generated 10 overarching themes and 13 specific themes derived from existing STROBE categories. The highest-rated overall themes for inclusion were: “Disease/exposure identification algorithms”; “Characteristics of the population included in databases”; and “Characteristics of the data”. In the quantitative survey (71 responses from 135 surveys sent), the importance assigned to each of the compiled themes varied depending on the manuscript section to which they were assigned. Following the working committee meeting, online ranking by stakeholders provided feedback for revision of the final checklist, which will be presented at the Farr Institute International Conference. Following publication, the checklist will be available at record-statement.org and further feedback will be requested to ensure that these guidelines represent a “living document”.

Conclusions: The RECORD statement addresses issues specific to observational research using routinely-collected health data. Stakeholder responses ascertained unique aspects of research using these data, including the need for better reporting of methodological issues. Through implementation of RECORD, authors, journals editors, and peer reviewers can encourage transparency of research reporting and improve our ability to judge validity and generalizability of findings and potentially allow for greater reproducibility.
Abstract 1629

IDENTIFYING IMPORTANT RISK FACTORS FOR PHENOTYPING PATIENTS WITH ARTHROPATHY CONDITIONS FROM HIGH DIMENSIONAL IMBALANCED ROUTINE DATA

Authors: F Fernandez-Gutierrez, J Kennedy, S Zhou, M Atkinson, S Brophy

College of Medicine. Swansea University

Presenting Author: J Kennedy

Corresponding author email: F.FernandezGutierrez@swansea.ac.uk

Introduction: Linkage of routine and administrative databases from multiple sources provides an advantageous form of understanding chronic diseases, such as arthropathy conditions. However, selecting good potential predictors, given a certain condition from a patient’s history with huge health records, can be challenging, particularly with small prevalence proportion, which leads to a high-dimensional imbalanced data space. Data mining feature selection (FS) methods can be used to overcome this problem. In this study, a general practitioner database, the Abertawe Bro Morgannwg University (ABMU) Health Board (Wales, UK) with Rheumatoid Arthritis (RA) patients was linked to a rheumatoid specialised database CELLMA (Riomed Ltd., UK) for RA identification. We compared the performance of three FS methods – Binomial distribution, $\chi^2$-statistic and Information Gain – on selecting risk factors to classify whether a patient is diagnosed with RA or not.

Methods: The linked dataset containing 13659 patients (1904 with RA, 14%) was split in training (60% of the data), testing (20%) and validation (20%) datasets. An initial list of 36243 possible predictors described by Read codes was included. These initial risk factors were ranked by each FS method. The decision tree C5.0 was used for classification. The training dataset was used for constructing the classifier and the testing was used for selecting the important factors by evaluating them in an incremental manner with one predictor in each step until all ordered predictors were included in the model. Once an optimal classifier is identified, its performance is confirmed using the validation dataset. Accuracy and positive predictive value (PPV) measurements were used to compare the performance.

Results: All methods selected similar predictors although with different order; only Binomial and $\chi^2$-statistic differ in almost 10% predictors among the first 1000. IG, with a selected combination of 169 predictors, and $\chi^2$ with 125, achieved optimal performance earlier than other methods, achieving accuracy of 93.19% and 93.15% and PPV of 80.98% and 80.86% respectively for the testing dataset. In contrast to other methods that selected lab codes as top predictors, IG and $\chi^2$ selected drug codes for Methotrexate –‘h341.’- and folic acid –‘i332.’, which are often prescribed together to decrease side-effects during RA treatments.

Conclusion: These preliminary results showed the importance of choosing an appropriate method to be able to discriminate between potential predictors. Ongoing tests are being carried out to compare results when applying this method to Ankylosing Spondylitis, which has a prevalence of 1.9% over the ABMU population.
DEVELOPING AND VALIDATING ALGORITHMS TO IDENTIFY RHEUMATOID ARTHRITIS FROM ROUTINE ADMINISTRATIVE DATA – A DATA MINING APPROACH.

Authors: S Zhou, J Kennedy, F Fernandez-Gutierrez, M Atkinson, R Cooksey, S Denaxas, S Brophy
Swansea University

Abstract 1693

Introduction: This investigation concentrates on constructing an algorithm for identifying cases of Rheumatoid Arthritis (RA) from General Practitioner (GP) data. This will be achieved by employing data mining techniques on secondary care data from rheumatology clinics within Abertawe Bro Morgannwg University (ABMU) Health Board as a gold standard. This approach will allow the identification of significant predictors from the GP data to determine whether a person is positive for RA. By employing these techniques we can minimise the amount of clinician time required for the algorithm development and speed up the development cycle. To determine if the algorithm is suitable for purpose, comparisons to a clinical derived algorithm will be made.

Methods: Patients within rheumatology clinics in ABMU had their GP records linked and data mining performed on them, which comprised the following steps:

- Scoring and selecting preliminary predictor variables based on their relative frequency.
- Aggregation of predictor variables utilising random forests, supported by expert knowledge from clinicians with expertise of the condition.
- Development of the initial algorithms to classify disease: The aggregated predictor variables are modelled utilising a decision tree approach in order to generate inference rules which can classify the occurrence of disease.
- Validation and selection of appropriate algorithms. Sensitivity, specificity and predictive values are utilised to determine which algorithm performs best and under which circumstances.

Results: All patients had outpatient records with a specialist rheumatologist and linked GP data. The population from ABMU comprised of the training/testing dataset, while the Cardiff population was employed for the validation.

The Cardiff (validation) dataset had a RA prevalence of 27.3 %, the proposed method achieved a 85.6 % positive predictive value, 86.2 % sensitivity and 94.6 % specificity. This improved the performance of the clinician derived algorithm with 78.4 % positive predictive value, 86.2 % sensitivity and 91.1 % specificity in the validation dataset.

Applying the model to the entire ABMU population comprising of 77 practices from 2000 to 2010, estimated an overall prevalence of 0.84 % (95 % confidence interval [0.798 %, 0.873 %]), an incidence of 0.053 % (95 % confidence interval [0.042 %, 0.069 %]), and the PPV of 76.7 %.

Conclusion: The use of data mining to develop disease identification algorithms from routine administrative data enables high quality algorithms to be developed rapidly and cost-effectively. The data mining approach achieved comparable results to the algorithm derived from clinician expert knowledge.
Abstract 1744

**SUB-CLASSIFYING ISCHAEMIC STROKE WITH RECORD LINKAGE AND GENETIC INSTRUMENTS**

**Authors:** Wilfred Bonney, Kennedy C. McCaig, Philip Appleby, Alex S. F. Doney
*University of Dundee*

**Presenting Author:** Kennedy C. McCaig

**Corresponding author email:** k.c.mccaig@dundee.ac.uk

**Background and Objectives:** Ischaemic stroke is a common endpoint of several different and complex disease pathologies. Sub-classification into patho-aetiological subtypes is critical to subsequent secondary prevention and clinical management. Established methods of stroke classification such as the TOAST (Trial of ORG 10172 in Acute Stroke Treatment), rely on purely clinical information. This study aims to (a) use large bio-resources linked to electronic medical records (EMRs) to define ischaemic stroke patho-aetiological sub-types; and (b) use this resource to investigate the extent to which genomic data could potentially aid the clinical classification of ischaemic stroke and hence, improve choice of therapeutic intervention and overall patient care.

**Methods:** A population of ischaemic stroke patients was extracted from the GoDARTS bio-resource and further sub-classified based on the TOAST classification using available EMR. Matched case-control studies based on age and gender were assembled and analysed using logistic regression models. PITX2 (rs6843082) and HDAC9 (rs2107595) single nucleotide polymorphisms (SNPs), which are robustly associated with CE stroke and LAA stroke respectively, were used as “genetic instruments” to validate and optimise the stroke sub-type populations. Beta weighted genetic risk scores, specific to stroke sub-type, were calculated and grouped into low, medium and high risk categories. Longitudinal analysis was used to assess the genetic risk scores’ ability to predict specific stroke subtype.

**Results:** 1,140 patients with ischaemic stroke were identified from the GoDARTS bio-resource who had genomic data, and further sub-classification yielded 249 patients with cardioembolic (CE) stroke and 464 patients with large artery atherosclerotic (LAA) stroke. In matched case-control analyses, significant associations were found between the PITX2 SNP and CE stroke ($OR=1.36, 95\% CI=1.06-1.76, P=0.016$); and the HDAC9 SNP and LAA stroke ($OR=1.43, 95\% CI=1.11-1.86, P=0.007$). The genetic risk scores also showed significant associations for CE stroke ($OR=1.93, 95\% CI=1.40-2.66, P=0.0001$) and LAA stroke ($OR=1.69, 95\% CI=1.19-2.39, P=0.003$). Longitudinal analyses revealed that the specific genetic risk scores were accurate predictors of stroke sub-type. Intriguingly, the CE genetic risk score was also associated strongly with a population of patients with cryptogenic stroke (i.e. Non-CE and Non-LAA stroke).

**Conclusion:** It is possible to derive ischaemic stroke subtypes using EMR and validate them using genetic instruments. Such instruments could potentially enhance existing ischaemic stroke sub-type clinical classification and as a result improve targeted (personalised) therapy.
DATA LINKAGE METHODS

Abstract 1694
GRAMPIAN DATA SAFE HAVEN: IMPLEMENTING A TECHNICAL SOLUTION TO FACILITATE SAFE AND SECURE ACCESS TO LINKED HEALTH AND CARE DATA.

Authors: K Wilde¹, S Gordon¹, S Hall¹, A Coutts¹, A Souter¹, A Wozniak¹, P Rebecca¹, J Evans¹, J Rodger¹, C Black²
¹University of Aberdeen, ²Farr Institute @ Scotland

Presenting Author: K Wilde

Corresponding author email :

Background: The growth in electronic data capture within health and social care offers huge potential for population health benefits through the use of linked clinical datasets for health surveillance, improved care planning and epidemiological research. Scotland has a world class reputation for rich, routinely collected population data, particularly in relation to health. Key institutional challenges are to support both data custodians and researchers to maintain patient confidentiality and ensure best practice in research governance; whilst enabling important health and population research.

Aims: To establish the Grampian Data Safe Haven (DaSH) to provide a safe and secure technical environment to facilitate access to linked data for health research; and to ensure best practice in research governance for data linkage research.

Methods: A shared NHS and University of Aberdeen infrastructure enabled identifiable information to be retained only within the NHS. Research Co-ordinators support researchers to plan data management within the technical environment. An experienced team of analysts support data linkage and provide project-specific linked anonymised datasets. A secure analytics platform, accessible from an approved researcher’s desktop, ensures rigorous access controls to the linked anonymised data.

Results: Grampian DaSH facility has supported 108 approaches for potential research projects. Seventeen novel linkages have been created with varying degrees of complexity, involving 30 different data sources. A growing suite of software is available for researchers. Data security and Data Management Planning have been placed at the heart of every project designed around based practices in safe: settings, researchers, projects and outputs.

Conclusion: Analysis of linked health and population datasets offers unprecedented opportunities for health surveillance and research to improve health and health services. IT infrastructure has supported the ability to deliver best practices in research governance for managing the linkage, access to and storage of data.
NEW CHALLENGES AND ISSUES WITH THE RECORD LINKAGE OF AUSTRIAN HEALTH INSURANCE DATA OF DIFFERENT SOURCES AND THE NEED FOR A RECORD LINKAGE FRAMEWORK

Authors: Barbara Glock\textsuperscript{1}, Florian Endel\textsuperscript{2}, Gottfried Endel\textsuperscript{3}, Niki Popper\textsuperscript{4}

\textsuperscript{1}Vienna University of Technology, Institute for Analysis and Scientific Computing, \textsuperscript{2}Institute for Analysis and Scientific Computing, Vienna University of Technology, Wiedner Hauptstrasse 8, 1040 Vienna, Austria, \textsuperscript{3}Main Association of Austrian Social Security Organisations, Kundmannsgasse 21, 1031 Vienna, Austria, \textsuperscript{4}dexhelpp, Neustiftgasse 57-59, 1070 Vienna, Austria

Presenting Author: Barbara Glock

Corresponding author email: barbara.glock@dwh.at

Due to data privacy issues, routinely collected data of different sources is pseudonymized (e.g., MBDS minimum basic data set from the Federal Ministry of Health, which up to 2015 even don’t have a personal identifier). This makes statistical analysis for decision support and health care planning very difficult, since no statements on patient pathways can be made. Data from insurance carriers (FoKo) is event based: each time a hospital reports to the insurance carrier a new data entry is generated. To enable efficient, significant and quality assured data analysis for patient centred assertions a record linkage of these episodes is required.

For historical data a linkage has been done before, but there are some challenges for the new data sets: in MBDS data for the whole of Austria is available, but in FoKo only data for Lower Austria; in FoKo a hospital stay may be split off in more data entries, due to intermediate reportings from the hospital; reported diagnoses of split episodes may differ from each other; and many more. We propose as first step of the record linkage to determine the already quality assured matching variables (e.g., birth year, gender), together with a quality check for same episodes (equal in all variables) with different person-IDs. Then a sequencing for split off episodes in FoKo, which are then joined, can be established. The minimum on variables that identify the data entry uniquely is determined. Finally, the record linkage is done, by firstly checking for matches in all chosen variables. If a unique match exists, the episodes are matched. Then an iterative process starts where different variables are varied. Quality checks after each run are included and this process can be applied more than once on the still remaining unmatched episodes.

The results of this record linkage are on the one hand the linked episodes for further statistical analyses and on the other hand statements on the quality of the given data, that are established through the preparing phases and also in the matching phases of this process.

Once the record linkage is done for existing data, the work is not finished yet, since more data will be available soon. This pleads for the establishment of a record linkage framework, for which the authors of this paper are preparing semi-automated fundamentals within the K-Project dexhelpp in COMET – Competence Centers for Excellent Technologies, funded by BMVIT, BMWGJ and transacted by FFG.
Abstract 1495

ACCURACY AND COMPLETENESS OF PATIENT PATHWAYS: FINDINGS FROM A NATIONAL LINKAGE STUDY

Authors: James Boyd¹, Ap Brown¹, Sm Randall¹, Am Ferrante¹, M Gillies¹, Jb Semmens²
¹Curtin University, PHRN, ²Curtin University, CPHR

Presenting Author: James Boyd

Corresponding author email: j.boyd@curtin.edu.au

Background: The technical challenges associated with national record linkage in Australia, and the extent of cross-border population movements, have been explored as part of a pioneering research project. The project involved linking State-based hospital admission records and death registrations across Australia for a study exploring mortality after hospital admissions between 1999 and 2008.

Methods: The project required linkage of over 44 million morbidity and mortality records from four Australian States using probabilistic methods. Complexity measures were calculated to measure the efficiency of the linkage. Accuracy of the linkage was measured through a comparison with jurisdictional keys sourced from individual States and the extent of cross-border population movement between these States was also assessed.

Results: Data matching identified almost twelve million individuals across the four States. The percentage of individuals from one State with records found in another ranged from 3-5%. Using jurisdictional keys to measure linkage quality, results indicate a high matching efficiency (F measure 97 to 99.

The project identified a sizeable ‘mobile’ population with hospital records in more than one State. Research studies that focus on a single jurisdiction will under-enumerate the extent of hospital usage by local population. It is important that researchers understand and are aware of the impact of this missing hospital activity on their studies.

Conclusions: The results demonstrate the feasibility and accuracy of undertaking cross jurisdictional linkage for national research. The benefits are substantial, particularly in relation to capturing the full complement of records in patient pathways as a result of cross-border population movements. The project highlights the need for an efficient and accurate data linkage system to support national research across Australia.

DATA SCIENCE

Abstract 1519

The Farr Commons: Healthcare assets all in one place - linked and discoverable!

Authors: John Ainsworth, Norman Morrison, Gamble Matthew, Carole Goble, Iain Buchan
University of Manchester

Presenting Author: Carol Goble

Corresponding author email:

The Farr Commons (www.farrcommons.org) is an online shared space for the aggregation and exchange of the digital assets created by the researchers and data scientists of the Farr Institute. Its goal is to enable knowledge to be exchanged and reused.

Data science is an end-to-end digital activity. Every action, decision or assumption made by a data scientist can be captured electronically. The knowledge captured in these digital assets may be the result of intensive human activity. However, because they are digital, given the appropriate infrastructure, the costs for dissemination and reuse are negligible in comparison.
The Farr Commons is a virtual space into which these digital assets of scientific research can be published, indexed, discovered and consumed. This can provide a step change in data science for research by enabling the resuability, reproducibility and transparency of scientific assets.

Here we present an initial prototype Farr Commons that builds upon the popular CKAN data catalogue. This initial phase the commons focuses upon discovery and citation by ensuring that there is a persistent unique identifier for each digital asset. The identifier will enable digital assets to be found, shared and attributed enabling them to be cited by other scholarly works. Each digital asset will have associated (i) provenance, which minimally, identifies the creator(s) of the asset, those that have subsequently modified it, and how it was modified, and (ii) descriptive metadata is required for each asset to facilitate indexing, discovery and reuse. These three basic rules are sufficient to create a functioning data science commons that will enhance the working practices of Farr Data Scientists.

The commons aims to fundamentally change the way scientists perform their work. It requires cultural change within the community. Currently, there is no mechanism or incentive for reusable knowledge objects nor the capability to reproduce findings by the exchange of digital objects. For most if not all scientists, the thought of exposing their data and working practices is a daunting prospect. However the goal of the commons is to make sure that the ultimate beneficiaries of open science will be the scientists themselves who will be able to scale and accelerate discovery through the use of the Farr Commons. This will lead to less duplication of effort and so enable funders to leverage and multiply their past investments in future funded activities.

Abstract 1584  
GETTING THERE IS 50.0% OF THE FUN (95% CI [47.7, 52.3]): SOFTWARE DEVELOPMENT METHODOLOGIES FOR RESEARCH

Authors:  
Daniel S Thayer¹, Richard Noyce², Arfon Rees¹, Sarah Rees¹, Luca Ruschetti¹, Ting Wang¹
¹Secure Anonymised Information Linkage (SAIL) Databank, College of Medicine, Swansea University, ²Administrative Data Research Centre - Wales, College of Medicine, Swansea University

Presenting Author:  
Daniel S Thayer

Corresponding author email:  
d.s.thayer@swansea.ac.uk

Background: Dataset analysis increases in complexity as the size and number of data sources grow. A suboptimal analysis design can be hard to correct later. Preparing data frequently takes far longer than expected, stretching deadlines. Code written for one project is not always easy to repurpose for another, even when requirements are similar. A small mistake may invalidate results. Often the project team has no visibility of whether the analyst’s work is correct. This is a major risk for data-intensive research, as high-profile cases of research errors illustrate.

These challenges have long been known in software development. The analyst team for the Secure Anonymised Information Linkage (SAIL) Databank has applied software development methodologies to health data research.

Objectives:
- Predict how long tasks will take and complete projects on schedule.
- Develop and document clear methods.
- Allow method reuse, supporting continuous improvement.
- Verify results and provide evidence of verification.

Methods: The SAIL analyst team uses scrum methodology to manage projects. Planning occurs in two-week intervals. Effort is estimated for tasks and compared against actual effort upon completion. A protocol is prepared, describing outputs and intermediate steps. Modular design divides the work into steps that accomplish one simple task. Anything useful for multiple projects can be implemented...
as reusable code. Parameterization allows variables like date ranges to be changed without rewriting the code. Separation of definitions, such as lists of clinical codes, from code allows easy change and reuse.

A test plan is prepared, describing verification for each requirement. Tests are automated, allowing them to be re-run whenever changes are made to the analysis. A test report shows whether tests have passed. Testing is supplemented by code review and review of the results by domain experts.

**Results:** The team’s ability to estimate and complete tasks on schedule has improved over time. Clear specification of the planned work and outputs has helped reduce the number of design changes needed later. Reusable code has increased the efficiency of the team and other researchers. Testing is challenging and requires a large amount of effort. However, the tests have caught many errors, and automation of the tests allows a fast response when changes to the analysis are requested.

**Conclusions:** The application of software development methodologies to research data analysis has been a learning process, but it has already been very beneficial to the SAIL team. Other researchers may find these methods useful.

---

**PUBLIC HEALTH AND EPIDEMIOLOGY METADATA: INFORMING DEVELOPMENT OF A QUALITY ASSESSMENT FRAMEWORK**

**Authors:**

Christiana McMahon¹, T Castillo², D Kehoe³, S Denaxas¹

¹Farr Institute of Health Informatics Research, London; UCL Institute of Health Informatics, ²University London Hospitals NHS Foundation Trust, London, United Kingdom, ³AIMES Grid Services CIC, Liverpool, United Kingdom

**Presenting Author:** Christina McMahon

**Corresponding author email:** christiana.mcmahon.11@ucl.ac.uk

**Background:** Researchers are increasingly linking combinations of longitudinal cohort study, genomic, and other administrative data together to produce enriched datasets through their inherent complexity and longevity. Research artefacts, such as data dictionaries (an example of metadata), are indispensable to researchers and support secondary use of clinical data.

Metadata describe the data and often the process through which the data were collected. Metadata quality plays an important role across the research data lifecycle for epidemiological and public health research. Harmonising phenotypic data can greatly benefit from good quality and standardised metadata. However, a lack in ubiquitous uptake of standards, coupled with inconsistent quality of metadata, can render analyses of these data problematic. Currently, there is a lack of robust mechanisms through which biomedical metadata quality can be assessed.

**Objectives:** To develop a novel framework for assessing metadata quality for epidemiology and public health research datasets.

**Methods:** We designed and implemented a comprehensive online survey aimed at stakeholders such as: data users, clinicians, data providers, and funding agencies. The survey ran from November 2014 until February 2015. Data were collected anonymously using the survey management software, REDCap.

**Results:** Data were collected from over 90 individuals, most of whom were located in Europe, and employed by a university as data users. Results show that most of the respondents routinely handle descriptive metadata and that quality is only sometimes assessed. Survey results also show, of those who submitted data, that accuracy was the most important dimension of metadata quality; the second and third most important qualities were accessibility and discoverability respectively.

Through the survey, we identified the following challenges associated with assessing metadata quality in biomedical research: a) lack of guidance to assist quality assessment and how quality is defined; b) lack of biomedical knowledge, subsequently negatively impacting how well metadata are understood; c) knowledge of and access to supportive software; and d) limited availability of
resources, such as time, to help facilitate quality assessment.

Conclusions: Results of the survey will inform development of a high level quality assessment framework for biomedical metadata. Given the breadth of public health and epidemiology research, application of this framework would promote a more in-depth and focused assessment of the metadata. Future steps will include evaluating the framework by applying it to a series of case studies. Recommendations for changes in research data management policy and practice will be made following the evaluation.

Abstract 1668

IDENTIFYING PATIENTS FROM ELECTRONIC MEDICAL RECORDS: A SYSTEMATIC REVIEW OF THE USE OF MEDICAL TEXT FOR CASE DETECTION

Authors:

Elizabeth Ford¹, John Carroll², Helen Smith¹, Tim Williams³, Jackie Cassell¹

¹Brighton and Sussex Medical School, ²University of Sussex, ³Clinical Practice Research Datalink

Presenting Author: Elizabeth Ford

Corresponding author: e.m.ford@bsms.ac.uk

Background: Electronic medical records (EMRs) are revolutionising health-related research, audit and service provision. In most EMRs, clinicians can code clinical information in a structured format, or enter information in free text. One key issue for study quality is the accurate identification of patients with the condition of interest in the dataset. The majority of research studies have used only coded parts of EMRs for case-detection, which may bias findings, miss cases, and reduce study quality. This review examines whether incorporating automatic text analysis into case detection algorithms can improve the quality of case identification.

Methods: Following a systematic search which returned 1815 papers, 59 studies were identified which reported on the extraction of information from free text with the stated purpose of detecting cases of a named clinical condition. Methods for extracting information from text, the technical performance of extraction algorithms and added benefits to studies were reviewed.

Results: Studies were mainly carried out in the USA, using hospital-based EMRs. Information was extracted from text for a wide range of conditions using manual review, keyword searches, rule-based algorithms and machine learning methods. There was no overall difference in algorithm performance between rule-based and machine learning methods of extraction. Seventeen studies compared the performance of case detection algorithms using only codes to algorithms combining codes and information from text. Including text resulted in a significant improvement in algorithm performance (median positive predictive value (PPV) 86% (code+text) vs 72% (codes); p=0.05; median sensitivity 75% (codes+text) vs 62% (codes), p = 0.05).

Conclusions: The extraction of information from EMR free text appears to improve the quality of case detection in comparison to algorithms using only codes, although the number of studies that could be compared was small. More harmonization within the field, such as standardised reporting of algorithm performance metrics like PPV and sensitivity, would make it easier to assess the benefits of extracting information from EMR text across studies. Some open-source information extraction systems such as MedLEE performed well across a range of conditions. The re-use of these systems would save time and effort in comparison to creating new algorithms for each study.
DATA SETS CONTAINING NATURAL LANGUAGE STRINGS ARE INCREASINGLY BECOMING AVAILABLE AS OUTPUTS FROM VARIOUS INTERNATIONAL INITIATIVES TO DIGITIZE HISTORICAL POPULATION RECORDS. ANALYSIS OF SUCH RECORDS, FOR EXAMPLE, HISTORICAL CAUSES OF DEATH, IS FACILITATED BY CLASSIFICATION TO STANDARD SYSTEMS SUCH AS ICD-10.

WE PRESENT RESULTS OBTAINED USING AN AUTOMATED CLASSIFICATION SYSTEM, WITH THE ABILITY TO EXTRACT MULTIPLE CAUSES FROM NATURAL LANGUAGE TEXT SUCH AS "BRONCHITIS AND WHOOPING COUGH".

THE SYSTEM USES A MACHINE LEARNING (ML) CLASSIFIER THAT CLASSIFIES TEXT STRINGS TO SINGLE CAUSES. THIS IS TRAINED USING TEXT FRAGMENTS EACH DESCRIBING A SINGLE CAUSE OF DEATH, AND THE CORRESPONDING SINGLE CLASSIFICATION. TO EXTRACT MULTIPLE CAUSES, EACH DEATH RECORD IS SPLIT INTO INDIVIDUAL WORDS, AND UNINFORMATIVE COMMON WORDS SUCH AS CONJUNCTIONS DISCARDED. EVERY POSSIBLE SUB-SEQUENCE OF CONSECUTIVE WORDS IS THEN EXTRACTED, FOR EXAMPLE: "BRONCHITIS", "WHOOPING", "COUGH", "BRONCHITIS WHOOPING", "WHOOPING COUGH", AND "BRONCHITIS WHOOPING COUGH".

EACH SUB-SEQUENCE OF WORDS IS CLASSIFIED, YIELDING A CLASSIFICATION AND CONFIDENCE VALUE FOR EACH ONE. IF A SUB-SEQUENCE IS ALREADY PRESENT IN THE TRAINING SET THEN THE CLASSIFICATION SPECIFIED THERE IS RETURNED WITH 100% CONFIDENCE. OTHERWISE, THE ML CLASSIFIER IS USED. SOME SUB-SEQUENCES ARE LIKELY TO RESULT IN SENSIBLE CLASSIFICATIONS, WITH HIGH CONFIDENCE, FOR EXAMPLE “BRONCHITIS” AND “WHOOPING COUGH”. OTHERS SUCH AS “BRONCHITIS WHOOPING” ARE LIKELY TO BE CLASSIFIED WITH LOWER CONFIDENCE, DUE TO THE LACK OF SIMILAR EXAMPLES IN THE TRAINING SET.

FROM THE SET OF CLASSIFIED SUB-SEQUENCES, ONE OR MORE ‘VALID’ SETS OF CLASSIFICATIONS IS CONSTRUCTED. FOR A SET OF CLASSIFICATIONS TO BE DEEMED VALID, IT MUST NOT CONTAIN MULTIPLE CLASSIFICATIONS LYING IN THE SAME BRANCH OF THE CLASSIFICATION HIERARCHY. FOR EXAMPLE, IN THE ICD-10 CLASSIFICATION THE CODES J20 “ACUTE BRONCHITIS” AND J206 “ACUTE BRONCHITIS DUE TO RHINOVIRUS” WOULD NOT BE ALLOWED TO OCCUR TOGETHER.

FURTHERMORE, THE SUB-SEQUENCES FROM WHICH THE CLASSIFICATIONS DERIVE MUST NOT OVERLAP. ONE OF THE VALID CLASSIFICATION SETS IS SELECTED, FAVOURING THOSE WITH MORE CLASSIFICATIONS AND HIGHER ASSOCIATED CONFIDENCE VALUES.

WE REPORT RESULTS OBTAINED USING HISTORICAL AND MODERN DATA SETS. WE DISCUSS HOW ACCURACY MEASURES VARY DEPENDING ON THE HIERARCHY LEVEL CONSIDERED, WHEN USING A HIERARCHICAL CLASSIFICATION SYSTEM, AND REVIEW SEVERAL APPROACHES TO ASSESSING ACCURACY WHEN CODING TO MULTIPLE CLASSIFICATIONS. FINALLY, WE PRESENT SOME RESULTS SHOWING HOW UNIQUE RECORDS ARE DISTRIBUTED WITHIN LARGE NATURAL LANGUAGE DATA SETS, AND SPECULATE ON HOW THESE COULD BE USED TO PREDICT FUTURE WORKLOAD IN LONG-RUNNING DIGITALISATION AND CLASSIFICATION PROCESSES.
DATA-DRIVEN METHODOLOGIES FOR DEVELOPING PROGNOSTIC MODELS IN CARDIOVASCULAR DISEASE

Authors: Aylin Cakiroglu1, Spiros Denaxas2, Harry Hemingway2, Raphaëlle Luisier1, Nicholas Luscombe1, Anoop Shah2, Andrew Steele1, Alessandra Vigilante1

1 The Francis Crick Institute, London, 2 Institute of Health Informatics, UCL London & Farr Institute, London

Presenting Author: Alessandra Vigilante

Corresponding author email: vigilantealessandra@gmail.com

The increasing use of electronic health records means that large quantities of information about a patient's medical history will are readily available for automated analysis and could be of potential benefit in guiding therapeutic decisions. Traditional epidemiological prognostic models use linear combinations of small subset of predefined variables to predict patient outcomes. Linked national electronic health records from primary and secondary care, such as the ones provided in the CALIBER platform, provide far richer information about patient phenotypes which data-driven methodologies can exploit in a manner that requires substantially less manual input. We will demonstrate results of applying of machine learning techniques for finding new clinically relevant variables, prediction of clinical outcomes, and exploring relationships between currently known and newly identified biomarkers.

DEMENTIAS PLATFORM UK: COHORT IDENTIFICATION AND ENGAGEMENT

Authors: K.s. Tingay1, K. McAllister2, S. Atkinson3, A. Hanbury4, J. Gallacher5, R.a. Lyons1

1 Swansea University, 2 University of Edinburgh, 3 Cambridge University, 4 Cardiff University, 5 Oxford University

Presenting Author: K.s. Tingay

Corresponding author email: k.s.tingay@swansea.ac.uk

Background: The Dementias Platform UK (DPUK) is an MRC-led, multi-million pound collaboration to facilitate research into dementias. A key part of this process aims to collate informatics from cohorts and to provide efficient means by which researchers can identify populations which might be of use, and could be approached for involvement in their research. By adding to information that we already know about the participants, including existing lifestyle, we hope to identify cognitive, genetic, physiological and imaging measures (biomarkers) to understand who is at risk of developing dementia and why the progression of dementia varies from person to person.

The DPUK is taking a new approach that considers what is happening to the health of the person – looking not just at the brain but at the brain in the context of the whole body – to identify those changes that may best be associated with the early stages of dementia. Emerging evidence has linked the development of dementia with inflammatory, cardiovascular and metabolic events and disorders, and the DPUK will be able to test these interactions, potentially identifying new treatment options to both manage symptoms, in the near term, and for preventative approaches in the longer term.

Methods: For the Informatics Platform work package, an integrated informatics environment will be established to bring together cohort and linkage data for the two million DPUK participants and help to co-ordinate the use of imaging and tissue sample resources for experimental medicine studies.
Aim of this presentation: This presentation will describe the challenges experienced so far in the Informatics Platform work package. Issues raised include engaging cohort owners, and the governance issues involved. The capturing of cohort metadata using an existing online tool, the EMIF Fingerprint Browser, will be discussed, particularly in relation to how the Browser has been adapted to fit the DPUK proforma, without compromising existing metadata standards.

Abstract 1651

ESTABLISHING A LINKED DATABASE & EVALUATING CHILDSMILE, THE NATIONAL ORAL HEALTH IMPROVEMENT PROGRAMME, WITHIN THE REMOTE NATIONAL SAFE HAVEN.

Authors: Jamie B.R Kidd, David I. Conway, Sherriff Andrea, Hodgins Faith, Lorna M.D. Macpherson, Alex D. McMahon

Community Oral Health, Dental School, College of MVLS, University of Glasgow, United Kingdom

Presenting Author: Jamie B.R Kidd

Corresponding author email: jamie.kidd@glasgow.ac.uk

Background: Childsmile, the national child oral health improvement programme for Scotland, established in 2006, delivers preventive interventions including: fluoride varnish application, toothbrushing and dietary advice, and toothbrushing programme in nurseries, schools and dental practices.

Aim: To assess the impact of the Childsmile programme on child oral health and oral health inequalities.

Design: The National Services Scotland (NSS) National Safe Haven (NSH) provides a secure portal allowing researchers to manage, link and analyse Electronic Patient Records (EPR). A linked database was assembled from multiple EPR datasets including: National Dental Inspection Programme (NDIP); Child Health Surveillance Health Visitor 6-8 week-old assessment (CHS); Management Information & Dental Accounting System; General/Acute and Inpatient Day Case (SMR01); Health Informatics Centre Childsmile database; and other historic Childsmile intervention databases.

An initial cohort, linked by weighted probability, consisted of children with both a CHS record from 2003-2012 and a 5-year-old NDIP result from 2009-2013. This was expanded to also include individual level data on: Childsmile interventions; dental registrations/treatment; hospital inpatient cases; as well as geographical and area-based Scottish Index of Multiple Deprivation (SIMD) data.

NDIP results were re-categorised and the endpoint was defined as either ‘obvious decay’ or ‘no obvious decay’. A second endpoint of dental extractions performed under general anaesthetic was established. Exposure variables included were: Childsmile dosage (single or combinations of interventions); age at first and last intervention; dental practice registration/attendance history; and SIMD data. Analysis was conducted using logistic regression, computing odds ratios and 95% confidence intervals using SAS Enterprise Guide version 6.1.

The study also provided an early opportunity for researchers to pilot the remote NSH and provide feedback to the electronic Data Research and Innovative Service (eDRIS).

Results: Preliminary analysis has been undertaken and final analysis will be completed upon receipt of the final tranche of data.

Problems relating to the installation of the remote NSH software and use of the SAS software contained within it were encountered. Working in partnership with eDRIS, resolutions were found and changes made to both the software and guidance notes to support future users of the remote NSH.

Conclusions: The assembly of a linked database accessed and analysed via the remote NSH, will provide an essential tool in the ongoing evaluation of Childsmile. Work to date has identified ongoing challenges to this process, which are being addressed.

Childsmile is funded by the Scottish Government Health Department and the NSH (in part) by NHS NSS Information Services Division.
Abstract 1701  
**DESIGN AND IMPLEMENTATION OF A COHORT EXPLORATION SYSTEM: THE HEALTH INFORMATICS CENTRE DATA AGGREGATION AND SUMMARISATION TOOL (HICDAST)**

**Authors:** Mike Ghattas  
*Health Informatics Centre (HIC) University of Dundee*

**Presenting Author:** Mike Ghattas

**Corresponding author email:** mghattas@dundee.ac.uk

**Background and Objective:** One of the important steps in conducting cohort studies is selecting the appropriate cohort based on specific research criteria. The main concerns at the beginning of any medical research is the availability of sufficient number of patients with enough data to complete the research at hand. Patients’ medical notes, questionnaires or Electronic medical records (EMRs) are usually used at this early stage of research to choose the project specific cohort. However, patient notes, EMRs or even questionnaires are not always available for all medical researchers, where many restrictions, mainly around patient’s confidentiality and privacy, have to be considered carefully. Additionally, other information would be required to identify the research cohort, which is not readily available to researchers and requires specific approved access to certain datasets. The objective of the Health Informatics Centre Data Aggregation and Summarisation Tool (HICDAST) is to provide researchers with easily accessible and fully anonymised longitudinal EMRs so as to facilitate cohort selection and data linkage.

**Method:** This talk will describe the design and implementation the HICDAST.

**Results:** HICDAST links several clinical datasets hosted and managed by the Health Informatics Centre (HIC) at the University of Dundee. HICDAST has been implemented and tested by many researchers. Preliminary feedback from researchers was very encouraging and HIC is currently piloting it in a secure safe haven environment.

**Conclusion:** HICDAST not only assures that data provided are fully anonymised with no identifiable data, but also enables researchers to enquire and explore a wide range of HIC-managed datasets with minimum efforts.

Abstract 1758  
**TUNNELING BETWEEN SILOS: ACCESSING DATA ACROSS DATA CENTRES**

**Authors:** Jim Mintha  
*Population Data BC*

**Presenting Author:** Jim Mintha

**Corresponding author email:** jim@popdata.bc.ca

As various data and linkage centres have formed and grown, it has become common to offer some sort of secure research environment where researchers can conduct their analysis within a protected environment. This allows some level of control over how the data is used and released. The environment and data is housed within a protected network, usually with very limited access to the outside Internet. Access is tightly controlled, often with sophisticated authentication, and any analysis that is to be exported can be subject to review.

What then happens when some of the data needed for a project is only available in a separate data centre? For a small set of data, permission can be arranged to transfer the data, but what if the data needed is part of a 10 petabyte genome database?

This paper will explore the various mechanisms that can be employed to connect disparate data centres in a secure manner that can allow the security, policies and procedures for all the data centres to be respected. Based on the experience connecting a remote data centre to Population Data BC’s
Secure Research Environment, I will outline the technical measures used to securely connect the centres, handle remote authentication to our servers, and allow data import/export. Beyond the technical side, it is just as important to have agreements in place to detail exactly who is responsible for what. I will look at these agreements as well as the various policies and procedures that are necessary to ensure smooth cooperation between the parties.

Abstract 1769

PREPARING NHS DATA ON PEOPLE WITH DIABETES FOR RESEARCH

Authors: Michael J Bluett, Scott G Cunningham
University of Dundee, Dundee, UK

Presenting Author: Michael J Bluett

Corresponding author email: mbluett@nhs.net

Aims: The Scottish Care Information – Diabetes Collaboration (SCI-DC) system collects data on over 275,000 people with diabetes. A system is required for researchers to be given extracts of the data to study the condition. Depending on the researcher, the extracted data may be identifiable or pseudonymised. Researchers may require linkage of the data to additional geographical look-ups, as well as cleaning and coding of the prescribing data the system holds. We aim to describe a series of processes to meet these objectives, within prescribed clinical governance frameworks.

Methods: A database instance called ‘Audit’, receives restores of data from the live database each weekend. Extracts for researchers requiring identifiable data access read-only views that return only results from their health boards (using row-level security). The views are dynamically regenerated after restores to reflect any changes made to the live database schema. Audit also stores patient identifiers for pseudonymised researcher cohorts. Extracts of either identifiable data based on the health board or pseudonymised cohort can be performed, with the data encrypted and transferred to the researcher. For researchers requiring additional processing, data can be further processed at the Farr Institute at the University of Dundee prior to re-encryption and transfer.

Results: Data has been extracted to the SDRN epidemiology group for use in a number of universities and separately for four research projects. Geographical data look-ups and cleaned and coded prescribing data have been provided to the SDRN epidemiology group. Extracts of identifiable Tayside and Fife data are being provided regularly to the Farr Institute at the University of Dundee for provision to researchers. Regular extracts have been partly automated, eliminating most of the labour involved in extracting, encrypting and transferring data.

Conclusions: A process has been defined to provide routinely collected clinical data to research teams for further analysis and epidemiological research. Further work is continuing to ensure all data useful to researchers is processed and extracted, as the live system stores, but doesn't process, data not deemed of clinical interest (including some data on the deceased).
Abstract 1450

SPATIAL METHODS FOR INFECTIOUS DISEASE OUTBREAK INVESTIGATIONS: SYSTEMATIC LITERATURE REVIEW

Authors: Catherine Smith¹, A C Hayward¹, S C Le Comber², H Fry³, M Bull⁴, S Leach⁴
¹UCL Department of Infectious Disease Informatics, Farr Institute of Health Informatics, ²School of Biological and Chemical Sciences, Queen Mary University of London, ³Centre for Advanced Spatial Analysis, University College London, ⁴Microbial Risk Assessment and Behavioural Science, Emergency Response Department, Public Health England

Presenting Author: Catherine Smith
Corresponding author email: catherine.smith.13@ucl.ac.uk

Background: Detecting and responding to outbreaks of infectious diseases is a key role of front-line public health organisations. Investigations are conventionally framed in terms of person, time and place, but spatial methods are used relatively infrequently. This study aimed to describe the extent to which these techniques have been implemented in previous investigations and the utility of the results that they provided.

Methods: We conducted a systematic review of published reports of outbreak investigations that used spatial methods. Reports were summarised according to the location, context and type of infection investigated. Methods used were categorised into broad classes of visualisation, description, cluster analysis and modelling; and the stage of the outbreak investigation to which they were applied was identified.

Results: We identified 80 reports using spatial methods; approximately 0.4% of the total number of published outbreak investigations. Water-borne (20 studies) or environmental (14) infections, including cholera and Legionnaires’ disease, were the most commonly investigated. The UK (10) and USA (8) had the most reports, with fewer in African countries (10 total). A range of spatial methods were used, with different tools applied at different stages of investigations. Cluster detection methods were used in establishing the existence of an outbreak, with the spatial scan statistic the most commonly applied (13 studies). Visualisation methods were used to describe cases, for example through simple dot maps (68 studies); or using maps of disease rates, either aggregated by region or using smoothed point locations. Geographic attributes of cases, such as their spatial average or proximity to potential sources, were used to generate hypotheses about the origins of outbreaks. These hypotheses were tested through modelling, for example by identifying areas that may have been exposed to contaminated air from suspected environmental sources; or using regression models to compare attack rates in areas at varying distances from potential sources. Finally, maps were used to communicate findings of investigations to the scientific community, policy makers and the public.

Conclusions: Spatial tools have been usefully applied to outbreak investigations across a broad range of contexts and infections. They have provided valuable insights that led to public health actions, but barriers to effective use include availability of accurate location data, lack of specificity of results, and provision of appropriate training. There is therefore scope for wider implementation of these methods, as well as development of new tools.
Abstract 1532  

THE EFFECT OF ULTRAVIOLET RADIATION ON BIRTH WEIGHTS AND GESTATIONAL LENGTH IN A SCOTTISH BIRTH COHORT

Authors:  
Lauren Megaw¹, Tom Clemens², Richard Weller², Chris Dibben², Sarah Stock²  
¹University of Western Australia, ²University of Edinburgh

Presenting Author: Lauren Megaw

Corresponding author email: lauren.megaw@health.wa.gov.au

Background: Seasonality effects have long been demonstrated on pregnancy outcomes including birth weight, preterm birth, stillbirth and pre-eclampsia. However, “season” is a surrogate marker for a complex collection of variables both social and geographical, and as such the outcomes of these studies around the world have varied. Ultraviolet radiation (UVR) exposure is one variable that may have an effect on pregnancy. Recent work has shown UVR exposure impacts blood pressure, obesity and immune function, as well as having a role in vitamin D metabolism. Epidemiological studies have shown UVR has positive associations with multiple sclerosis, overall mortality and cardiovascular disease. We hypothesise that UVR exposure in pregnancy will be associated with improved fetal growth accounting for some of the seasonal variation effects.

Aim: Our aim is to investigate the effects of UVR on birth weight, preterm birth, perinatal mortality and pre-eclampsia through linkage of data from the Scottish Birth Record (SBR) and SMR02 with historical sunshine measurements.

Methods: Obtain from the SBR and SMR02 the following outcomes on all singleton pregnancies delivered after 24 weeks from 2001 - 2013: partial post code, weight, gestational length, week of birth, maternal age, ethnicity, smoking status, parity, previous pregnancy outcomes, incidence of abruption, hypertension, pre-eclampsia, diabetes and social deprivation. Through the partial post code, link this to the sunshine hours data for that pregnancies specific monthly sunshine amount for each month of pregnancy calculated from date of delivery backward to conception. Calculate birth weight centiles and perform survival analysis with appropriate regression for levels of sunshine experienced at differing time points in pregnancy for term and pre-term births to assess for relative risks for small for gestational age (birth weight centile <10th). Adjust for known confounders including sex, birth order, smoking and socioeconomic deprivation using multivariate regression. Secondary analysis will be performed for preterm birth and hypertensive complications. The size of this cohort would allow a nested sibling analysis to help control for socioeconomic confounders. This work will be complemented by concurrent mechanistic studies investigating the positive direct effects of UVR exposure on uterine artery blood flow in pregnancy.

Conclusion: Seasonal effects on birth outcomes are complex, but isolating sunshine hours – a variable that is easily measured and historically recorded – and by linking this to a large Scottish birth cohort in both space and time we aim to investigate its association with pregnancy outcomes to guide future research and development of therapies.
Abstract 1705

A REAL-TIME DECISION SUPPORT SYSTEM FOR THE EARLY DETECTION OF GASTROINTESTINAL DISEASE OUTBREAKS IN SMALL ANIMAL PRACTICES IN THE UK.

Authors:
Fernando Sánchez-Vizcaíno1, B Rowlingson2, A.d. Radford1, A. Hale2, E. Giorgi3, S.J. O’Brien1, S Dawson4, R.m. Gaskell1, P.h. Jones1, T Menacere1, P.-J.M. Noble4, M Wardeh1, P Diggles2
1University of Liverpool, Institute of Infection and Global Health, Leahurst Campus, Chester High Road, Neston, UK, 2University of Lancaster, Faculty of Health and Medicine, Furness College, Lancaster, UK, 3University of Liverpool, Institute of Infection and Global Health, Waterhouse Building (Block F), 1-5 Brownlow Street, Liverpool, UK, 4University of Liverpool School of Veterinary Science, Leahurst Campus, Chester High Road, Neston, UK

Presenting Author: Fernando Sánchez-Vizcaíno

Corresponding author email: fsvb@liv.ac.uk

Introduction: The small companion animal (pets) sector currently lacks a coordinated disease surveillance system, creating a knowledge gap for endemic/emerging diseases in these species. The Small Animal Veterinary Surveillance Network (SAVSNET) captures large volumes of electronic health records (EHR) in real-time, including demographic, disease and syndrome information, from diagnostic laboratories and veterinary practices throughout the UK. Companion animals are commonly presented by their owners to veterinary surgeons in the UK with gastrointestinal disease (GI). Here we describe (1) descriptive statistics for GI in companion animals and (2) a novel real-time decision support system for early detection of changes in GI incidence in companion animals.

Material and Methods: EHR were collected from UK veterinary practices throughout 2014, including results from a short questionnaire administered to veterinary surgeons and appended after ~25% of GI consultations. A Bayesian spatio-temporal mixed effects binary regression model was used to model the incidence of GI in dogs and cats as a proportion of all presentations. The model was fitted to data between 01/11/2014 and 15/11/2014 using a bespoke Markov chain Monte Carlo algorithm to generate samples from the predictive distribution of the underlying spatio-temporal incidence surface. These samples were then used to compute predictive probabilities for exceedance policy-relevant relative risk thresholds; a high predictive probability at a particular time and place gives an early warning of a possible GI outbreak.

Results: EHR were obtained from 291,951 consultations (204,591 dogs, 73,604 cats) from 61 practices (120 premises). GI comprised 4.59% of canine and 3.63% of feline consultations, respectively. In total, 3,005 GI questionnaires (2,225 dogs, 665 cats) were collected. The most frequent GI syndrome reported in dogs was diarrhoea without blood (40.22% of GI consultations); in cats, the most common syndrome was vomiting without blood (36.69%).

The final model included as explanatory variables age, species, weekend indicator and a measure of deprivation.

To assess the ability of the model to detect outbreaks, we simulated a temporally localized outbreak of GI in one of the 120 premises. Predictive probabilities for a relative risk threshold of 1.2 or more were all comfortably greater than 0.5.

Conclusions: This is the first demonstration of the feasibility of real-time syndromic surveillance in UK small animal practices. In future work, we intend to adapt the model to early detection of human GI outbreaks, and to investigate dependence between the spatio-temporal variations in risk of GI for companion animals and people.
Abstract 1460

AN INTEGRATED GENOMICS AND CLINICAL RESOURCE FOR DATA-DRIVEN HEALTH SERVICES POLICY AND PRACTICE DECISION-MAKING

Authors: ML McBride, S Aparicio, JJ Spinelli, S Tyldesley

British Columbia Cancer Agency

Presenting Author: ML McBride

Corresponding author email: mmcbride@bccrc.ca

Background: Biomedical understanding and approaches to cancer are evolving rapidly. Scientists are decoding the genomes of different cancers and linking these to biological and medical/clinical data on disease development and progression.

In order to translate these findings into effective and efficient healthcare throughout the cancer care trajectory, risk stratification is needed to group cancer patients by levels of needs-based healthcare intensity and settings, based on individual health risks modified by current health conditions, genetic predispositions, demographics and lifestyle.

Objective: To link genomic and patient-reported outcome data to registry, medical/clinical, and population-based individual longitudinal healthcare administrative data on breast cancer patients in British Columbia, in order to generate evidence and a model of care to inform personalized care for these patients throughout the cancer care trajectory.

Methods: We have linked population-based, longitudinal, individual registry, clinical, and health administrative records for the approximately 32 thousand breast cancer patients diagnosed to BC residents from 1989-2008, with a population-based comparison group; and assessed cumulative and relative risks of long-term mortality, morbidity, healthcare utilization, access and quality of care, and predictors of these outcomes. We are currently updating the linkages to include patients diagnosed to end 2011, and followed to end 2013. For a subset of approximately 1000 patients diagnosed between 2005 and 2009, we have information on education, ethnicity, health/medical and reproductive history, family history of cancer, lifestyle characteristics, and lifetime occupational and residential histories; and patients are being genotyped using the 600,000 marker OncoChip. For another subgroup of approximately 720 patients, data is available on ten genetically-distinct molecular subgroups with different survival, including a high-risk, estrogen-receptor-positive 11q13/14 cis-acting subgroup and a favourable prognosis subgroup.

Results: Strengths of the resource are the detailed clinical and treatment information, available for 82% of patients; comprehensive health care utilization data (including all physician-ordered outpatient services, including physician visits; hospitalizations; mental health, longterm care, prescription drugs), available for 88% of the cohort; and length of follow-up. Analyses to date show multiple excess long-term health risks (premature mortality, second cancer, multimorbidity) and healthcare utilization relating to age at diagnosis, cancer stage, and treatment.

Conclusions: Record linkage of cancer registries, genomic and clinical information, and administrative databases, is a cost-effective and comprehensive means to conduct cancer care research. Findings will add to knowledge of the implications of genomics on care. They will also inform development of care guidelines, management strategies and health policy, to support targeted risk-based surveillance and interventions.
Abstract 1567

DEVELOPING UK DEMENTIAS PLATFORM BASED ON UKSeRP

Authors: Simon Thompson
Swansea University

Presenting Author: Simon Thompson

Corresponding author email: simon@chi.swan.ac.uk

The MRC Dementias Platform UK (DPUK) is a multi-million pound public-private partnership, developed and led by the Medical Research Council, to accelerate progress in, and open up, dementias research. The DPUK’s aims are early detection, improved treatment and ultimately, prevention, of dementias.

The dementias are a progressive, debilitating and presently incurable set of conditions characterised by neurodegeneration with a severe decline in cognitive functioning, including memory and communication. Over 600,000 people in the UK suffer from dementia, at a socioeconomic cost of over £17 billion a year.

The analysis portal contributes towards the first and second DPUK strategic objectives which are the integration of data across cohorts and create a single portal for data access. This involves the collection, integration, curation and analysis of the substantial volumes data obtained during the course of project.

The project will bring together 22 dementia cohorts which brings some of the big data challenges due to the dataset size but especially the complexity of the data such as baseline surveys, clinical measures, imaging, omics, sensors, routine data: structured and unstructured. Each cohort has collected and stored a large number of different variables and where there are common variables the coding or units of measurements will differ. Data harmonisation will be a major challenge with the project to ensure maximum utility can be gained from a holistic view.

Through the use of UKSeRP (FARR UK Secure Research Platform) and multiple deployments of NRDA (FARR National Research Data Appliance) the portal will be established and all data management, loading and cataloguing will be provided through these initiatives. The NRDA will also allow for devolved control to the cohorts for dataset access and sharing of data between cohorts or research projects.

The governance issues and concerns within the portal are very complex due to the fact that each cohort will already have governance principles and agreements in place, therefore by combining several cohorts for a specific research question what governance model allows the cohorts to stay compliant.

Abstract 1642

WILL THE REAL JOE BLOGGS PLEASE LINK UP? ASSESSING THE QUALITY OF INDEXING IN RECORD LINKAGE.

Authors: S Iqbal, D N Clark, C Morris
NHS National Services Scotland

Presenting Author: S Iqbal

Corresponding author email: suhail.iqbal1@nhs.net

Background: Research that uses linked datasets relies on the linkage process matching up the correct records for an individual from various data sources e.g. Joe Bloggs who is taking part in a randomised control trial is linked to the same Joe Bloggs in NHS Scotland health datasets. The process of linking to a population spine, such as the NHS Scotland Community Health Index, is called indexing. Robust indexing is key to providing high quality data to researchers who wish to make use of linked datasets for their research. This is also important for our NHS Scotland internally linked datasets. However, indexing is
not an exact science therefore we need ways of measuring the quality of the process.

**Impediments to the indexing process:** The most accurate way to link data would be if all individuals had one universal unique identifier that never changed, was never reused, was impervious to corruption or input errors and was recorded across all datasets, meaning that this fantastical identifier would be sufficient to link Joe Bloggs’ records across any datasets the researcher was interested in. However, because in practice no such identifier exists, we rely on a number of identifiers such as name, date of birth, gender and postcode to match individuals. In most instances this allows for individuals to be matched up correctly in a self evident way as the identifiers provided to us match up with our records for a given individual. Problems creep in when, e.g., missing or erroneous identifiers such as names are supplied to us to perform indexing, or people are not on the population spine.

**Assessing the quality of indexing:** In the minority of cases where an exact match is not found, we rely on probabilistic algorithms to assist us in identifying the best match. But how do we assess the completeness and accuracy of these ‘best matches’? We have proposed methods of performing quality assurance on our indexing process to allow us to measure the precision (total number of correct pairs matched out of the total numbers of pairs matched) and recall (total number of correct pairs matched out of the total number of correct pairs). This quality assurance process can be used to assess linkage quality across different indexing population spines, systems and human processors, and in turn lead to performance improvements.

Abstract 1684  
**IMPROVING TRANSPARENCY AND CONSISTENCY OF DATA ACCESS REVIEWS IN BC**

Authors:  
Nancy Meagher¹, Kaitlyn Gutteridge¹, Kim McGrail²

¹Population Data BC, ²University of British Columbia

Presenting Author: Kim McGrail

Corresponding author email: nancy.meagher@popdata.bc.ca

**Objective:** Requests for secondary uses of administrative data in British Columbia, as in many jurisdictions, typically require three layers of review: 1) peer review, 2) Research Ethics Board review, and 3) data steward review. Each of these agents or review bodies acts independently, though sometimes with reference to other reviewers. There is some uncertainty about the nature and extent of overlap in the lens and ambit for review. This project sought to identify the elements being assessed during each of the three reviews involved in a data access request.

**Approach:** We assessed relevant legislation, policy and procedure documents, and guidelines associated with each reviewing agent. Findings were corroborated with the experience of staff of Population Data BC who have an oversight and coordination function for research data access requests in BC. Thirteen core elements were identified and grouped into six categories, which are being framed as principles of review:

- Science: scientific merit and potential impact
- Approach: questions and analyses
- Data: granularity, sensitivity and justification for the data requested
- People: experience, affiliation
- Environment: technical and network infrastructure
- Interest: public and societal interest

The scope of review for each of these elements by each of the three agents was identified based on current practice and mandate. The extent or priority of review was recorded in a grid and heat-mapped to make it easy to see areas of overlap among the agents.

**Results:** There is significant overlap in reviews on each of the core principles, pointing to the possibility of improving consistency and transparency in reviews with better coordination and transparent roles. This structured investigation has supported development of a harmonized set of review principles, and
opens up the initial dialogue amongst review bodies as to who is responsible for which principle, which means there is some potential for improvement in efficiency as well. Further discussions will have to address how each principle translates into an adjudication decision.

**Conclusion**: Reviews by each of the three agents or bodies engaged in data access adjudication in British Columbia have significant overlap. Clarification of the principles being reviewed is a first step in reducing the overlap and through that improving transparency, consistency and potentially efficiency.

---

**Abstract 1691**

**PULLING TOGETHER, NOT APART: WORKING PARTNERSHIPS ARE KEY TO GOOD GOVERNANCE AND HIGH QUALITY RESEARCH IN SAFE HAVENS**

**Authors**: Corri Black, S Gordon, A Coutts, S Hall, P Rebecca, A Soutar, A Woznik, K Wilde

1Farr Institute @ Scotland. 2University of Aberdeen

**Presenting Author**: Corri Black

**Corresponding author email**: corri.black@abdn.ac.uk

**Background**: In recent years, significant investment in health informatics research has enabled substantial expansion of the field. At one NHS Scotland and academic regional research facility, the Grampian Data Safe Haven [DaSH], we have seen rapid development with: new data linkages, new datasets, increased project complexity and a sizeable growth in the number of researchers new to the field. The governance and permission pathways continue to evolve. The challenge has been to facilitate best practices in governance while supporting increasingly diverse research needs. The importance of trusted partnership working has been recognised as crucial.

**Aim**: To identify opportunities to evolve from a service provision model to one embedding working partnerships with researchers to promote best practice in governance and research methods.

**Methods**: In 2015, Grampian DaSH undertook a review of the opportunities for partnership working with researchers across the life of a research project in the safe haven. Pivotal process points were mapped. A range of approaches to support good partnership working were defined. Using lessons learnt from the first 100 DaSH research contacts, we are developing new ways to embed best practice.

**Results**: Six research milestones were identified: Project Initiation; Planning; Pre-Linkage; Linkage; Analysis and Archive. Approaches to supporting partnership working ranged from a single point of contact (DaSH email), a named lead research coordinator and analyst to oversee the project journey, a data checking partnership, core partners team meetings and a secure research ‘laboratory’, with immersive videoconferencing, to enable team working. Two key developments have emerged. Firstly is the move from a static, phased project management framework to an ‘agile’ project management approach; seeing projects as a continuous journey and using ‘scrum and sprint’ approaches to identify and implement actions. Secondly is the recognition of the importance of the researchers, research coordinators and data analysts working as a team to ensure good governance, improve data quality and understanding, and share learning in a changing research environment.

**Conclusion**: Grampian DaSH aims to work in partnership with researchers to build knowledge and understanding of health, care and society and share learning. By working as a team, researchers, research co-ordinators and analysts can embed best practices in project planning and communication across the life of project.
DATA SAFE HAVENS IN HEALTH RESEARCH AND HEALTHCARE

Authors: Paul R Burton¹, Madeleine J Murtagh¹, Andy Boyd², Edward S Dove³, Bartha M Knoppers⁴

¹Data 2 Knowledge, University of Bristol, ²ALSPAC, University of Bristol, ³JK Mason Institute for Medicine, University of Edinburgh, ⁴Centre of Genomics and Policy, McGill University

Presenting Author: Andy Boyd

Corresponding author email: p.burton@bristol.ac.uk

The health research community are engaged in projects which require a wealth of information. This information can be drawn directly from the population, from routinely collected records and neighbourhood and environmental sources. Frequently, investigators require information from multiple sources with multiple legal owners. A fundamental challenge for data managers – such as those maintaining cohort study databanks - is to site such data in repositories that can easily be accessed under appropriate technical and governance controls which are effectively audited. This demands socio-technical solutions that may easily become enmeshed in protracted debate and controversy as they encounter the norms, values, expectations and concerns of diverse stakeholders. In this context, the development of what are called ‘Data Safe Havens’ has been crucial. Unfortunately, the origins and evolution of the term have led to a range of different definitions being assumed by different groups. There is, however, an intuitively meaningful interpretation that is often assumed by those who have not previously encountered the term: a repository in which useful but potentially sensitive data may be kept securely under governance and informatics systems that are fit-for-purpose and appropriately tailored to the nature of the data being maintained, and may be accessed and utilised by legitimate users undertaking work and research contributing to biomedicine, health and/or to ongoing development of healthcare systems.

In this paper we explore a fundamental question: “what are the specific criteria that ought reasonably to be met by a data repository if it is to be seen as consistent with this interpretation and viewed as worthy of being accorded the status of ‘Data Safe Haven’ by key stakeholders”? We propose 12 such criteria relating to the themes of: data maintenance and release; veritable data; and, data security. We will discuss the criteria in turn.

THE YIN & YANG OF DATA MANAGEMENT - FOUR INTRACTABLE CONTRADICTIONS AND THEIR HARMONIOUS RESOLUTIONS

Authors: Alex Hacker

Nuffield Department of Population Health, University of Oxford

Presenting Author: Alex Hacker

Corresponding author email: alex.hacker@ndph.ox.ac.uk

Background: The China Kadoorie Biobank study (www.ckbiobank.org) provides a uniquely rich and powerful resource for investigating environmental and genetic determinants of chronic disease. It is a long-term prospective cohort study of ~513,000 men and women aged 30-79 years recruited from ten geographically diverse urban and rural regions in China. Participants are actively followed-up through established morbidity and mortality registries, and by linkage to the national health insurance system to identify disease events.

Importing, integrating and distributing such a large and diverse dataset presents practical challenges, but also more fundamental ones. When individually reasonable data management requirements come into conflict, can a solution be found that satisfies both without compromising either? In this session I present, with examples, four such conflicts and the ways in which we harmonise and resolve
Methods:

Gather perfect data / Handle imperfect data
A data manager does not want to allow erroneous data into their database, but nor do they want to discard data that are imperfect but still meaningful. The approach that CKB takes is to implement automatic validation at the point of entry, tailored to the data source in question.

Fix data issues / Don’t make assumptions
Ideally every data issue that is detected should be fixed or flagged, but correction is not always possible and it’s rarely clear where to draw the line between error and outlier. We address this via comprehensive data documentation, empowering each analyst to identify, assess and handle the values that might be problematic for them.

Be flexible / Be consistent
Every researcher has different data requirements, definitions and exclusions. Data management must support this without needless duplication of effort, or leaving everyone working on incompatible datasets. CKB’s solution is centralised distribution: a single core database, centrally maintained and updated, from which all analyst’s datasets are derived, and into which the work of individuals is incorporated for the benefit of all.

Keep it simple / Include everything
Most analyses examine some areas of the data in great detail but require only basic summaries of others. CKB supports this using well-chosen data aggregation, offering multiple levels of detail so that each analyst can decide how far they wish to “zoom in” on each element.

Conclusion: Some apparently contradictory requirements of data management can be resolved with the above techniques, creating a resource suitable for a wide range of applications without compromise.

Abstract 1754  FUNDING MODELS FOR RECORD LINKAGE CENTRES AND THEIR IMPLICATIONS: AN INTERNATIONAL COMPARATIVE ANALYSIS

Authors: John Frank1, Claudia Pagliari1, Andrew James Williams1, Felicity Flack2, Rebecca Glauert2, Nick de Klerk2, Fiona Stanley2, Noralou Roos3

1Farr Institute @ Scotland, University of Edinburgh, 2Telethon Kids Institute/University of Western Australia, 3Manitoba Centre for Health Policy, University of Manitoba

Presenting Author: John Frank

Corresponding author email: john.frank@ed.ac.uk

Background: Across Australia, Canada, and the United Kingdom a number of models have evolved in recent decades for funding “centres” (either single centres or networks of centres) for the linkage and analysis of largely public-sector administrative datasets. The various funding models have major implications for centre organizational structure and governance. The datasets linked in these facilities are typically anonymized in the process of linkage, to create comprehensive longitudinal datasets comprising two or more administrative or research data resources.

Purpose and Methods: This paper uses interview information from key respondents, across the three countries, knowledgeable about the various funding models in use, and their pros and cons. Comparative analysis of the current funding models is used to derive guidance for those embarking on such joint ventures, or considering organizational change in existing centres/networks.

Summary of Findings: Broadly speaking, most existing centres are funded, and therefore substantially directed by the mandates arising from that funding, by either:

- mostly research grants, typically awarded by national research funding agencies to consortia
of university-based academics, to create and analyse ongoing “linked dataset platforms” for learning about entire populations’ life-course trajectories, and their determinants – as well as for evaluating policies, programmes and practices intended to change these trajectories, in regards to the wide variety of outcomes available in the administrative data.

- mostly long-term, renewable contracts with the public sector, typically intended to maintain and update core datasets and linkage capacity (including the support of sound career pathways for the highly skilled staff employed in such centres); these contracts often include specific, pre-negotiated “deliverables” which the centre must produce during the funding period, for decision-makers in government agencies and ministries who have contributed the datasets in the first place.

Many centres have mixed models that allow both sorts of funding simultaneously, which itself has some advantages, especially when one form of funding falters. Our respondents reported a complex mixture of advantages and disadvantages for each of these models, related in part to the adage “who pays the piper calls the tune,” as well as the differing “cultures of data use and research” across academia and public sector administration. They also reported that sometimes one model is relied on in the early years of a centre’s existence, but later events force a diversification of funding models. Our presentation will summarize the pros and cons of the various models, according to the experience of key informants.

**PATIENT AND PUBLIC INVOLVEMENT**

**Abstract 1377**

**ESTABLISHMENT OF A PUBLIC PANEL FOR THE SCOTTISH HEALTH AND ETHNICITY LINKAGE STUDY (SHELS)**

**Authors:**

Hester Ward¹, Anne Douglas², Arma Sayed-Rafiq³, Alex Stannard⁴, Graham Bissell⁵, Genevieve Cezard², Raj Bhopal², Laurence Gruer²

¹NHS National Services Scotland, ²University of Edinburgh, ³NHS Health Scotland, ⁴National Records Scotland

**Presenting Author:** Hester Ward

**Corresponding author email:** hester.ward@nhs.net

**Introduction:** The Scottish Health and Ethnicity Linkage Study (SHELS) is an on-going collaboration examining ethnic differences in hospitalisation and mortality for various diseases. Publicly funded research should reflect public interests and values; specifically linkage of health records to the self-reported ethnicity from the census merits direct public engagement. To date, SHELS methods and results have been communicated through publications, community events, a website and mass media releases. After much consideration of approaches, we sought to establish a panel enabling dialogue between the researchers and people representing multi-ethnic Scotland.

**Aim:** To create a panel of 10-20 members of the public from different ethnic groups to meet with the SHELS team to:

- Learn about the SHELS study and comment on methods and findings;
- Contribute to dissemination methods and future research plans.

**Methods**

**Expectations:** Panel members were expected to attend three half-day meetings in Edinburgh or Glasgow over 18 months; read papers for each meeting; work with other panel members; and contribute to group discussions.

**Payment:** Up to £35 travel expenses and a £40 ‘High Street Stores’ voucher per meeting.

**Recruitment:** Applicants aged 18 and over were sought via NHS and community groups representing minority ethnic populations and other research networks; and via SHELS, Edinburgh University and
Volunteering websites. A balance of ethnic groups was attempted and persons not involved in research were given preference. Successful applicants completed and signed a Terms & Agreement form.

**Results:** There were 29 applicants: 19 accepted a place on the panel, with four on a reserve list. Ethnicity of the 19 included: Indian, Indian-White Scottish, Pakistani, Chinese, Persian, Portuguese, American, Irish, and White Scottish; 11 were female, ages ranged from 29–69.

The first SHELS Public Panel took place in March 2015 with 12 panel members and nine SHELS team members.

**First meeting:** This covered: introduction of all participants; background, purpose and aims of SHELS; creating the linked databases; data security and confidentiality; methods of analysis; and presentation of a sample of results. The meeting was successful, with active engagement of the panel members. Many pertinent questions were asked and discussed, e.g. how complete and reliable was ethnic recording; were illegal immigrants included; what differences in health were found; and had the results influenced policy or improved health?

**Conclusion:** A multi-ethnic panel has been established, with the first meeting indicating considerable potential for exploring public attitudes to this type of research and so meeting its aims.

---

**Abstract 1679**

DEVELOPING A PROGRAMME OF PUBLIC ENGAGEMENT FOR A REGIONAL RESEARCH SAFE HAVEN

**Authors:** Fiona E Strachan1, Mhairi Aitken1, Sophie Ilson1, Sarah H Wild1, Alison McCallum2, Janet Hanley2, Sarah Cunningham-Burley1

1University of Edinburgh, 2NHS Lothian

**Presenting Author:** Fiona E Strachan

**Corresponding author email:** f.strachan@ed.ac.uk

The importance of public engagement in research is well recognized and forms a key element in developing our approach to record linkage for research purposes. As part of this work, we contributed to a workshop with the Farr Institute @ Scotland public panel to discuss plans for our local research safe haven. This took the form of a short presentation with questions for discussion, panel members also had the opportunity to ask questions and raise issues of concern or interest.

The presentation covered the policy context and national guidance, principles of using routine data for healthcare research and plans for the Lothian Research Safe Haven. The specific topics discussed were:

- the use of routine healthcare data for healthcare research
- privacy and security standards
- access to data
- linking to other datasets e.g. social care
- commercial companies as partners in research.

The presentation stimulated considerable discussion amongst the panel members and questions were taken throughout the presentation rather than tabling at the end. The discussion highlighted a number of areas of interest, including consent and if consent was ‘opt out’, how people would know about this process; identifiers and whether data would be anonymous was important; data security was highlighted not just in relation to IT hardware but also to staff codes of conduct and supervision; access to the data and who would approve researchers and specific projects; the involvement of commercial partners in research raised a number of concerns, in particular the concern that data would be “sold to commercial companies”. There was some interesting discussion around the term “routine data” and sensitivity of personal data. There was keen interest in ongoing public engagement, perhaps in the form
of project review as part of the work of the safe haven, and not just to inform this early stage in development.

The workshop provided a valuable opportunity to inform the panel of plans for the development of the Lothian Research Safe Haven and to discuss concerns raised. This work will be followed up with a programme of public engagement. This could include similar workshops and discussions with specific patient groups or representative organisations. The role of lay members in project review will also be considered.

The development of meaningful ways of including the perspective of patients and members of the public in local and national record linkage projects is vital to ensure public confidence in healthcare research linking routinely collected clinical data.

LINKED OBSERVATION FOR HEALTH POLICY

**Abstract**

**ANALYSIS OF BIRTHS AT HOME IN ENGLAND AND WALES BY TIME OF DAY AND DAY OF WEEK USING LINKED ADMINISTRATIVE DATA**

**Authors:** Nirupa Dattani, Alison J Macfarlane  
*City University London*

**Presenting Author:** Alison J Macfarlane

**Corresponding author email:** n.dattani.1@city.ac.uk

**Background:** Clinical policies from the 1970s onwards led to rising levels of intervention and resulted in a concentration of births on to weekdays. Data from Scotland and the Netherlands suggests that births 'in office hours' may be safer for mothers and babies. From the 1990s onwards, the proportion of births at home in England and Wales has risen after falling below 1 per cent in the late 1980s. This has led to questions about the timing of spontaneous birth and its relation to planning of services.

**Method:** This analysis was undertaken as part of a larger series of data linkage studies in which data from civil registration in England and Wales were linked to data recorded when an NHS Number, a unique identifier, is allocated to new born babies. These data have been routinely linked and are available from 2005. Day of the week was derived from date of birth of the baby and home births were identified using place of birth recorded at civil registration. Time of birth was taken from the NHS Numbers for Babies dataset. Data about births from 2005 to 2012 were analysed in the secure facilities Virtual Micro-Data Laboratory (VML) at ONS. Further linkage to care at delivery will enable analysis of births at home by complications in the mother and the baby.

**Results:** The majority of births at home occur at term and numbers vary by time of day and day of the week. This pattern has remained consistent throughout the period 2005 to 2012.

**Discussion:** Further analysis of the linked data will cover all births, the majority of which occur in hospital, by onset of labour and method of delivery by time of day and day of week to see if there are differences in outcome. The aim of our analysis is to inform the NHS when planning midwifery and obstetrics staffing levels by time of day and day of the week.
USING DATA LINKAGE TO BUILD A NATIONAL DATABASE OF DATA ABOUT BIRTH AND ITS OUTCOME FOR POLICY ANALYSIS IN ENGLAND AND WALES

Authors: Alison J Macfarlane, Nirupa Dattani
City University London

Presenting Author: Alison J Macfarlane

Corresponding author email: A.J.Macfarlane@city.ac.uk

Background: In England and Wales, the data recorded at civil birth registration are mainly socio-demographic, while systems used to record hospital care in general and care at delivery in particular, contain very few demographic data items. Another dataset, NHS Numbers for Babies (NN4B), recorded when a newborn baby’s NHS number, a unique identifier, is allocated contains key items not present in the birth registration or hospital systems. A series of collaborative projects was undertaken to build a linked database. The first two aimed to pilot the linkages and explore their potential for use in research and the production of national statistics, while two further projects were funded to extend the linkage to answer specific questions.

Method: The first project piloted linkage of NN4B to birth registration data for births in 2005, using the NHS number, date of birth, postcode and sex. The second project, using linked data for births in 2005-07, linked in maternity records from the Hospital Episode Statistics (HES) for England and the National Community Child Health Database / Patient Episode Database Wales, using the mother’s NHS Number, where available, or several combinations of other indirect patient identifiers such as mother’s date of birth, postcode and baby’s date of birth. The next project, funded to answer questions about the outcome of pregnancy by time of day and day of the week, will extend the database to include births from 2005 to 2012 and link in data about subsequent admissions of mothers and babies to hospital. A further project has been funded to link these data to educational records of children born in 2005-06.

Results: After 99.8 per cent of birth registration records of children born in 2005, were linked to NHS numbers for Babies records, this linkage was mainstreamed from 2006 onwards. Overall 91 per cent of delivery records in England and Wales were linked to the birth registration/NHS Numbers for Babies records.

Discussion: To benefit fully from this linkage, improvements are urgently needed in the quality of the data contained in Maternity HES. Results are not yet available for the current project which has been severely delayed by problems at the Health and Social Care Information Centre. Building this linked database is improving the scope and range of national maternity data for England and Wales, but there have been many challenges to overcome, including changes to data access, organisational change and intermittent availability of funding.

USE, EFFECTIVENESS AND SAFETY OF ORAL ANTIPLATELET MEDICINES DISPENSED IN THE COMMUNITY IN SCOTLAND: COHORT DEFINITION AND PRELIMINARY FINDINGS

Authors: Grant M A Wyper1, Dr. Samantha Alvarez-Madrazo1, Dr. Kim Kavanagh1, Dr. Martin Denvir2, Prof. Marion Bennie1
1University of Strathclyde, 2University of Edinburgh

Presenting Author: Grant M A Wyper

Corresponding author email: grant.wyper@strath.ac.uk

Introduction: Antiplatelet drugs are a key secondary-prevention therapy in cardiovascular disease and are recommended in many national and international guidelines. Until 2009 the antiplatelet agents available were aspirin, dipyridamole and clopidogrel. Recently two new antiplatelet agents, prasugrel
and ticagrelor, were approved in Scotland. These new agents are considerably more expensive and carry a higher risk of adverse bleeding events. Data sources in Scotland provide a unique opportunity to assess the impact of these new agents in routine clinical practice.

**Methods:** This study will investigate: (i) Antiplatelet utilisation in the community in accordance to Health Board of residence, clinical indication, demographics, urban/rural classification and socioeconomic status; (ii) Effectiveness of individual antiplatelets in relation to safety outcomes including haemorrhagic events, subsequent coronary or cerebrovascular events including interventional procedures; (iii) The risk of cancer associated with the use of antiplatelets.

The patient cohort will be identified as patients being dispensed antiplatelets in the community, or with a hospital discharge diagnosis for one or more of the known antiplatelet indications (e.g. stroke, acute coronary syndrome, peripheral artery disease) during 2009-2014. Follow-up after the index event will be undertaken using acute hospital medical diagnoses and procedures identified from the Scottish Morbidity Record 01 for a period of 12 months after the index event. Cancer registrations will also form part of follow-up using data from the Scottish Cancer Registry. To evaluate drug-drug interactions, the full prescribing history will be obtained from the Prescribing Information System. Patient death records will be sourced from National Records of Scotland and migration dates from the GP registration system will be used to censor patients.

**Results:** Preliminary findings showed that 465,044 patients were issued antiplatelets from primary care in 2014. The number of patients by individual therapy was: 384,236 (82.6%) aspirin; 6,489 (1.4%) dipyridamole; 109,940 (23.6%) clopidogrel; 427 (0.1%) prasugrel and 5,551 (1.2%) ticagrelor. There was a two-fold increase in the number of patient’s dispensed ticagrelor, whereas there was only a slight increase in the number of patient’s issued prasugrel from 2013 to 2014. Both mono and dual antiplatelet therapy will be explored during this study including any changes in therapy, particularly following cardiovascular and cerebrovascular events.

**Conclusions:** This retrospective observational study focusing on the utilisation of antiplatelets in the Scottish community via the linkage of patient-based health records will provide important insights into the clinical impact of novel antiplatelet agents.

---

**Abstract 1599**

**THE USE OF ROUTINE HEALTH DATA AND RECORD LINKAGE IN A NATIONAL BURDEN OF DISEASE STUDY FOR SCOTLAND**

**Authors:**

1 I Grant, 2 D Stockton, 3 M McFadden, 4 G Wyper, 5 O Mesalles Naranjo

1Information Services Division, National Services Scotland, 2NHS Health Scotland

**Presenting Author:**

Grant M A Wyper

**Corresponding author email:**

ian.grant@nhs.net

**Background:** In a resource constrained environment there is increasing pressure for strategic planning and priority setting for public health and health services to be informed by robust evidence. Burden of Disease studies can provide a comprehensive assessment of the health status of the population, through record linkage and use of routine health data, one that integrates fatal and non-fatal outcomes, yet also allows them to be examined separately.

The National Burden of Disease, Injuries and Risk Factors Study in Scotland is a two year project funded by the Chief Scientist Office. The study is being conducted by the Scottish Public Health Observatory (ScotPHO)

**Methods:** Years of life lost (YLL) measures the years of life lost due to premature death. Mortality and Life Expectancy data from National Records Scotland have been used to calculate YLL for over 200 diseases. Years Lived with Disability (YLD) is a measure of health loss resulting from non-fatal health outcomes from diseases and injuries. YLD is calculated from the estimation of incidence or prevalence of the health condition. The study design is to link two or more datasets (dependent on disease) to derive disease prevalence and/or incidence estimates by linking primary and secondary
care datasets. YLD, together with YLL, are used to calculate the Disability Adjusted Life Years (DALY).

**Results**
- Years of life lost to premature mortality for over 200 diseases in Scotland 1981-2013, by age, gender and local geographies.
- Annual prevalence and DALYs, (numbers and rates) in Scotland by gender and age group for each year between 2000 and 2013 for a selection of diseases including: Ischaemic heart disease, Liver cirrhosis, Stroke, Chronic Obstructive Pulmonary Disorder, Dementia, Diabetes, Uni-polar depressive disorders, Gall bladder and bile duct disease, Pancreatitis, Alcohol use disorders

**Conclusions:** This study will provide comprehensive evidence to support rational resource allocation and to identify inequalities in the disease burden across patient demographics and socio-economic groups. This information can be used to compare the relative impacts of interventions in reducing the burden of disease and to inform economic evaluation of those interventions, thus helping to address the future challenges posed by the ageing of the population, changes in disease and risk factor patterns, and the increasing costs of health services through the production of projections of the disease burden.

**Abstract 1619**

**TO WHAT EXTENT IS THE BURDEN OF DISEASE IN SCOTLAND ATTRIBUTABLE TO A RANGE OF MODIFIABLE RISK FACTORS?**

**Authors:**
E Tod, G McCartney
Scottish Public Health Observatory

**Presenting Author:**
Grant M A Wyper

**Corresponding author email:**
elaine.tod@nhs.net

**Background:**
In a period of limited resources and competing health priorities, the availability of robust health data to inform effective public health decision making is becoming increasingly important. Burden of Disease studies are useful in this regard as they provide a comprehensive assessment of the health of the population through the quantification of fatal and non-fatal outcomes resulting from disease. A key element of this is the comparative risk factor assessment which seeks to understand the contribution of modifiable risk factors in explaining the disease burden in a population.

The National Burden of Disease, Injuries and Risk Factors Study in Scotland is a two year project funded by the Chief Scientist Office and is being conducted by the Scottish Public Health Observatory (ScotPHO). The linkage of Scottish Health Survey data with routine health outcome data will provide the basis for the comparative risk factor assessment.

**Aim and objective:**
To provide comprehensive data on health needs to support rational resource allocation.

- Estimation of the proportional reduction in disease that would occur if exposure to a range of causally linked modifiable risk factors were reduced to an optimum exposure distribution.

**Methods:**
This is a prospective cohort study design, enabled by linkage of the Scottish Health Survey (for provision of risk factor data) to routine administrative data on health outcomes for selected disease areas. Data will be analysed from eight waves of the Scottish Health Survey between 1995 and 2012. The framework for the risk factors is based on the Dahlgren and Whitehead model of the social determinants of health and evidence was obtained from literature reviews on the specific risk factors for each disease outcome. A mix of self-reported and objective measurements is included in this study.

The key outcome measure is the population attributable fraction. This can be defined as the proportional reduction in population disease that would occur if exposure to a risk factor were reduced to an alternative ideal exposure scenario. Cox proportional cause-specific hazards regression will provide hazard ratio estimates as an input to the population attributable fraction formulae.

**Results:**
Depending on data availability, population attributable fractions will be presented for a
Abstract 1602

ESTABLISHING A COHORT OF PATIENTS WITH CHRONIC INFLAMMATORY CONDITIONS TREATED WITH BIOLOGICS IN GREATER GLASGOW & CLYDE USING A NEW HOMECARE SERVICE DATABASE

Authors: Samantha Alvarez-Madrazo1, Yvonne Semple2, Kim Kavanagh3, Kevin Milne4, Gayle Robertson4, Norman Lannigan4, Daniel R. Gaya5, Stefan Siebert6, Duncan Porter7, Marion Bennie8
1Strathclyde Institute for Pharmacy and Biomedical Sciences, University of Strathclyde, Glasgow, 2Strathclyde Institute for Pharmacy and Biomedical Sciences, University of Strathclyde, Glasgow; Medicines Information, Pharmacy Department, Glasgow Royal Infirmary, 3Department of Mathematics and Statistics, University of Strathclyde, Glasgow, 4Pharmacy and Prescribing Support Unit, Glasgow, 5Gastroenterology Unit, Glasgow Royal Infirmary, 6Institute of Infection Immunity and Inflammation, University of Glasgow, 7Gartnavel General Hospital, Glasgow, 8Strathclyde Institute for Pharmacy and Biomedical Sciences, University of Strathclyde, Glasgow; Public Health and Intelligence Strategic Business Unit, NHS National Services Scotland, Edinburgh

Presenting Author: Samantha Alvarez-Madrazo

Corresponding author email: samantha.alvarez-madrazo@strath.ac.uk

Background: Treatment of chronic inflammatory conditions such as rheumatoid arthritis, inflammatory bowel diseases and dermatological disorders is rapidly evolving. Biologics are the newest class of drugs, recommended by the UK and international guidelines, for those unresponsive, intolerant or with contraindications to conventional therapies. Biologics are high cost medicines, often supplied directly to patients by external healthcare providers. Several biologics are licensed for inflammatory conditions and there can be patient specific variability in response and adverse effects. Thus there is a need to evaluate the use and response of these drugs in real life settings.

Objectives: This study aims to describe the drug utilisation and prescribing patterns of subcutaneous injections of biologic therapy for patients with inflammatory conditions in the largest Scottish health board.

Methods: The Homecare Service Database (HSD) contains information of all patients in the Greater Glasgow and Clyde (GGC) health board receiving medications by healthcare delivery companies. This innovative database has electronic records from 2012 to the present day. All the patients have their unique identifier, known as the Community Health Index (CHI) number. The GGC Biologics Cohort (GGC-BC) will be identified as any patient who had a prescription of a subcutaneous preparation of a biologic from HSD between 2012 and 2015. The records of the GGC-BC patients will be linked via CHI with other primary care from the Prescribing Information System (PIS), hospital admissions (SMR01), hospital based maternity and birth data (SMR02), cancer registrations (SMR06) and death records from the National Records of Scotland (NRS). The demographic characteristics of patients, prevalent and incident use of biologics, switching and discontinuation of these medications, use of concomitant medications, comorbidities and adverse events will be evaluated.

Results: A total of 3,502 patients (99% were more than 18 years old) received a biologic in the first quarter of 2015, a 1.4 fold increase in numbers since 2012. The average number of patients prescribed a sub-cutaneous biologic according to therapy group were: 82% rheumatology, 10% Crohn’s disease and 8% dermatology.

Conclusions: Through the novel record linkage of the homecare service database in GGC with other
healthcare datasets, the GGC Biologics Cohort will be created and the utilisation of biologics in chronic inflammatory conditions evaluated. This will provide a platform for future studies to better inform healthcare providers and authorities regarding patterns of use of biologics, their clinical effectiveness and safety in the short and long-term.

Abstract 1726

EXPERIENCES OF USING ROUTINE DATA TO DERIVE ESTIMATES OF BURDEN OF ASTHMA IN ENGLAND, NORTHERN IRELAND, SCOTLAND AND WALES AND DATA-GAP ISSUES

Authors:

M Mukherjee1, M Heaven2, R Gupta3, A Farr4, A Stoddart1, A Bandhopadhyay2, BNwaru5, M Shields6, R Lyons7, C Dibben8, D Strachan3, B McKinstry1, G Davies9, A Sheikh10

1Edinburgh Health Services Research Unit, Usher Institute of Population Health Sciences and Informatics, The University of Edinburgh, Edinburgh, UK, 2CIPHER—Centre for the Improvement of Population Health through e-Records Research, Centre for Health Information, Research and Evaluation (CHIRAL), College of Medicine, Institute of Life Science 2 (ILS2), Swansea University, Swansea, UK. 3Population Health Research Institute, St George’s, University of London, London, UK, 4Swansea Centre for Health Economics (SCHE), College of Human and Health Science, Swansea University, Swansea, UK. 5Usher Institute of Population Health Sciences and Informatics, The University of Edinburgh, Edinburgh, UK. 6Centre for Infection and Immunity, School of Medicine, Dentistry and Biomedical Sciences, Queen’s University Belfast, Belfast, UK. 7CIPHER - Centre for the Improvement of Population Health through e-Records Research, Centre for Health Information, Research and Evaluation (CHIRAL), College of Medicine, Institute of Life Science 2 (ILS2), Swansea University, Singleton Park, Swansea, SA2 8PP, UK. 8School of Geography & Geosciences, Department of Geography & Sustainable Development, The University of Edinburgh, Edinburgh EH8 9AG, UK. 9Asthma & Allergy Group, Institute of Life Science, College of Medicine, Swansea University, Swansea, UK. 10Centre of Medical Informatics, Usher Institute of Population Health Sciences and Informatics, The University of Edinburgh, Edinburgh, UK

Presenting Author: M Mukherjee

Corresponding author email: mome.mukherjee@ed.ac.uk

Introduction Since asthma is now one of the most common long-term conditions in the UK, we were commissioned by Asthma UK to build on earlier estimates of national prevalence and costs from asthma, filling data gaps previously identified in relation to healthcare and broadening the field of enquiry to include wider-societal costs. We have worked across the member countries i.e. England, Northern Ireland, Scotland and Wales, and have (1) produced a detailed overview of estimates of incidence, prevalence and healthcare utilisation; (2) estimated health and wider-societal costs; (3) identified any remaining information gaps and explored the feasibility of filling these and (4) provided insights into future research that has the potential to inform changes in policy leading to the provision of more cost-effective care.

Methods and analysis This involved secondary analyses of both stand-alone and linked datasets from national health surveys, general practice, prescribing, emergency care, out of hours, ambulance service, hospital, mortality and administrative data sources, to estimate the most comprehensive prevalence, healthcare utilisation and outcomes from asthma in the UK and member countries to date. Comparisons across nations were not always possible, especially where data were unavailable for certain outcomes in a country, time point, or were differentially defined across nations. The presentation will describe the experiences of the study team, the practicalities of this kind of UK wide study utilising available routine data. The lessons learnt have important bearings for mapping the epidemiology, disease burden and costs of other long-term conditions. The insights gleaned will be taken forward into the creation of the UK Asthma Observatory.
Abstract 1767  
**COSTS OF HOSPITAL CARE OVER TEN YEARS FROM A DIAGNOSIS OF CANCER**

**Authors:** Peter Hall¹, Armando Vargas-Palacios², Katie Spencer³  
¹University of Edinburgh, ²University of Leeds, ³Leeds Teaching Hospitals NHS Trust

**Presenting Author:** Peter Hall  
**Corresponding author email:** p.s.hall@ed.ac.uk

**Background:** Improvements in the efficiency of healthcare are dependent on data describing the costs of care. The costs of breast cancer care to healthcare providers are poorly reported, particularly beyond the period of initial diagnosis. For the purpose of service evaluation we now have the opportunity in the United Kingdom to cost healthcare activity in secondary care through the linkage of diagnostic and clinical events coded within local electronic clinical records and the financial data collected to facilitate commissioning and reimbursement.

**Methods:** 1,000 consecutive patients diagnosed with breast cancer (ICD10 C50) from January 1999 were identified retrospectively from the electronic clinical record of a large cancer centre and the Northern and Yorkshire Cancer Registration and Information Service (NYCRIS). Eligible patients were followed up for minimum of 10-years or until death if earlier, and had complete data linkage. Clinical events were verified by 100% clinical audit. Linked data was obtained from the local NHS finance records to provide costs of care based on the NHS England Payment-By-Results national tariff. Costs are reported in 2011 GBP and represent the cost to the local commissioner. Multivariate ordinary least squares regression of log-cost was used to identify predictors of cost. Predictive factors were selected on the basis that they are known predictors of clinical outcome.

**Results:** There were 805 patients with complete linkage. There were 43 local recurrences and 144 distant recurrences. The mean 10 year cost per patient was £17,707 (95% CI 15,662 – 19,752). The mean 10 year cost was £32,198 (25,721 – 38,674) from diagnosis of a distant recurrence and £31,290 (16,969 – 45,611) from local recurrence. The majority of the cost was incurred within the first 18 months after initial diagnosis or recurrence. Stage at diagnosis was the only statistically significant (p = <0.05) predictor of 10 year cost. Costs after a diagnosis of ovarian and colorectal cancer will also be reported along with the hospital costs of end of life care.

**Conclusions:** Cancer recurrence places a financial burden on hospitals which is of a noteworthy magnitude. Investment in the prevention of recurrence should be a priority for financial as well as clinical reasons. It is possible to use routine hospital finance data to assign costs to care at patient level with potential for utility in both prospective and retrospective evaluation.

Abstract 1783  
**DATA LINKAGE IN THE AUSTRALIAN LONGITUDINAL STUDY ON WOMEN’S HEALTH: USING DATA TO INFORM WOMEN’S HEALTH POLICY**

**Authors:** Deirdre McLaughlin¹, Caroline Jackson²  
¹School of Public Health, The University of Queensland, Herston, Queensland, Australia, ²University of Edinburgh, Usher Institute of Population Health Sciences and Informatics

**Presenting Author:** Caroline Jackson  
**Corresponding author email:** d.mclaughlin@sph.uq.edu.au

**Objectives:** The Australian Longitudinal Study of Women’s Health (ALSWH) has been investigating the health of more than 40,000 women since 1996. These survey data have been complemented by linkages to national and state administrative datasets. This presentation discusses our experiences and how these data have been used to inform and evaluate policy.
Approach: The ALSWH currently involves women in three age groups (1973-78 birth cohort, aged 18-23 in 1996; 1946-51 birth cohort, aged 45-50 in 1996; 1921-26 birth cohort, aged 70-75 in 1996) who were randomly selected from the Medicare Australia database, with deliberate oversampling from rural and remote areas. The ALSWH has approval to link participants’ data with data from the National Death Index, Medicare Australia, Admitted Patients data collections, Cancer Registries, Perinatal Data Collections and Aged Care datasets. However, the process of obtaining access to administrative datasets has been time consuming and costly.

Results: The use of longitudinal survey and linked administrative data has allowed the ALSWH to provide information on a broad range of outcomes for women and inform the development of the 2010 Australian Women’s Health Policy. In addition, findings have contributed significantly to policy development in areas including continence management, quality use of medicines, mental health among women and the impact of physical activity and BMI on health care costs in Australian women. Despite this, the process of obtaining ethical and data custodian approvals has provided many challenges and numerous iterations. Data linkages in the ALSWH have been approved by twelve Human Research Ethics Committees, twenty-two data custodians and seven data linkage units.

Conclusions: It is important to be able to evaluate the impact of health care policy and practice. Such evaluation cannot be undertaken from within any single system, but follow individuals within and across systems over time. Simplification of the process to access administrative data would allow more efficient use of resources.

NATURAL EXPERIMENTS

Abstract 1588 IMPACT ON INITIATION AND CESSATION OF STATINS AMIDST THE MEDIA INTEREST IN STATIN THERAPY SIDE EFFECTS PRIOR TO THE 2014 NICE LIPID MODIFICATION GUIDELINES

Authors: A Matthews, E Herrett, K Bhaskaran, L Smeeth
London School of Hygiene and Tropical Medicine

Presenting Author: A Matthews

Corresponding author email: Anthony.Matthews@lshtm.ac.uk

Introduction: In July 2014, the National Institute for Health and Care Excellence (NICE) amended their guidelines for statin therapy for the primary prevention of cardiovascular disease, lowering the cardiovascular disease risk threshold for which a patient should be recommended statins from 20% to 10% ten year risk. In the context of debate surrounding these impending changes, Abramson et al.[1] published a paper in October 2013 claiming that prescribing statins to people with a low risk of cardiovascular disease would increase the number of adverse events, without providing overall health benefit. A high volume of media coverage followed. Anecdotal evidence suggests that this has impacted initiation and cessation of statins, which would lead to increased cardiovascular disease risk.

Objective: To estimate changes over time in initiation and cessation of statins for both primary and secondary prevention, and the public health impact of media interest in statin side effects.

Methods: We will use the UK Clinical Practice Research Datalink (CPRD) to perform an interrupted time series analysis using segmented linear regression to assess trends in initiation and cessation of statins between 2011 and 2015, testing for changes in trend at the height of media interest (November 2013 - March 2014). We will assess trends for both primary and secondary prevention, and examine whether any effects differ by patient characteristics, cardiovascular risk factors or overall cardiovascular risk.

Results: The main analysis is currently in progress. Initial analyses suggest that after March 2014, the point at which media interest in statins side effect was at its highest, there was an increase in the proportion of patients stopping their statin prescription for both primary and secondary prevention.

Conclusion: This analysis will determine whether the media coverage of statin side effects was
associated with decreased initiation and increased cessation of statins, and the impact on subsequent cardiovascular disease risk.


Abstract 1716

TRENDS IN CAESAREAN SECTION RATES: ARE THRESHOLDS FOR DELIVERY FALLING AND NEONATAL OUTCOMES IMPROVING?

Authors: Anna Rose, E. Raja, S Bhattacharya, M Black

University of Aberdeen

Presenting Author: Anna Rose

Corresponding author email: a.rose.11@aberdeen.ac.uk

Aims: To analyse trends in caesarean section (CS) rates at Aberdeen Maternity Hospital (UK) since 1950, explore potential determinants of changes in primary, term unplanned CS rates, including indicators of clinical thresholds, and investigate the relationship between increased CS rates and neonatal outcomes.

Methods: Time trends analysis of CS rates for all deliveries ≥24 completed weeks gestation, 1950-2012 (N=261,425); and first births ≥37 weeks from 1982 onwards (n=67,053) recorded in the Aberdeen Maternity and Neonatal Databank. The relationship between changed CS rates and changes in maternal and clinical characteristics was explored (maternal body mass index greater than 25kgm⁻², maternal age over 35 years, smoking rate and rate of induction of labour). CS rates were also compared to indicators of clinical thresholds for unplanned CS (median duration of labour (in hours) before CS for delayed labour and percentage of CS for suspected fetal distress following confirmed abnormal fetal pH on fetal blood sampling, indicating fetal hypoxia) and neonatal outcomes (neonatal mortality rate (NMR); birth trauma; rate of neonatal unit (NNU) admission and 5-minute Apgar score <7 indicating poor condition at birth). These relationships were explored through univariate and multivariate linear regression. Binary logistic regression was used to calculate overall change in risk of CS per year.

Results: From 1950-2012, CS rates increased significantly from 3.0% to 28.2%. From 1982, the increase appeared to be largely explained by increased maternal age and higher induction of labour rates (R²=0.92). CS thresholds for delayed labour progression and suspected fetal distress decreased significantly (p<0.01). Decreased NMR (p=0.03) and low Apgar scores (p<0.01) appeared to be explained by increased CS rates (R²=0.16 and R²=0.25 respectively). There was a non-significant decreasing trend in birth trauma (p=0.15), but a positive trend in NNU admission (p=0.09).

Conclusions: CS rates have dramatically increased since 1950, especially unplanned CS which have more than doubled in three decades. This was mainly attributable to increases in maternal age and induction of labour rates – the latter being likely to reflect increasingly complicated pregnancies. There is a suggestion of a reduction in clinical thresholds for CS for delayed progress in labour and suspected fetal distress, but an improvement in neonatal outcomes during the study period (other than NNU admission) is clear, which may justify lower clinical thresholds. This study suggests increased CS rates may have net neonatal benefit.
Abstract 1729

SCI-DIABETES COLLABORATION: SUPPORTING QUALITY IMPROVEMENT AND SERVICE DELIVERY FOR NHS SCOTLAND

Authors: S G Cunningham¹, W M Urquhart², R R Mcalpine¹, A Emslie-Smith²
¹University of Dundee, ²NHS Tayside

Presenting Author: S G Cunningham

Corresponding author email: scott.cunningham@nhs.net

Background: Scottish Care Information-Diabetes Collaboration (SCI-DC) is the national diabetes information technology programme for NHS Scotland. Its key application, SCI-Diabetes, is a web-based clinical information system, supporting the care of people with diabetes. It captures data from primary, secondary and tertiary care, laboratory systems, specialist screening services and directly from patients to create a truly “shared” record. We aimed to demonstrate that SCI-DC supports quality improvement and delivers benefits to all users.

Methods: We aligned SCI-DC functionality to NHS Scotland’s Healthcare Quality Strategy and its six dimensions of quality. These six dimensions were taken in turn, with the functionality of SCI-DC systems matched to one or more. We then performed a quantitative analysis of 2014 data from system audit logs and process monitoring tools. National analyses showing changes in diabetes prevalence and clinical outcomes were referenced from the annual Scottish Diabetes Survey.

Results: 4,689 users logged in to SCI-Diabetes 326,517 times (average 69.6 logins/user) and accessed 9,465,862 web pages in 2014 (2018.7/user; 29/login). Effectiveness is demonstrated through routine reporting. Diabetes prevalence increased from 103,835 (2%) in 2002 to 276,430 (5.2%) in 2014. Despite this increase, data completeness continues to rise: Hba1c 94% complete in previous 15 months; cholesterol 90.7%; BMI 87.7%; ethnic group 81.1%. Patient education and records access is provided by the My Diabetes My Way website. 3,696 patients had logged in to access their records by the end of 2014 (92.2% increase since 2013; n=1,923). Automated systems lead from diabetes diagnosis to retinopathy screening appointment booking in two days. SCI-DC supports single data entry by transmitting secondary and tertiary care data to primary care. In 2014, 3,302,196 data items were sent to ~1000 GP sites in 1,520,051 XML messages (2.2 data items/message) for 269,982 patients (12.2 data items/patient; 5.6 messages/patient).

Conclusions: SCI-DC supports diabetes services and the six dimensions of quality, leading to improved collaborative working and clinical outcomes. Recommendations for national health care policy and planning are based on SCI-DC data. Secondary use contributes to the clinical evidence base, local, regional and national reporting and facilitates patient education and self-management. Across the UK and beyond, SCI-DC is considered a ‘gold-standard’ system unrivalled amongst other long-term conditions.

Abstract 1730

BIRTH AND EARLY CHILDHOOD OUTCOMES ASSOCIATED WITH ASSISTED REPRODUCTIVE TECHNOLOGIES: THE MASSACHUSETTS OUTCOMES STUDY OF ART (MOSART) FINDINGS, 2004-2010

Authors: Milton Kotelchuck¹, E Declercq², H Diop³, B Luke⁴, J Stern⁵, H Cabral²
¹Massachusetts General Hospital, ²Boston University School of Public Health, ³Massachusetts Department of Public Health, ⁴Michigan State University College of Human Medicine, ⁵Dartmouth School of Medicine

Presenting Author: Milton Kotelchuck

Corresponding author email: mkotelchuck@mgh.harvard.edu

Background: In the U.S., efforts to explore the impact of ART on reproductive and early childhood health have been hindered by the absence of a non-ART-treated subfertile comparison population (to
distinguish between ART treatment and underlying subfertility), and limited population-based samples assessing health beyond the birthing period. The Massachusetts Outcomes Study of Assisted Reproductive Technologies (MOSART), which addresses these prior concerns, utilizes the linkage of the Massachusetts Pregnancy to Early Life Longitudinal public health data system to the Society for Assisted Reproductive Technology Clinical Outcomes Reporting System, a treatment-cycle-based clinical database, to create a unique, longitudinal database to assess the impact of ART on reproductive and child health. This presentation will synthesize and update findings of prior MOSART reports on birth outcomes with our newer findings on early (0-3 years) childhood impacts.

**Methods:** Longitudinal cohort design; comparing outcomes of three groups of women (fertile, subfertile without ART, and subfertile with ART treatment); with births to the fertile group as the reference, stratified separately among singletons and multiple births. Unadjusted prevalence, adjusted odds ratios, and 95% CI are calculated (adjusted for maternal age, education, race/ethnicity, marital status, nativity, parity, insurance source, and chronic diabetes and hypertension; childhood outcomes were further adjusted for C-sections and birth outcomes). The study sample includes all Massachusetts births from 7/2004 through 12/2010 (N = 481,588; ART = 17,673; subfertile = 9,792; fertile = 454,123). Outcomes examined included: distribution plurality status; reproductive health measures (LBW, prematurity, infant mortality); and early childhood health measures (birthing hospital length of stay (LOS), childhood rehospitalizations, rehospitalization LOS, birth defects, specific childhood morbidities, childhood mortality, and enrollment in Early Intervention programs). (Data presented is from 2004-2008, to be updated through 2010 for the Conference.)

**Selected Results:** Unadjusted prevalence, for singletons only, presented here; with data listed sequentially for ART, subfertile without ART, and fertile populations. Singleton status (57%, 82%, 97%); LBW (7.8%, 5.7%, 5.4%); preterm delivery (10.2%, 8.1%, 6.4%); neonatal mortality per 1000 births (3.72, 8.01, 2.56); birthing hospital LOS in days (4.8, 4.1, 3.8); rehospitalization (9.1%, 9.2%, 10.6%); rehospitalization LOS in days (4.9, 5.0, 4.8); birth defects (2.0%, 2.1%, 1.6%); and EI program enrollment (18.1%, 16.6%, 15.4%).

**Conclusions:** From a population perspective, beyond increased plurality, ART appears to have some impact on birth outcomes [especially among singletons, but not twins (data not shown)], but much less than effects of subfertility. There were only minor subsequent impacts on early childhood health measures.

---

**Abstract 1752**

**CO-PRESCRIBING MENTAL AND PHYSICAL HEALTH MEDICATION LEADS TO HIGHER USE OF PHYSICAL HEALTH SERVICES: A RECORD LINKAGE STUDY FOR THE WHOLE POPULATION OF AUSTRIA**

**Authors:**

Heinz Katschnig¹, Simone Sauter¹, Florian Endel¹, Christa Strassmayr¹, Gottfried Endel²

¹IMEHPS.research, ²Main Association of Austrian Social Security Institutions

**Presenting Author:**

Heinz Katschnig

**Corresponding author email:**

heinz.katschnig@meduniwien.ac.at

**Background:** Awareness of the high comorbidity of mental and physical disorders is increasing. Research based on large routine health care databases offers a possibility to study such comorbidity on a population level. One approach, used in this study, is to analyze the co-prescription of medication for physical and mental disorders and to follow the identified patients up by record linkage methods.

**Methods:** In Austria (8.5 million inhabitants) a database is under construction linking patient records from different data sources covering contacts with primary care and specialist outpatient physicians, hospital admissions and filling of prescriptions in community pharmacies for the entire population since the year 2006. For the years 2006 and 2007 the database (called the GAP-DRG) is finalised and was used for the present study. A cohort of patients was identified who had at least one prescription (of any type) filled in the 4th quarter of 2006, and was then divided into subgroups of psychotropic medication only, physical health medication only, and a combination of both over the three month baseline.
period. The whole cohort and each subgroup was studied for contacts with general practitioners, psychiatric and non-psychiatric specialised outpatient services, as well as hospital admissions to psychiatric and non-psychiatric hospital beds in the year 2007.

**Results:** In the year 2007 only one in 20 patients of the whole cohort (3,540,001 patients) had at least one contact with an outpatient psychiatrist, but 64% contacted a non-psychiatric outpatient specialist. Only one in 100 patients was admitted to a psychiatric hospital bed, but 25% to a non-psychiatric hospital bed. The subgroup of patients with co-prescription of physical and mental health medication (547,847 patients) had the highest contact rates with non-psychiatric outpatient services and was substantially more often admitted to non-psychiatric inpatient services (33%) than the group with physical medication only (19%).

**Conclusions:** Co-prescription of medication for physical and mental disorders leads to substantially higher utilization of physical health care services than prescription of physical health medication without psychotropic co-prescription does. Whether this is due to a specific help-seeking behaviour of comorbid patients, or due to a psychotropic co-prescription being an indicator of severity of an existing physical health problem, could be potentially clarified by more in-depth analyses of specific medications and the time sequence of different types of prescriptions.

---

**Abstract 1757**  
**QUANTIFYING THE RISK OF ACUTE KIDNEY INJURY ASSOCIATED WITH NON-STEROIDAL ANTI-INFLAMMATORY DRUGS IN PEOPLE WITH CHRONIC KIDNEY DISEASES**

**Authors:** Xinyu Zhang, Bruce Guthrie, Peter Donnan  
Division of Population Health Sciences, School of Medicine, University of Dundee

**Presenting Author:** Xinyu Zhang

**Corresponding author email:** x.l.zhang@dundee.ac.uk

**Background:** Non-steroidal anti-inflammatory drugs (NSAIDs) are very widely used painkillers, and are both prescribed and can be consumed over-the-counter. NSAIDs are a common cause of adverse drug events (ADEs). However, compared to gastrointestinal (GI) and cardiac ADEs, renal risks of NSAIDs are less well characterised and quantified because manifestation of GI and cardiac AEs are often obvious whereas there are no specific signs and symptoms of AKI.

It is believed that people who already have chronic kidney disease (CKD) are also at higher risk from NSAIDs, but there is relatively little data that measures exactly how risky an NSAID is in people with CKD. Clarifying when the renal risks of NSAIDs are significant enough to require either avoidance or additional intensive monitoring has the potential to significantly alter practice, and to improve safety.

**Aims:** To measure the rate of NSAID use among people with CKD and how it has changed over time. To estimate the crude and adjusted rate ratio of AKI in people with CKD exposed to NSAIDs, and by calculating baseline risk to estimate the absolute risk of AKI for people with CKD exposed to NSAIDs.

**Methods:** Nested case-control study analysis of cohort data will be used to more accurately estimate the AKI risk of NSAIDs in people with CKD, using routinely held clinical and laboratory data extracted by the Health Informatics Centre (HIC). The population studied will be people ≥18 years old with CKD resident in Tayside between 2000-2014 and/or resident in Fife from 2009-2014. CKD and AKI are determined as glomerular filtration rate (GFR) <60 ml/min per 1.73m² for >3 months and increase in serum creatinine (Scr) by 50% within 7 days, OR increase in Scr by 0.3 mg/dl (26.5mmol/l) within 2 days, respectively based on Kidney Disease Improving Global Outcomes (KDIGO) definitions. NSAID exposure will be determined using dispensed prescribing data. SPSS 21.0 and SAS 9.2 will be used for data analysis.

**Results:** The study is under the process of cohort defining and data cleaning. Once the cohort is defined, cases and controls will be matched and rate ratio as well as absolute risk of AKI in people with CKD exposed to NSAIDs will be calculated. Further findings will be presented in August.
Conclusion: This large population based study using current international definitions of AKI to measure outcomes will estimate the absolute risk of harm of NSAIDs to better inform clinical decision making.

Abstract 1762 RETROSPECTIVELY DEFINING A STUDY COHORT USING HOUSING INTERVENTION DATA

Authors: Rhodri Johnson¹, S Rodgers¹, R Bailey¹, A Lacey¹, J Demmler¹, W Poortinga², F Dunstan², S Lannon², R Smith², P Anderson¹, R Lyons¹
¹Swansea University, ²Cardiff University

Presenting Author: Rhodri Johnson

Introduction: Housing quality has been linked to changes in health though there is a noted lack of rigorous epidemiological evidence. Welsh Government have invested in housing quality improvements through the Welsh Housing Quality Standard (WHQS). Carmarthenshire County Council’s housing improvement project aimed to achieve WHQS and formed the basis of this health evaluation study. We linked housing intervention, demographic and health data in SAIL to evaluate the housing interventions.

Methods: Intervention datasets from two local authorities were processed by a trusted third party and anonymised into SAIL which formed our cohort and comparator groups. An iterative review process was completed to create metadata describing and classifying the interventions which consisted of one to eleven separate elements of work.

Final datasets were created linking participants resident within intervention properties for relevant time periods. A regional comparator dataset was created using all properties and residents within Carmarthenshire over a specific time period.

Results: The data review demonstrated intervention durations ranged from one day to eight years and informed our study design; a ‘multiple interrupted time series’. A cohort qualification rule was created stipulating that people must be resident for a minimum 90 days pre-intervention. The database structure and linkage methods allowed migrating participants to be retained.

Final datasets were created with one row per month for each person remaining within the study, allowing changes in interventions, demographics and health outcomes to be available for statistical analysis. Following application of qualification rules the cohort consisted of 20,000 participants in around 8,500 homes equating to 128 thousand person years with 1.5 million rows of data. The first comparator group contained 50,000 people with 298 thousand person years, the second comparator group contained 240,000 people with 1,716 thousand person years.

Discussion: Sourcing and interpreting secondary data can involve significant researcher time and resource, especially when sourcing from multiple organisations; it would be beneficial if national data collection, storage and dissemination standards could be developed to reduce resource requirements and increase research quality.

This work has resulted in a reliable dataset on which to evaluate our health outcomes of interest.
Abstract 1357  

SOCIOECONOMIC STATUS, COMORBIDITY, AND MORTALITY IN PATIENTS WITH DIABETES IN SCOTLAND 2004-2011

Authors:  
Jeremy J Walker, Sarah H Wild  
The University of Edinburgh

Presenting Author:  
Sarah H Wild

Corresponding author email: jeremy.walker@ed.ac.uk

Background: Mortality in people with diabetes (as in many populations) often exhibits marked social patterning, risk of death being greater in deprived groups. This may reflect deprivation-related differences in comorbid disease (conditions additional to diabetes itself). This study sought to determine whether the social patterning of mortality in a type 2 diabetic population is explained by differential comorbidity.

Methods: Data drawn from the Scottish Care Information - Diabetes Collaboration (SCI-DC) dataset were used to identify 70,197 men and 56,451 women diagnosed with type 2 diabetes at 25 years of age and above in Scotland during the period from 1st January 2004 to 18th May 2011. Linked hospital admission records covering the five years prior to diabetes diagnosis were used to construct comorbidity histories for this cohort. Additional linkage to death registration records permitted determination of vital status (dead vs. alive) at one year after diagnosis with diabetes. All linkage was performed by the Information Services Division (ISD) of NHS National Services Scotland. Logistic models were fitted to estimate mortality at one year after diagnosis with diabetes, predicted initially by age and socioeconomic status (SES). The representation of SES used was the Scottish Index of Multiple Deprivation (SIMD; a small-area based multi-dimensional measure of deprivation). Separate models were fitted for each sex. These models were then extended to include in turn five representations of comorbidity: the Charlson Index; the Elixhauser comorbidity set; number of hospital in-patient bed-days; number of unique British National Formulary codes in the patient’s prescribing history; and a set of comorbidities developed by ISD to guide coding of Scottish hospital admission records. The capacity of comorbidity to explain social mortality gradients was assessed by observing the change in regression coefficients for SES following the addition of each measure of comorbidity.

Results: The number of deaths observed in the cohort was 3,059. Adjustment for comorbidity attenuated the association between SES and mortality. However, in nine of ten models fitted (five comorbidity measures X two sexes), SES remained significantly predictive ($p < 0.05$) of mortality after comorbidity was taken into account.

Conclusion: The social patterning of mortality in a diabetic population is not completely explained by differing levels of disease additional to diabetes itself. Other dimensions of deprivation are implicated in the elevated death rates observed in deprived groups of people with diabetes.
Abstract 1465

ELECTRONIC LONGITUDINAL ALCOHOL STUDY IN COMMUNITIES (ELASiC)

Authors: Simon Moore
Cardiff University

Presenting Author: Simon Moore

Corresponding author email: mooresc2@cardiff.ac.uk

There exists a unique opportunity to exploit existing UK data to better understand alcohol misuse across the life course. The overall goals of this project are to 1) incorporate data into the UK Secure eResearch Platform (UKSeRP) and develop facilities to enable research access, 2) undertake hypothesis driven research using this platform to provide critical insights into alcohol use, its effects and pathways into harm, 3) exploit the current research group’s expertise to capitalise on the most recent developments in analytical methodology, and 4) make explicit the policy relevance of the work and exploit opportunities to interface with possible intervention development. This project therefore aims to leverage the value of a broad set of longitudinal studies and data linkage facilities to construct an analytical platform within UKSeRP that facilitates the investigation of harmful alcohol use across the life course. The proposal is balanced, it combines a broad skill set, with considerable expertise with the available data together with techniques in the analysis of such data.

The data itself covers youngsters, those in mid- and later-life. These data are then linked to routine administrative data including those from schools (e.g. educational attainment), the NHS (health), the police (anti-social behaviour) and other sources including births and deaths. These then allow us to understand how alcohol affects general health, health service use, mental health, educational attainment and criminality. The nature of these data also allow us to understand how alcohol use affects cognitive decline in later life, and emotional and cognitive development in young people.

There has already been considerable activity generating opportunities to use longitudinal data for research, although only limited work has looked at alcohol specifically. We are therefore able to exploit these on-going investments and collaborate with the UK Dementia Platform, Centre for the Improvement of Population Health through e-Records Research (CIPHER) at Swansea University (funded by a consortium led by the Medical Research Council) which will provide technical and financial support for data linkage. In addition, the project will explicitly examine the policy and impact relevance of outputs, use outputs to investigate opportunities to develop novel interventions with, for example, the Centre for the Development and Evaluation of Complex Interventions for Public Health Improvement (funded by the Economic and Social Research Council, amongst others). These collaborations bring together a unique set of skills that are required for such work as well as technical support.

Abstract 1471

LINKING SERVICE USE AT A PSYCHIATRIC HOSPITAL TO UNEMPLOYMENT AND SYMPTOMATOLOGY

Authors: S J Kiddle, R Patel, R Jackson, R Stewart, R J B Dobson
King’s College London

Presenting Author: S J Kiddle

Corresponding author email: steven.kiddle@kcl.ac.uk

In this work I will present preliminary work from two studies, both looking at service use in a large psychiatric hospital, the South London and Maudsley (SlaM) NHS foundation trust. Pseudoanonymised Electronic Health Records (EHRs) for SlaM are available through the groundbreaking Clinical Record Interactive Search (CRIS) tool/database. CRIS allows access to EHR covering around 250k patients, including around 18 million free-text documents. For these studies service use will be extracted from structured EHR data.
Project 1 will focus on out-patient service use for patients with Schizophreniform disorders, mood disorders and substance abuse disorders. SLAM patients come from the Southwark, Lewisham, Lambeth or Croydon boroughs. Annual out-patient service use will be compared to the proportion of households claiming unemployment benefits within small areas (1,000 – 1,500 residents), to see if areas with higher unemployment also require greater psychiatric out-patient support. In a natural experiment, data is available before and after the recession of 2008, giving us a great opportunity to relate changes to employment and mental health using routinely collected data.

Project 2 will focus on how psychosis symptoms evolve over time, and the relation between psychosis symptoms, medication and service use. Methods will be developed and applied to model the temporal evolution of symptoms, medication and service use.

Abstract 1494

**LINKING HOSPITAL DATA WITH BIRTH RECORDS TO EXAMINE SECOND TRIMESTER PREGNANCY OUTCOMES**

Authors: Charles Algert, Christine L Roberts, Margaret Harpham, Jane B Ford

*Kolling Institute, University of Sydney*

Presenting Author: Charles Algert

Corresponding author email: charles.algert@sydney.edu.au

**Background:** Studies of pregnancy using population data typically use birth datasets, often linked to hospital admissions. The study populations are “births”, with birth usually defined by some minimum gestational age such as 24 weeks. This is a significant constraint when pregnancy outcomes from the second trimester (14-27 weeks) are important. Linked hospital admissions for women can be used to identify admissions for miscarriage/early fetal death. As one example, this approach is necessary for a full assessment of the use of the cerclage procedure in the maternal population. Insertion of a stitch around the cervix (cerclage) can be used to reduce the risk of premature birth and is typically inserted before 18 weeks.

**Methods:** Data were obtained from linked birth and hospital records in NSW Australia. Births are recorded from 20 weeks gestation, or 400 grams birthweight. The study period was over two years: pregnancies with an estimated date of conception between 1 January 2010 and 31 December 2011. This allowed for birth of a term infant before December 2012. The hospital records include >20 fields each for diagnosis and procedure codes. The primary outcome of interest for evaluation of cerclage was spontaneous birth and fetal death, which could be identified from diagnosis codes; medical terminations were identified from diagnosis and procedure codes. Gestational age at <20 weeks (from the hospital data) is only coded by category: <14 weeks, 14-19 weeks, or unknown.

**Results:** There were 191,989 deliveries in the birth records during the two years, of which 1227 were delivered before 28 weeks; 600 of these were at 20-23 weeks. However, the birth records include medical terminations. Excluding terminations (identified from linkage to the hospital data), the birth records reduced to 408 spontaneous deliveries for the four week interval at 20-23 weeks gestation. The hospital data identified 575 spontaneous miscarriages in the six weeks comprising the 14-19 week gestational age interval. Linking the hospital data with the birth records resulted in 983 spontaneous miscarriages/births at 14 to 23 weeks over the two years. Among these, 88 (9.0%) women had a cervical cerclage.

**Conclusions:** Linkage of hospital admissions with birth datasets improves ascertainment of pregnancy outcomes. The linkage is essential when early pregnancy outcomes are important.
Abstract 1548

GESTATION OF DELIVERY OF TWINS – INFLUENCE ON PERINATAL MORTALITY AND MORBIDITY AND CHILDHOOD EDUCATIONAL OUTCOMES

Authors:
Sarah Murray1, J Pell2, S Stock1, J Norman1
1University of Edinburgh, 2University of Glasgow

Presenting Author: Sarah Murray

Background: Twin pregnancy is associated with an eight to tenfold increase in perinatal mortality. Reduction of stillbirth and perinatal mortality is a major priority for the NHS in both England/Wales and Scotland. The incidence of multiple births is rising. Twin pregnancy is associated with a significantly higher risk than singleton pregnancies including increased rates of preterm birth, intrauterine growth restriction and stillbirth. Data from singleton pregnancies suggest that in terms of perinatal death it is safer for the baby to be delivered at 38 weeks compared to continuing the pregnancy. In contrast to this data, there is little evidence on the optimum gestation of delivery of twins.

Aims: The aim of this project is to make a major contribution to determining the optimum timing of delivery of twins to reduce perinatal mortality and morbidity. We aim to use existing epidemiological cohorts to determine gestation specific perinatal and neonatal mortality and morbidity in twin pregnancies following spontaneous delivery and elective delivery in the absence of maternal complications. We also aim to determine the long term educational outcomes of twin delivery by gestational age including record of additional educational support need, type of need, educational attainment level and school leaver destination.

Methods: We will use a number of databases to carry out the research aims listed above. The Scottish Morbidity Record (SMR02) collects information on all women discharged from Scottish maternity hospitals and will be used to carry out a population based retrospective cohort study of all twin deliveries at 34 weeks gestation or greater delivered from 1980 onwards. Outcomes will be analysed by chorionicity and trajectory to delivery. A number of maternal and neonatal outcomes will be recorded including perinatal mortality, admission to neonatal unit, mode of delivery, postpartum haemorrhage, shoulder dystocia and uterine rupture. Multivariate logistic regression modeling will be used to examine the relation between outcomes of elective induction of labour and expectant management for each gestation from 34 weeks. SMR02 records for all twin deliveries will then be linked to the ScotXed school census and the SVQ databases to explore the relationship between gestation of delivery and educational support need.

Abstract 1554

USING LINKED ADMINISTRATIVE HEALTH DATABASES TO ESTIMATE INCIDENCE AND PREVALENCE OF MULTIPLE SCLEROSIS IN BRITISH COLUMBIA, CANADA (1991-2008)

Authors:
Elaine Kingwell1, Feng Zhu1, Ruth Ann Marrie2, John D. Fisk3, Christina Wolfson4, Sharon Warren5, Joanne Profetto-McGrath6, Lawrence W. Svenson4, Nathalie Jette7, Virender Bhan3, B. Nancy Yu2, Lawrence Elliott2, Helen Tremlett1
1University of British Columbia, 2University of Manitoba, 3Dalhousie University, 4McGill University, 5University of Alberta, 6Surveillance and Assessment Branch, Alberta Health, Government of Alberta, 7University of Calgary

Presenting Author: Elaine Kingwell

Canada has a universal publicly funded health care system. In the western-most Canadian province, British Columbia (BC), health claims data are captured for the entire population (approximately 4.5 million). We used population-based administrative health databases and a previously validated case definition to estimate the incidence and prevalence of multiple sclerosis (MS) in BC over 13 and 18 year
periods, respectively.

We accessed the BC Hospital Separations, Medical Services Plan (physician visits), Vital Statistics (deaths), and Health Registration files to identify all BC residents meeting the case definition of > 3 International Classification of Disease codes for MS (ICD-9: 340 or ICD-10-CA: G35). Point prevalence was estimated annually on July 1, 1991-2008. Incidence and prevalence were calculated per 100,000 people using the BC mid-year population, and were age-standardized to the 2001 Canadian population. Incidence estimates were generated based on the year of the first demyelinating disease claim for 1996-2008; a preceding 5-year demyelinating disease claim-free period was required. We investigated changes in incidence, prevalence and sex ratios over the observation periods using Poisson and log binomial regression.

On July 1, 2008, an estimated 11,184 people with MS were living in BC, giving a standardized point prevalence of 235.8 per 100,000 persons (95% CI:231.4-240.3). The prevalence in 1991 was 90.3 (95% CI:87.0-93.7), and increased by approximately 5.5% per year on average over the 18 year observation period (p<0.001). The female: male prevalence ratio increased gradually from 2.27 in 1991 to 2.71 in 2008 (p<0.001) and the peak prevalence age increased from 45-49 years in 1991 to 55-59 years in 2008. From 1996 to 2008 there were 5,876 incident MS cases in BC, with an average annual age-standardized incidence rate of 10.9 (95% CI:10.6-11.2) per 100,000. The annual incidence rate was essentially stable over the 13 year period while the incidence sex ratio decreased somewhat over time (p=0.02), but remained above 2 for all years (averaging 2.72:1).

The incidence and prevalence of MS in BC are among the highest in the world, although comparable to estimates from eastern and central Canadian provinces. Neither the incidence of MS nor the incidence sex ratio showed an increase over the study period; however, MS prevalence and the prevalence sex ratio increased significantly. Population-based administrative health databases and validated case definitions using health claims data can provide a reliable, accessible and cost-effective means of monitoring the incidence and prevalence of a chronic disease such as MS.

Abstract 1589 LONGITUDINAL HOSPITAL UTILISATION BY ABORIGINAL VERSUS NON-ABORIGINAL CHRONIC DISEASE PATIENTS IN WESTERN AUSTRALIA

Authors: David Whyatt¹, J Marsh², G Pearson², R Tenneti¹, A Vickery¹

¹University of Western Australia, ²Telethon Kids Institute

Presenting Author: David Whyatt

Corresponding author email: david.whyatt@uwa.edu.au

Australia’s Aboriginal populations have significantly different age of onset of disease and health utilisation patterns. Aboriginal people have a higher burden of illness and a life expectancy at birth that is 15-20 years less than non-Aboriginal people.

This study used linked data from all patients in Western Australia between 2002 and 2010 who had a diagnosis of chronic disease at a hospital admission and/or ED presentation. The use of hospital services, in terms of inpatient days and emergency department (ED) presentations, was longitudinally modelled in the six years prior and the four years following a cardinal event. A cardinal event was defined as a hospital admission with a diagnosis of chronic disease (i.e. type 2 diabetes, heart failure, and/or chronic obstructive pulmonary disease), where that event has not occurred in the previous two years.

It was found that Aboriginal patients use the same number of inpatient days as non-Aboriginal patients in the years prior to and following a cardinal event. However, Aboriginal patients had approximately three to four times higher rate of ED presentations (for any cause) in this ten year period. It could be suggested that differences in age, socioeconomics, access to services, and co-morbidity could be associated with this phenomenon. However, using propensity scoring analysis, it was found that these factors only explained a third of the difference in ED presentation rates between Aboriginal and non-
Aboriginal patients. It is concluded that social norms and behaviours, both amongst Aboriginal patients and health care providers, may be more important determinants of ED utilisation than previously thought. The potential impact on chronic management in the Australian Aboriginal community will be discussed.

Abstract 1598

**UTILISATION OF EVIDENCE-BASED STATIN TREATMENT FOR THE PREVENTION OF ACUTE CORONARY SYNDROMES IN ABMU HEALTH BOARD**

Authors: Jo Davies  
Swansea University

Presenting Author: Jo Davies

Corresponding author email: jo.davies@swansea.ac.uk

Despite major advances in treatment over the past decade, Acute Coronary Syndromes (ACS) remain to be a major cause of morbidity and mortality in Wales. Guidelines published by the National Institute for Clinical Excellence (NICE) recommend initiating statin therapy in all high-risk patients, which has proven to be highly effective in preventing ACS. Despite this, previous research has highlighted that treatment is often underutilised due to under prescribing, resulting in adverse health outcomes and an increased burden on the healthcare system. The aim of this study was to determine the extent to which evidence-based, primary prevention statin therapy is being utilised in patients with a ≥20% 10-year risk of ACS and to identify any gaps in treatment provision. A retrospective cohort study was carried out using linked, administrative health data to assess statin prescribing in high-risk patients in a Welsh health board and to determine whether certain factors are associated with under-prescribing of statin medication. Time to event analyses along with the Cox proportional hazards model were used to assess the time between being identified as high-risk and receiving statin medication, and to examine the association of variables with statin prescribing such as age, gender, socioeconomic status, risk level, and practice size. An overview of the methods used in this study will be presented along with the study results.

Abstract 1605

**A RECORD-LINKAGE STUDY TO INVESTIGATE POSSIBLE ASSOCIATIONS BETWEEN HPV-INFECTION AND PRE-TERM LABOUR AND BIRTH**

Authors:  
MC. Aldhous1, R Bhatia2, J Caldwell3, Je Norman1, Sj Stock1  
1Centre for Reproductive Health, University of Edinburgh, 2HPV Research Group, Dept Pathology, University of Edinburgh, 3eDRIS, National Services Scotland

Presenting Author: M C Aldhous

Corresponding author email: m.aldhous@ed.ac.uk

**Background:** Preterm labour and birth (<37 weeks of pregnancy) affects 5-18% of all pregnancies. Uterine bacterial infection accounts for ~30% of cases but ~50% of cases are of unknown cause. Preterm labour and birth cause complications for the new-born baby and longer-term disabilities (e.g. cerebral palsy, learning difficulties and sensory impairments). Human papilloma virus (HPV) infects cervical epithelial cells and is normally cleared by the immune system. In some women HPV-infection persists, causing cervical intraepithelial neoplasia (CIN), leading to cervical cancer and is associated with specific HPV-strains (HPV-16 and -18). Treatment of CIN is by removal of abnormal cells. Previous studies suggested that HPV-infection may be associated with preterm birth: whether this is due to HPV-infection itself or the development and/or treatment of CIN is unclear. Cervical cells have
barrier and immune functions, which protect the uterus from infection and prevent preterm labour. In pregnancy, prior or current HPV-infection may alter the functions of cervical cells, thus affecting pregnancy outcomes. CIN and its treatment might also affect the cervical barrier and/or immune functions, leading to poor pregnancy outcomes.

**Aims**

1. To investigate whether HPV-infection is associated with spontaneous preterm birth, miscarriage or stillbirth
2. To investigate whether HPV-subtypes (HPV-16, HPV-18 or others) are associated with spontaneous preterm birth, miscarriage or stillbirth
3. To investigate whether excisional treatments for CIN are associated with spontaneous preterm birth, miscarriage or stillbirth

**Study Design and Methods:** The Scottish HPV Archive is a biorepository of residual cervical samples from routine screening, with linked HPV results, clinical data and identifiers which are available for HPV related research, through application.

We are using record-linkage of the HPV Archive with SMR01, SMR02 and NRS Birth, Stillbirth and Death records, in a cohort of women who have had a pregnancy subsequent to index sample collection. Anonymised HPV-data (HPV-status, HPV-genotype, degree of dyskaryosis and CIN result) will be linked to data from SMR01 (miscarriage), SMR02 and/or NRS Birth, Stillbirth and Death Records regarding pregnancy outcome (miscarriage, gestation at delivery, onset of labour, birth-weight, gender, perinatal mortality) potential confounders and previous pregnancy details.

If an association between preterm labour and HPV-infection is found, this gives a potential pathogenic mechanism on which to base future clinical, epidemiological and laboratory studies. If a specific HPV-strain is implicated this would also give potential pathogenic mechanisms to investigate, but also could suggest a potential screening factor for risk of preterm labour.

---

**Abstract 1637**

**A TWO-COUNTRY ELECTRONIC HEALTH RECORD COHORT STUDY IN ENGLAND (CALIBER) AND NEW ZEALAND (PREDICT): ASSOCIATION OF WHITE CELL COUNT WITH ALL-CAUSE MORTALITY**

**Authors:**

Anoop D Shah¹, Simon Thornley², Sheng-Chia Chung¹, Spiros Denaxas¹, Rod Jackson³, Harry Hemingway¹

¹Farr Institute of Health Informatics Research, UCL Institute of Health Informatics, University College London, London, UK, ²Counties Manukau District Health Board, New Zealand, ³University of Auckland, New Zealand

**Presenting Author:** Spiros Denaxas

**Corresponding author email:** a.shah@ucl.ac.uk

**Background:** Total white blood cell count (WBC) is a commonly performed blood test. Associations with all cause mortality have been suggested, but there have been no large studies of associations between clinically recorded WBC and mortality.

**Methods:** We performed a cohort study in England and New Zealand using linked health records. The English study was CALIBER (primary care, hospitalisations, mortality and acute coronary syndrome records). The New Zealand study was PREDICT (cardiovascular risk assessments in primary care, hospitalisations, mortality, dispensed medication and laboratory results). People aged 30 to 75 years with no prior cardiovascular disease were eligible for inclusion, and were followed until death, transfer out of the practice (in CALIBER) or study end.

**Results:** The study included 686,475 individuals in CALIBER and 194,513 in PREDICT. Hazard ratios (HRs) showed ‘J’-shaped associations between baseline WBC and mortality; the second quintile was associated with lowest risk in both cohorts. The two highest WBC quintiles were associated with higher mortality (adjusted HR (95% confidence interval) for highest compared to the middle quintile 1.90 (1.82,
1.99) in CALIBER, 1.53 (1.31, 1.79) in PREDICT. Associations were strongest close to the time of WBC measurement (CALIBER: adjusted HR 4.12 (3.69, 4.60) for the first 6 months, 1.60 (1.53, 1.68) thereafter; PREDICT: adjusted HR 2.67 (1.82, 3.92) for the first 6 months, 1.49 (1.27, 1.76) thereafter). There was a weaker association of high white cell count with mortality among younger patients, but no statistically significant interactions with sex or ethnicity.

**Conclusions:** Elevated WBC as recorded in routine clinical care was associated with increased mortality even within the ‘normal’ range. Associations were strongest close to the time of WBC measurement, but a weaker association persisted for several years. Replication in independent populations increases confidence in these findings.

**Abstract 1733**

**PATTERNS OF FETAL AND BIRTH MEASUREMENTS: IS THERE AN ASSOCIATION WITH CHILDHOOD AUTISM SPECTRUM DISORDER?**

**Authors:** Markus Steiner¹, S Turner¹, G Devereux¹, J Williams², J Crum², R Wood⁴

¹University of Aberdeen, ²University of Aberdeen, ³NHS Grampian, ⁴NSS

**Presenting Author:** Markus Steiner

**Corresponding author email:** m.steiner@abdn.ac.uk

**Background:** Autism spectrum disorders develop early in life and there is a strong suspicion that early life exposures contribute to the development of the condition. A systematic review identified a broad range of risk factors including infections during pregnancy, maternal and paternal age, maternal asthma and allergies, and maternal haemorrhage during pregnancy.

**Methods:** This unmatched case control study has been utilising fetal scan measurements at 12 and 20 weeks gestation and birth in the Aberdeen Maternal and Neonatal Database (AMND) covering all births in Grampian from the 1985 to 2013 which have been linked for this study to cases of children diagnosed between 1995 and 2013 with autism spectrum disorders or Asperger’s (ASD) and learning or developmental difficulties in the Support Needs System (SNS). Additional information have been sought from SMR00 outpatient data and SMR04 mental health about possible ASD events and the CHI register to allow for censoring for people who left NHS Grampian. Additional information on possible pre- and peri-natal factors, e.g. prematurity and maternal smoking as mediating or confounding factors like pre-natal medication for antiepileptic drugs, antidepressants and immune-modulating drugs are included in the database. Deviation from normal fetal growth trajectories have been calculated as z-standardised residuals.

**Data:** 137067 live births have been collected between 1985 and 2013 in the AMND database, 58313 have valid scan data which can be used to assess fetal growth trajectories and the fetal development. 1234 children with a diagnosis of ASD or with learning or developmental difficulties in the SNS database have been linked to the AMND data. 979 of these children have a diagnosis of ASD, 255 only have learning and developmental difficulties. 483 children with ASD have a valid fetal scan and are feasible for further analysis in the case control design.

**Outlook:** The SMR00 and SMR04 are available to be added to the database in the nearest future and final results will be presented at the conference.

**Abstract 1736**

**DATABASE LINKAGE TO STUDY HEALTH SERVICES USE AND EDUCATIONAL OUTCOMES OF CHILDREN WITH BIRTH DEFECTS IN ARKANSAS**

**Authors:** James Robbins, A Goudie, W Nembhard

University of Arkansas for Medical Sciences

**Presenting Author:** James Robbins
Background: The Arkansas Reproductive Health Monitoring System (ARHMS) is an active birth defects surveillance program covering all births in the state of Arkansas. Using ARHMS as the foundation, this project will link important large-scale databases to enable studies of the proximal and longer term consequences of birth defects. This project represents a relatively new and exciting focus of birth defects research towards studies of outcome. Linking registry information to other existing data will allow studies of the financial burden of birth defects, health care use and disparities in access to care, contribution of comorbid conditions to health care use and survival, and longer term outcomes of birth defects including mortality, residual medical comorbidities, and achievement in school.

Method: To allow studies of public health questions, almost 14,000 cases of births from 2000 to 2011 in the ARHMS birth defects registry will first be joined to 28,000 unaffected control births drawn from Arkansas birth records. All-age mortality data will identify those who have died. The resulting file of ARHMS cases and unaffected controls will be linked in four separate processes to:

- **Arkansas Hospital Discharge Database for years 2000 through 2011** — Arkansas Health Department will link using birth certificate numbers as unique identifiers;
- **Arkansas Medicaid claims** — Birth certificate numbers will be used as unique identifiers for the 60% of the state’s births that are insured by Medicaid;
- **Arkansas standardized test data, school attendance data and special education referral data for children in grades 3 through 8** — Identifiers will be Social Security Number, date of birth, mother’s name and father’s name; and
- **Detailed clinical data for children receiving care at Arkansas Children’s Hospital**.

Finally, in a separate linkage process, 3,600 Arkansas cases and 1,200 Arkansas controls included in the National Birth Defects Prevention Study will be linked to ARHMS cases, Arkansas Hospital Discharge data, Arkansas Medicaid claims, and standardized test data.

Product: Linkage results at each step will document percentage linked, methods of handling unmatched records, measures taken to handle missing data, and personal identifiers used. A de-identified dataset will be compiled for studies of the post birth consequences of birth defects.

**RESEARCH COHORT EXTENSION**

Abstract 1348 USING DATA LINKAGE TO INVESTIGATE INCONSISTENT REPORTING OF SELF-HARM AND QUESTIONNAIRE NON-RESPONSE

Authors: Becky Mars¹, Jon Heron¹, Andrew Boyd², Rosie Cornish², Catherine Crane³, Keith Hawton³, Glyn Lewis⁴, John MacLeod¹, Kate Tilling¹, David Gunnell¹

¹School of Social and Community Medicine, University of Bristol, ²ALSPAC, University of Bristol, ³Department of Psychiatry, University of Oxford, ⁴Department of Psychiatry, UCL

Presenting Author: Rosie Cornish

Corresponding author email: Becky.Mars@bristol.ac.uk

Objectives: To examine agreement between self-reported and medically-recorded self-harm, and investigate whether the prevalence of self-harm differs in questionnaire responders vs. non-responders.

Methods: Data from 3027 participants from the Avon Longitudinal Study of Parents and Children (ALSPAC) birth cohort, 2363 of whom had responded to a self-harm questionnaire at age 16, were
linked to England’s Hospital Episode Statistics (HES) database. Rates of HES recorded self-harm were compared amongst those who completed (n=2,363) and did not complete (n=644) the self-harm questionnaire. We also compared HES recorded self-harm with participant self-reports.

**Results:** Fifty-four individuals (1.8%) had a self-harm event recorded in HES, with 41 (1.4%) recorded as having a hospital admission for self-harm. The prevalence of hospital admissions for self-harm recorded in HES was higher amongst those who did not complete the self-harm questionnaire than amongst those who did (self-harm hospital admissions: 2.0% in non-responders vs. 1.2% in responders, difference=0.8%, 95% CI -0.4% to 1.9%).

Fifteen self-harm events were recorded in HES prior to completion of the self-harm questionnaire, Three of which (20%) were not reported by participants (1/12 hospital admissions and 2/3 A&E only attendances).

**Discussion:** Our sample size is small, however, the results provide preliminary evidence to suggest that self-harm prevalence estimates derived from self-report may be underestimated. Future work will examine these issues in larger sample.

**Conclusions:** Our findings demonstrate the utility of combining self-reported self-harm data with data from medical records.

Abstract 1398 LARGE-SCALE COGNITIVE TESTING IN BRADFORD PRIMARY SCHOOLS: DATA VALIDATION, HARMONISATION, AND LINKAGE WITH LOCAL EDUCATION AUTHORITY (LEA) DATA.

Authors: Elizabeth Andrews, John Corkett

Bradford Institute for Health Research

Presenting Author: Elizabeth Andrews

Corresponding author email: elizabeth.andrews@bthft.nhs.uk

**Background:** Twenty percent of children entering school have underlying cognitive deficits and behavioural problems. Unidentified problems mean children may receive sub-optimal support. Further, socially disadvantaged children have a higher risk of cognitive deficits, making them more vulnerable and reinforcing societal inequalities. To reduce inequalities, we need to transform the psychological assessment of children. We examined the feasibility of large-scale cognitive testing in school settings.

**Method:** Data were collected across children aged 4-5 years in 77 primary schools within Bradford. The children were assessed on: (i) manual dexterity – assessed and data collected using a computer-based system built with LabVIEW®; (ii) literacy skills – ‘one-on-one’ testing with data collected using MS Access; and (iii) behavioural health – evaluated by teachers on paper-based questionnaires and back-filled into a bespoke web application.

**Results:** Data sets were available for 6841 children. Data was cleaned for the 3 different data sets separately and together. 6487 sets obtained for manual dexterity and literacy, and 2161 complete sets (all 3 assessments collected for Born in Bradford [BiB] only). We anticipated that children would be uniquely identified across data sets by their Unique Pupil Number (UPN). Data issues included errors in transcribing, free text errors and opportunities for entering data in the wrong field, duplications and missing data. The linkage to education data was done using a deterministic match based upon first name, last name, gender, date of birth and postcode. The matching rule was to have a single definite match on each of these attributes in the LEA database for them to provide us with the UPN. The matches provided by the LEA were verified by us. Issues included late submission of UPNs, out-of-date postcodes and mismatch through use of first name and preferred names in data sets.

**Discussion:** Tests across Bradford primary schools were successfully implemented establishing that ‘cognitive screening’ is feasible. Barriers to implementation were identified that need addressing before cognitive testing can become routine. The time required to develop robust systems for collecting only valid data and for linkage should not be underestimated. We make practical suggestions for implementation to help future users of this approach. Schools currently prioritise which students are...
seen by specialists but this arrangement can be haphazard. The research reported opens up the possibility of empowering schools to identify and provide appropriate support for all children with learning needs early in their school career.

Abstract 1404  INFLUENCE OF EDUCATIONAL ATTAINMENT ON THE ASSOCIATION BETWEEN SOCIAL CLASS AT BIRTH AND MULTIMORBIDITY IN MIDDLE AGE

Authors:  Marjorie Johnston\textsuperscript{1}, Corri Black\textsuperscript{1}, Gordon Prescott\textsuperscript{1}, Stewart Mercer\textsuperscript{2}, Michael Crilly\textsuperscript{1}

\textsuperscript{1}University of Aberdeen, \textsuperscript{2}University of Glasgow

Presenting Author:  Marjorie Johnston

Corresponding author email: marjorie.johnston@abdn.ac.uk

Introduction: Evidence suggests social class at birth is associated with disease development in later life and that this may be attenuated by educational attainment. No large, prospective cohort study with social class measured contemporaneously at birth has examined this for the outcome of multimorbidity (the occurrence of two or more health conditions in an individual). The aim is to examine the association between social class at birth with the presence of multimorbidity at middle age and assess the way this is modified by educational attainment.

Methods: This work uses the Aberdeen Children of the 1950s (ACONF) cohort (n=12,150, 48% female, born in Aberdeen 1950-56). The cohort has previously been linked to administrative data and an application is currently being processed to update this. Early life variables available include social class based upon the father’s occupation at birth of the participant (using the General Register Office’s Occupational classification).

In 2001-03, 7,183 (64%) responded to a postal questionnaire (PQ), providing data on factors such as educational attainment (highest qualification gained) and health status (asked to self-report up to six health conditions deemed important to them). Multimorbidity is classed as those reporting two or more conditions.

There was complete case analysis of PQ responders. Logistic regression assessed the relationship between social class and multimorbidity, with the reference group being higher social class (I/II). This analysis was then adjusted for educational attainment. The process was repeated with adjustment for age at PQ and gender.

Results: Of 7,184 individuals (age 44-52 years at PQ, 52% female) 389 (5.4%) reported multimorbidity. In the unadjusted analysis, the odds ratio (OR) for multimorbidity was 1.59 in group “III non-manual” (95\% confidence interval 0.96-2.63) rising to 2.43 (1.46-3.69) in group “V” and 2.77 (1.60-4.81) in the “unemployed/unknown” group.

When adjusted for educational attainment the statistically significant association between social class and multimorbidity was attenuated and did not reach statistical significance in group “III non-manual” (1.34 (0.80-2.23)) and group “V” (1.57 (0.96-2.57)). When repeated with age and gender, odds ratios were generally unaffected.

Conclusion: Lower social class at birth is associated with higher odds ratios of developing multimorbidity in middle age. This effect is attenuated by educational attainment in adulthood. Age at PQ and gender has little impact. Given the rising prevalence of multimorbidity as our population ages and the associated impact upon patients and services, these results have important implications for researchers, public health practitioners and policy makers.
DEFINING CASES OF ADOLESCENT ANXIETY AND DEPRESSION USING ELECTRONIC PRIMARY CARE DATA: AN EXTERNAL VALIDATION USING LINKAGE TO A COHORT STUDY

Authors: Rosie Cornish¹, Ann John², Andy Boyd³, John Macleod³
¹ALSPAC, University of Bristol; ²Farr Institute of Health Informatics Research, Swansea University; Public Health Wales NHS Trust, ³University of Bristol

Presenting Author: Rosie Cornish

Corresponding author email: rosie.cornish@bristol.ac.uk

Background: Rates of anxiety and depression among children and adolescents are increasing and represent an important public health problem. Further, depression during adolescence is associated with risk-taking behaviours such as substance misuse and with poorer mental health during adulthood. The ability to correctly identify cases of anxiety and depression using routinely collected primary care data would facilitate research in this field.

Methods: The Avon Longitudinal Study of Parents and Children (ALSPAC) is a birth cohort. Pregnant women were recruited in the early 1990s and the offspring have been followed up since birth. Depression was measured when the children were aged 17-18 years using the CIS-R³, a widely validated tool for identifying individuals with depression.

Practices using software systems supplied by EMIS and Apollo were asked to provide their assent for the extraction of the electronic patient records of ALSPAC participants. The extracted records were anonymised and securely transferred into infrastructure developed by the Secure Anonymised Information Linkage (SAIL) project at Swansea University. Within this infrastructure de-identified ALSPAC data were linked to these GP records.

The data analysis has just started. We will use Read codes corresponding to diagnoses, symptoms and treatment for anxiety and depression to create multiple case definitions and identify individuals meeting these. We will then compare GP-recorded cases to ALSPAC-recorded cases and calculate the sensitivity, specificity and predictive values for each definition, using the CIS-R as the gold standard.

Results: GP data are available for 2146 singletons. Of these, 364 individuals have been excluded because they left their registered GP practice from which we extracted data before the age of 20. Among the remaining 1782, 1352 (76%) had depression measured in ALSPAC using the CIS-R. Of these 1352, 100 (7.4%) met the criteria for a diagnosis of depression using the CIS-R, of whom 20 (20%) had a GP-recorded diagnosis of depression or anxiety between the ages of 16 and 19 (inclusive). A further 15 individuals had received antidepressants, anxiolytics or hypnotics and an additional 5 had no diagnosis or treatment but GP-recorded symptoms of anxiety or depression between these ages.

These are preliminary findings. Further findings in relation to the different case definitions will be presented.

Conclusions: Recommendations will be made regarding the optimal choice of case definition when carrying out a research study on adolescent anxiety and depression using electronic primary care patient records.
Abstract 1624  
THE CREATION OF A NEW 1936 BIRTH COHORT IN SCOTLAND

Authors:  
Zengyi Huang1, Graham Kirby2, Ian Deary3, Frank Popham4, Roxanne Connelly2,  
Chris Dibben3  
1LSCS, University of Edinburgh, 2University of St Andrews, 3University of Edinburgh, 4University of Glasgow

Presenting Author: Zengyi Huang  
Corresponding author email: zengyi.huang@ed.ac.uk

This paper provides an overview of the methods used to create an administrative data based 1936 Birth Cohort, a collaboration between the National Records of Scotland (NRS), the University of Edinburgh Centre for Cognitive Aging and Cognitive Epidemiology (CCACE) and the Scottish Longitudinal Study (SLS) – Development and Support Unit. The cohort is structured around an existing study, the SLS, a 5% sample, based on 20 pseudo random birthdates, of the Scottish population followed between 1991 to the present day. We took the SLS birth date sample from the Scottish Mental Survey of 1947 (a cognitive ability test almost all Scottish children aged 11 sat in 1947) and linked it to the 1939 National Register, the National Health Service Central Register (NHSCR) and the SLS. The outcome is a powerful life-course dataset containing information from childhood to old age.

In this paper we will discuss and critically evaluate the methods used to create the study. Automatic and manual linkage and transcription of the 1939 registration data were undertaken by the clerical support team from the NHSCR. We will present the linkage results and discuss the strengths and weaknesses of the methods. We will outline the results of the machine learning techniques to automatically code occupation descriptions to standard frames (HISCO). And finally we will discuss methods used for automatically geocoding the 1939 addresses to street centroids. We will overall assess the challenges and benefits of retrospectively creating new birth cohort studies from routine and administrative data.

Abstract 1692  
LINKING THE SCOTTISH MENTAL SURVEY 1947 WITH HEALTH OUTCOMES

Authors:  
Iva Cukic1, C M Calvin1, C E Brett1, G D Batty2, I J Deary1  
1Centre for Cognitive Ageing and Cognitive Epidemiology, University of Edinburgh, 2Department of Epidemiology & Public Health, University College London

Presenting Author: Iva Cukic  
Corresponding author email: icukic@exseed.ed.ac.uk

This talk will describe results of linkage between data from youth and health outcomes of the Scottish Mental Survey 1947 (SMS1947, N=70805), especially its two representative subsamples: the 6-Day Sample and 36-Day Sample. All children born in 1936 and attending Scottish schools on June 4th 1947 took the same intelligence test. All children born on the first three days of each month of 1936 (36-Day Sample) were selected to complete, in addition, a “Random Sample Sociological Schedule”. This group was representative of the SMS1947 in terms of sex, geographical location, cognitive ability, and social background. Furthermore, children born on the 1st day of even-numbered months (6-Day Sample, N=1208) had the Schedule and a Stanford-Binet IQ test completed and were, in addition, subsequently followed up for 16 years to age 27. Where analyses are done on the 36-Day Sample excluding the 6-Day Sample we refer to this as the 30-Day Sample (N=5083).

The 30-Day Sample members were traced through the NHS Central Register. Childhood data were linked with mortality records, hospital admissions data (1980-2014 in Scotland, 1997-2013 in England/Wales), and cancer registrations (1980-2014 Scotland, 1984-2013 England/Wales). We explored the associations between childhood cognitive ability and later life morbidity and mortality. Lower
childhood cognitive ability was associated with higher risk of later life cardiovascular events such as fatal myocardial infarction and heart failure, and both fatal and non-fatal coronary heart disease. Furthermore, lower cognitive ability at age 11 was associated with higher risk of developing lung cancer later in life, and higher risk of death from all cancers combined.

During the 16-year follow-up of the 6-Day Sample, cognitive ability of the sample members’ younger siblings (N = 1554) was assessed at age 11 using the same revision of the Stanford-Binet cognitive test. We will use these data to investigate whether differences in childhood cognitive ability between siblings predict differences in later-life health outcomes.

The same linkage procedures as described for the 30-Day Sample are currently in progress for the entire SMS1947, i.e. for most of the 1936-born population who were in Scottish schools at about age 11. This will be the first time that linkage between childhood cognition and health outcomes has been attempted on almost an entire population.

Abstract 1776

ELECTRONIC FOLLOW UP OF A DOUBLE-BLIND RANDOMISED CONTROLLED TRIAL: PROBIOTIC SUPPLEMENTATION IN THE PREVENTION OF CHILDHOOD ATOPY

Authors: Gareth Davies1, M Gravenor1, C Brooks1, Daniel Thayer1, M Storey1, C Li2, G Morgan2, S Allen2, I Garaiova3, S Plummer3, S Jordan2

1The College of Medicine, Swansea University, Singleton Park, Swansea, Wales, UK, 2Department of Nursing, The College of Human and Health Sciences, Swansea University, Singleton Park, Swansea, Wales, UK, 3Research Department, Cultech Limited, Baglan Industrial Park, Port Talbot, West Glamorgan, Wales, UK

Presenting Author: Gareth Davies

Corresponding author email: g.i.davies@swansea.ac.uk

Background: The validity of clinical trials may be jeopardized by “loss to follow up”. One strategy to ameliorate this problem is electronic follow up using healthcare databases housing routinely collected data. However this method of follow-up has not, to our knowledge, been formally compared with traditional fieldwork.

Objectives: To explore the feasibility of electronic follow up of participants in a clinical trial of probiotics and atopy in young children.

To compare trial field data with electronic data routinely collected by the health services, for diagnoses of asthma at 2 years.

To complete 5 year follow up for asthma-related outcomes.

Participants and setting: 454 women, delivering 452 infants were recruited to a trial of a probiotic food supplements in antenatal clinics in SW Wales, 2005-2007. Women were asked to take probiotic supplements or placebo before birth and administer these to their infants in their first 6 months. Children were followed up by traditional fieldwork contacts for 2 years and on electronic medical records for 5 years.

Methods: The project utilises routinely collected, anonymised, person based electronic data collected by the health service submitted to the Secure Anonymised Information Linkage (SAIL). The data were examined for National Health Service (NHS) Read codes relating to eczema, asthma, and related medicines, including beta 2 agonists, oral corticosteroids and leukotriene receptor antagonists and antimicrobials. Use of asthma-related medicines was classified according to BTS guidelines.

Comparison of field and electronic data was undertaken using Cohen’s Kappa. 5 year outcomes were assessed using binary logistic regression and Kaplan Meier survival analysis.

Results: 435 of 452 (96%) recruits’ records were accessible in the electronic SAIL data, 422 (93%) with “good” coverage at 2 years and 370 (82%) at 5 years. Attrition was not associated with socio-economic status. Amongst those with electronic records, at 2 years, there was moderate agreement between the 2 methods of outcome ascertainment. There was no significant difference between trial arms in use of medicines prescribed for asthma.
**Conclusions:** Using SAIL, electronic follow up of trial participants was feasible, efficient and more comprehensive than traditional fieldwork. There is moderate agreement between primary care records and parents’ reporting asthma. Administration of a probiotic supplement in infancy does not appear to reduce the prescription of asthma treatment at age 5 years.

**TRIAL’S METHODOLOGY AND INNOVATION**

**Abstract 1513**

**STATIN WEB-BASED INVESTIGATION OF SIDE EFFECTS (STATIN WISE)**

**Authors:** Emily Herrett¹, E Williamson¹, T van Staa², L Smeeth¹

¹London School of Hygiene and Tropical Medicine, ²University of Manchester

**Presenting Author:** Emily Herrett

**Corresponding author email:** emily.herrett@lshtm.ac.uk

**Background:** Statins reduce cardiovascular disease risk and are recommended as part of the strategy for primary and secondary prevention in the UK. There has been widespread reporting of statin-related symptoms in the media, such as muscle pain and weakness. These reports have been prompted by data from non-randomized, non-blinded observational studies but have not been confirmed by randomised controlled trials. Accurate data on the cause of symptoms experienced during statin use are needed to inform patients’ and doctors’ treatment choices. The proposed study will address this important uncertainty about statin therapy for patients and the population.

**Aims:** To determine whether muscle adverse events attributed to statin use by patients are caused by statins.

**Methods and analysis:** This series of 200 N-of-1 randomised, double blind, placebo controlled trials will take place in UK primary care, among patients who have recently stopped using statins or are considering stopping statin use due to perceived side effects. Participants will receive six two-month treatment blocks of either atorvastatin 20mg or placebo, with the order of treatments randomly allocated. The primary outcome will be muscle symptoms (pain, weakness, tenderness, stiffness or cramp), measured daily during the last 7 days of each treatment period using a visual analogue scale on a bespoke mobile application. Secondary outcomes will be collected using a questionnaire at the end of each treatment period and will explore symptoms in more detail, including frequency, intensity and possible causes.

At the end of each participant’s 12 month follow-up, they will be shown summaries of their individual unblinded results, with estimates of the effect of statins on their own symptoms. This will help them to decide whether their symptoms are caused by statins, and whether to continue taking statins. Data from 200 participants will be aggregated to form a powerful dataset and analysed to assess, at the population level, the extent to which the adverse events attributed to statins appear to be causally related.

**Ethics and dissemination:** The trial will be approved by the National Research Ethics Committee. The results of the combined analysis will be published in peer reviewed medical journals and presented at national and international conferences. The results of this research will be disseminated widely in order to influence practice. With the support of our Patient and Public Involvement Panel, we will determine the best methods for dissemination to participants, general practices and national policy makers.
TRANSFoRm CONNECTIVITY INFRASTRUCTURE TO EMBED eCRFs INTO EHR SYSTEMS

Authors: Mark M McGilchrist¹, Vasa Curcin², Jean-Francois Ethier³, Theodoros Arvanitis⁴, Brendan Delaney⁵
¹University of Dundee, UK, ²King’s College London, ³Institut national de la santé et de la recherche médicale (INSERM), France, ⁴University of Warwick, UK, ⁵King’s College London, UK

Presenting Author: Mark M McGilchrist

Corresponding author email: m.m.mcgilchrist@dundee.ac.uk

As a comprehensive digital infrastructure to support the Learning Healthcare System (LHS), TRANSFoRm advances the integration between clinical research and routine healthcare. Following a semantic mediation approach for data integration, the software implements a service oriented architecture that is being evaluated with commercial partners in a full RCT.

Introduction and Background: Embedding clinical trial tasks within daily healthcare workflow and interoperating with local IT infrastructure is essential to improving the quality, throughput and affordability of clinical trials. The TRANSFoRm project has developed two core capabilities to enable study workflow and semantic interoperability of data. Using a platform ontology, a generic model for EHR exchange documents, and a mapping between the two, a generic query model grounded in openEHR archetypes can specify data access. Extensions to CDISC SDM/ODM standard specify data preloads for forms. Workflow and associated operations and messages are driven by local software components enhancing security. This architecture is evaluated in an EU-wide clinical trial on the efficiency of different treatment modes for GERD, using five different EHR systems at 40 practices.

Methods: TRANSFoRm components, implemented as RESTful Web services, work together to deliver: patient eligibility checks and enrolment, part-filling of eCRF data from EHRs, mobile/web data collection from patients, and storing a copy of study data in the EHR. A study system (TSS) coordinates study events and data collections using HTML form templates with bound queries for data pre-load. A Data Node Connector (DNC) manages the interaction between the TSS and EHR, including form requests, translations for presentation (SDM to HTML), data pre-loads using a semantic mediator, and storage in TSS and EHR in standard formats (ODM). The message protocol for this interaction is now finalized and a comparative analysis with IHE standards is ongoing.

Results: TRANSFoRm federated infrastructure for eCRFs has been GCP-validated and is currently being evaluated in the GERD randomized controlled trial, due to finish in October 2015. We shall present initial results alongside the technical, business and clinical aspects of conducting the study.

Discussion: Integration of clinical research studies into routine clinical practice has potential to bring a fundamental shift to the conduct of medical research. Facilitating subject identification and recruitment alone would significantly reduce the cost and complexity of running trials. Pre-loading EHR data into eCRF forms and coordinating the workflow for both study and EHR operations reduces effort for clinicians and decreases the likelihood of error.
Abstract 1671

STATISTICAL ISSUES IN DESIGNING SERIES OF N OF 1 TRIALS

Authors: Elizabeth Williamson, Emily Herrett, Liam Smeeth
London School of Hygiene & Tropical Medicine

Presenting Author: Elizabeth Williamson

Corresponding author email: elizabeth.williamson@lshtm.ac.uk

Ascertaining how many patients to recruit for a randomised trial is a common goal, but far from easy to achieve. Power calculations depend on untestable assumptions and require specification of quantities that are unknown prior to running the trial. These problems are greatly exacerbated in the context of a series N of 1 trials.

An N of 1 trial involves a single patient who is randomised to a sequence of the treatments to be compared, with the order of treatments randomly allocated. The aim of an N of 1 trial is to provide individualised information about the treatment effect in order to inform the patient’s subsequent treatment decisions. An attractive feature of N of 1 trials is that they can be embedded within the patient’s standard primary care, making this an ideal tool to use alongside more traditional prescribing practices. For the purposes of estimating the treatment effect at the population level, a series of N of 1 trials can also be combined to form a multiple crossover trial.

Designing a series of N of 1 trials is complicated by this two-fold inferential aim: we wish to estimate treatment effects both in individuals and in the population. Statistical power for both individual and population level inference can be achieved by adding more treatment periods; this, however, risks increasing non-adherence to the allocated treatment and participant dropout, thereby greatly complicating the population level analysis. A second complication arises due to the lack of closed-form power calculations for N of 1 trials. Power calculations must typically be performed through computationally-intensive simulation-based approaches. Finally, N of 1 trials typically use patient reported outcomes, which often follow distributions that cannot be well approximated by the standard normal-based models.

This talk will illustrate these statistical problems in the design of a series of N of 1 trials using a trial investigating the effects of statins on perceived side-effects including muscle. We will present some results showing how the power of a series of N of 1 trials is affected by the choice of number of periods versus the number of participants, the patterns of participant dropout, and the assumptions made regarding the variability of the outcome measure. Finally, we will discuss the goal of the individual-level inferences, and whether traditional hypothesis testing provides the most meaningful summary of results in this context.

Abstract 1690

USING ELECTRONIC ROUTINE LINKED HEALTH RECORDS IN ASSESSING INTERVENTIONS IN EMERGENCY CARE

Authors: Helen A Snooks, Sarah A Gaze, Ronan A Lyons, Ian T Russell, Alan J Watkins
College of Medicine, Swansea University

Presenting Author: Helen A Snooks

Corresponding author email: h.a.snooks@swansea.ac.uk

Background: SAFER2 was a pragmatic cluster randomised controlled trial that used a hierarchy of outcomes comprising death, unscheduled hospital admission, emergency department attendance, and emergency ambulance calls to evaluate new protocols which allow emergency ambulance paramedics treating falls in older people to assess and refer patients to appropriate community based care.

Objectives: Participants in England were asked to consent to follow up through both routine and
identifiable medical records. This presentation considers the consistency between data on hierarchy components available through these two routes. We also consider the proportion of cases with analysable outcomes through the two routes, and the characteristics of those participants providing data through two routes.

Methods: Routine medical records are available anonymously through the Secure Anonymised Information Linkage Databank in Wales and the Health & Social Care Information Centre in England. Identifiable data on consenting participants in England was provided by individual care centres. Following successful linkage with study participants, data on each component in the hierarchy of outcomes is assessed at one month (for safety) and six months (for effectiveness).

Results: Matched routine records, available in approximately 99% of cases, yielded analysable outcomes on 80% of all participants; in contrast, outcomes based on identifiable data sources were available on less than 40% of consenting participants in England. Participants consenting to identifiable follow up reflected overall gender and intervention group proportions, but were generally younger & healthier (in terms of the hierarchy of outcomes at both one & six months). Hierarchy component outcomes from the two data routes are broadly, but not invariably, consistent.

Conclusions: Linkage problems are diminishing as experience develops. Routinely recorded data provides more complete information on a greater proportion of study participants over multiple timepoints in follow-up, and allows assessment of the intervention on increased sample sizes. Our experience and findings on assessing the SAFER2 intervention with data from different routes will provide useful guidance for other studies using routine data.

Abstract 1765  PATIENT PHENOTYPING FROM ELECTRONIC HEALTH RECORDS DATA FOR CLINICAL TRIALS PROTOCOL FEASIBILITY: CASE STUDY IN BREAST CANCER

Authors: Dionisio Acosta¹, Paul Taylor¹, Mohanned Keshtgar²
¹CHIME, Institute of Health Informatics, University College London. ²Department of Surgery, Royal Free Hospital, London

Presenting Author: Dionisio Acosta

Corresponding author email: d.acosta@ucl.ac.uk

The IMI Electronic Health Records for Clinical Research (EHR4CR) project aims at the development of a standard based technological platform for the systematic and effective reuse of Electronic Health Records (EHR) data across Europe, in a trustworthy, adaptable and scalable fashion, supporting the life cycle of clinical trials. The first stage of this project is the Protocol Feasibility Scenario (PFS), an iterative process whose objective is to refine the design of the trial protocol and ascertain the existence or a population sufficiently large for the study. Currently, PFS is an expensive and imprecise process and as such is seem as a perfect candidate for substantial improvements by the use of this technology. This work presents a case study in the implementation and deployment of the EHR4CR PFS for a Breast Cancer dataset at the Royal Free Hospital in London. Based on this experience, we discuss a range of issues from dataset definition criteria and the impact of controlled terminologies, to clinical data transformation, loss of information and its impact in output patient counts. After obtaining Ethics, Data Protection and Information Governance approvals, a Breast Cancer dataset was curated and referenced against international controlled terminologies by using BioPortal. An Extraction, Transformation and Loading (ETL) process was implemented, using a functional approach, to transform the dataset existing clinical modelling into the EHR4CR Clinical Information Model (CIM). Patient anonymity was ensured by following masking strategies and anonymisation algorithms were also evaluated (k-anonymity, l-diversity). We analysed the information loss in each stage of the process, both clinical and temporal aspects, and the effect in patient counts in a set of clinical trials. The main result of the process was a Reference Breast Cancer Data Definition, referenced against controlled terminologies, that can serve as the basis for a standard in this clinical area. We found that the platform can be successfully deployed and provides patients counts in a fraction of time and costs compared to the standard process. We also established that the transformation between the EHR clinical model...
and EHCR CIM implied a loss of clinical and temporal information, and even in some cases, the inability to represent certain clinical concepts, depending on design and controlled terminology choices. It is important to note that patient privacy assurances can be provided by observing both technological and anonymisation strategies. Finally, we propose a series of future research directions to overcome some of the limitations detected.

Abstract 1770 USING ROUTINELY COLLECTED SECONDARY CARE DATASETS FOR MULTI-CENTRE CLINICAL TRIALS

Authors: Sharon Kean, R Wilson, I Ford
University of Glasgow

Presenting Author: Robbie Wilson

Corresponding author email: sharon.kean@glasgow.ac.uk

Background: Using primary and secondary electronic health records within a clinical trial context is becoming widely accepted practice that can streamline some trial processes and may reduce costs. The potential for using these records to inform trial design during the feasibility stage, to facilitate recruitment by identifying patients, to supplement trial data such as adverse events and prescribing information, and to follow-up patients long-term, if utilised properly, can give accurate trial results. The increased use of electronic data capture to facilitate all stages of clinical trial management, in combination with access to healthcare datasets, can provide a comprehensive clinical trial data platform.

Methods: Centralised secondary care datasets, including hospitalisation and deaths, were extracted for trial participants to identify or supplement trial reported adverse events. The system developed combined these secondary care data with trial visit data which had been captured via an eCRF. The combined datasets were reviewed to ensure no double-reporting and a clinical review was undertaken to assign causality and expectedness for those events where only minimal information was available. All events, whether self-reported or from healthcare records, were then used to flag possible endpoints to the trial adjudication committees via a web portal.

Conclusion: Resolving information governance issues and gaining approvals to use healthcare data can be a rate-limiting step. Developing clinical trial data management systems to combine healthcare data with trial data to report within regulatory timelines is not without challenge. Understanding the complexity and limitations of these data is crucial. We will illustrate the approaches used, the systems that were developed and the challenges overcome.