

## Electronic Medical Data

David V Ford  
Director, Health Informatics,  
College of Medicine,  
Swansea University, UK

College of Medicine  
Coleg Meddygaeth

Swansea University  
Prifysgol Abertawe

## What is Electronic Medical data?

- Many names, many differences
  - Electronic Medical Records / Electronic Health Records / Personal Health Records / Individual Health Records etc.
- An electronic medical record (EMR) is a [computerized medical record](#) created in an organization that delivers care, such as a hospital or doctor's office

College of Medicine  
Coleg Meddygaeth

Swansea University  
Prifysgol Abertawe

## What is Electronic Medical data (2)?

- Content can vary but can include:
- Diagnoses, procedures, medications, other therapeutics, radiology images and reports, labs orders and results, nursing and care data,
- Content usually structured using a coding or classification system, but operational EMR's often have significant free text data too (e.g. histopath and radiology reports)

College of Medicine  
Coleg Meddygaeth

Swansea University  
Prifysgol Abertawe

## What are the uses for EMR data?

- Observational studies
  - Post marketing surveillance
  - Burden of disease
  - Incidence and prevalence estimation
  - Natural history of disease / epidemiology
  - Pragmatic outcome studies
  - Vigilance studies
- Interventional Studies
  - Trial site identification
  - Power calculation
  - Electronic inclusion / exclusion criteria
  - Support patient recruitment
  - Long term e-follow up of trial participants

College of Medicine  
Coleg Meddygaeth

Swansea University  
Prifysgol Abertawe

## Benefits

- Data created and validated by clinicians for clinical use
- No (research) biases in the data collection
- Costs of data collection already covered
- Data already present – no collection lag
- Co-morbidities clearer and more precise (in full EMRs)
- Participant identification easier
- Long term (electronic) follow-up straight forward
- Large longitudinal populations
- Excellent for observational studies. Can also be good for trial site identification, accrual estimation and participant consenting.

College of Medicine  
Coleg Meddygaeth

Swansea University  
Prifysgol Abertawe

## Limitations

- Variety of different coding and classification systems in use, free text very hard to analyse at volume
- Data quality varies according to clinician / organisation / country. Demographics can be a problem
- Data can have been subject to "game playing"
- Combining datasets is difficult – mapping one data to another without loss is often impossible
- Analysing troublesome data is highly skilled – expertise and local knowledge is needed and it takes time.
- Often multiple EMRs for the same patient – problems of combining results and establishing truth. Can be a **very** substantial problem.
- Data protection / privacy laws are variable by jurisdiction

College of Medicine  
Coleg Meddygaeth

Swansea University  
Prifysgol Abertawe

## Case study: The SAIL Databank

- Secure Anonymous Information Linkage (SAIL) Databank
  - 3 million people of Wales (UK) - total population cohort
  - Data linkage system – all records linked to 'known' anonymous records, with address tracking
  - Data supplied to SAIL every month. Goes back 20 years (average)
  - Built on Blue C high performance supercomputing (IBM P7)
  - All inpatients, outpatients, day cases – every hospital in Wales
  - General practice data 9 c. 200 practices (1.6 million people)
  - Over 200 data sources including path, some rad, child health database, ambulances, school achievement, NHS Direct call centre, social care, congenital abnormalities, cancer registry, etc

College of Medicine  
Coleg Meddygaeth

Swansea University  
Prifysgol Abertawe

## Case study: The SAIL Databank (2)

- Very carefully designed probabilistic matching process (to the Welsh Demographics Service)
- Strong privacy protection architecture
  - Split file anonymisation – no confidential identifiable data flows to anyone
  - Secure transportation and storage
  - Access only via SAIL Gateway – remote desktop system – full toolset, freedom to work, but . . .
  - Data never leaves the Gateway – only result sets
  - Access via database views, as agreed in project spec.
  - All projects reviewed by the SAIL Information Governance Review Panel
  - All users registered, with strong user agreements

College of Medicine  
Coleg Meddygaeth

Swansea University  
Prifysgol Abertawe

## Case study: The SAIL Databank (3)

- Excellent for all observational studies
- Good for trial site identification / participant quantification
- Difficult to contact real patients (SAIL must maintain anonymous status)
- 30 + studies currently underway
- £35million in research income
- Team of 40+ information scientists working on SAIL
- Difficult data, not just pressing F7 and the answer comes out

College of Medicine  
Coleg Meddygaeth

Swansea University  
Prifysgol Abertawe

## Methodology references - Architecture

### The SAIL Databank: building a national architecture for e-health research and evaluation

David V Ford<sup>1</sup>, Kerina H Jones<sup>1</sup>, Jean-Philippe Verplancke<sup>1</sup>, Roman A Lyons<sup>1</sup>, Gareth John<sup>2</sup>, Ginevra Brown<sup>1</sup>, Caroline J Brooks<sup>1</sup>, Simon Thompson<sup>1</sup>, Owen Bodger<sup>1</sup>, Tony Couch<sup>2</sup> and Ken Leake<sup>2</sup>

#### Abstract

**Background:** Vast quantities of electronic data are collected about patients and service users as they pass through health service and other public sector organisations, and these data present enormous potential for research and policy evaluation. The Health Information Research Unit (HIRU) aims to realise the potential of electronically-held, person-based, routinely-collected data to conduct and support health-related studies. However, there are considerable challenges that must be addressed before such data can be used for these purposes, to ensure compliance with the legislation and guidelines generally known as Information Governance.

**Methods:** A set of objectives was identified to address the challenges and establish the Secure Anonymised Information Linkage (SAIL) system in accordance with Information Governance. These were to: 1) ensure data transportation is secure; 2) operate a reliable record matching technique to enable accurate record linkage across datasets; 3) anonymise and encrypt the data to prevent re-identification of individuals; 4) apply measures to address disclosure risk in data views created for researchers; 5) ensure data access is controlled and authorised; 6) establish methods for scrutinising proposals for data utilisation and approving output; and 7) gain external verification of compliance with Information Governance.

**Results:** The SAIL databank has been established and it operates on a DB2 platform (Data Warehouse Edition on ADX) running on an IBM P series Supercomputer: Blue-C. The findings of an independent internal audit were favourable and concluded that the systems in place provide adequate assurance of compliance with Information Governance. This expanding databank already holds over 500 million anonymised and encrypted individual-level records from a range of sources relevant to health and well-being. This includes national datasets covering the whole of Wales (approximately 3 million population) and local provider-level datasets, with further growth in progress. The utility of the databank is demonstrated by increasing engagement in high quality research studies.

**Conclusion:** Through the pragmatic approach that has been adopted, we have been able to address the key challenges in establishing a national databank of anonymised person-based records, so that the data are available for research and evaluation whilst meeting the requirements of Information Governance.

14

## Methodology references - Matching

### The SAIL databank: linking multiple health and social care datasets

Roman A Lyons<sup>1</sup>, Kerina H Jones<sup>1</sup>, Gareth John<sup>2</sup>, Caroline J Brooks<sup>1</sup>, Jean-Philippe Verplancke<sup>1</sup>, David V Ford<sup>1</sup>, Ginevra Brown<sup>1</sup> and Ken Leake<sup>2</sup>

#### Abstract

**Background:** Vast amounts of data are collected about patients and service users in the course of health and social care service delivery. Electronic data systems for patient records have the potential to revolutionise service delivery and research. But in order to achieve this, it is essential that the ability to link the data at the individual record level be retained whilst adhering to the principles of information governance. The SAIL (Secure Anonymised Information Linkage) databank has been established using disparate datasets, and over 500 million records from multiple health and social care service providers have been loaded to date, with further growth in progress.

**Methods:** Having established the infrastructure of the databank, the aim of this work was to develop and implement an accurate matching process to enable the assignment of a unique Anonymous Linking Field (ALF) to person-based records to make the databank ready for record-linkage research studies. An SQL-based matching algorithm (MACRAL, Matching Algorithm for Consistent Results in Anonymised Linkage) was developed for this purpose. Firstly the suitability of using a valid NHS number as the basis of a unique identifier was assessed using MACRAL. Secondly, MACRAL was applied in turn to match primary care, secondary care and social services datasets to the NHS Administrative Register (NHSAR), to assess the efficacy of this process, and the optimum matching technique.

**Results:** The validation of using the NHS number yielded specificity values > 99.8% and sensitivity values > 94.6% using probabilistic record linkage (PRL) at the 50% threshold, and error rates were < 0.2%. A range of techniques for matching datasets to the NHSAR were applied and the optimum technique resulted in sensitivity values of 99.9% for a GP dataset from primary care, 99.3% for a PEDW dataset from secondary care and 95.2% for the PARIS database from social care.

**Conclusion:** With the infrastructure that has been put in place, the reliable matching process that has been developed enables an ALF to be consistently allocated to records in the databank. The SAIL databank represents a research-ready platform for record-linkage studies.

11

## Thanks

For more information,  
please visit:

[Hiru.swansea.ac.uk](http://Hiru.swansea.ac.uk)

(Note: No 'www!')



College of Medicine  
Coleg Meddygaeth

Swansea University  
Prifysgol Abertawe