Health Information Research Unit

Secure Anonymised Information Linkage (SAIL) system

SHIP International Workshop

February 2010

Professor Ronan Lyons

chiral Centre for Health
- information
- research
- evaluation

School of Medicine
Swansea University

SAIL

- A system which links anonymised data at individual and household level across many health and health related datasets
- Uses split files, TTP, and encryption to preserve anonymity
- Utilises high performance computing infrastructure (EU/IBM)

- Created by the national e-health research facility for Wales (HIRU)
- Funded by NISCHR, Swansea University School of Medicine, various grants
- Part of the UK e-health research infrastructure

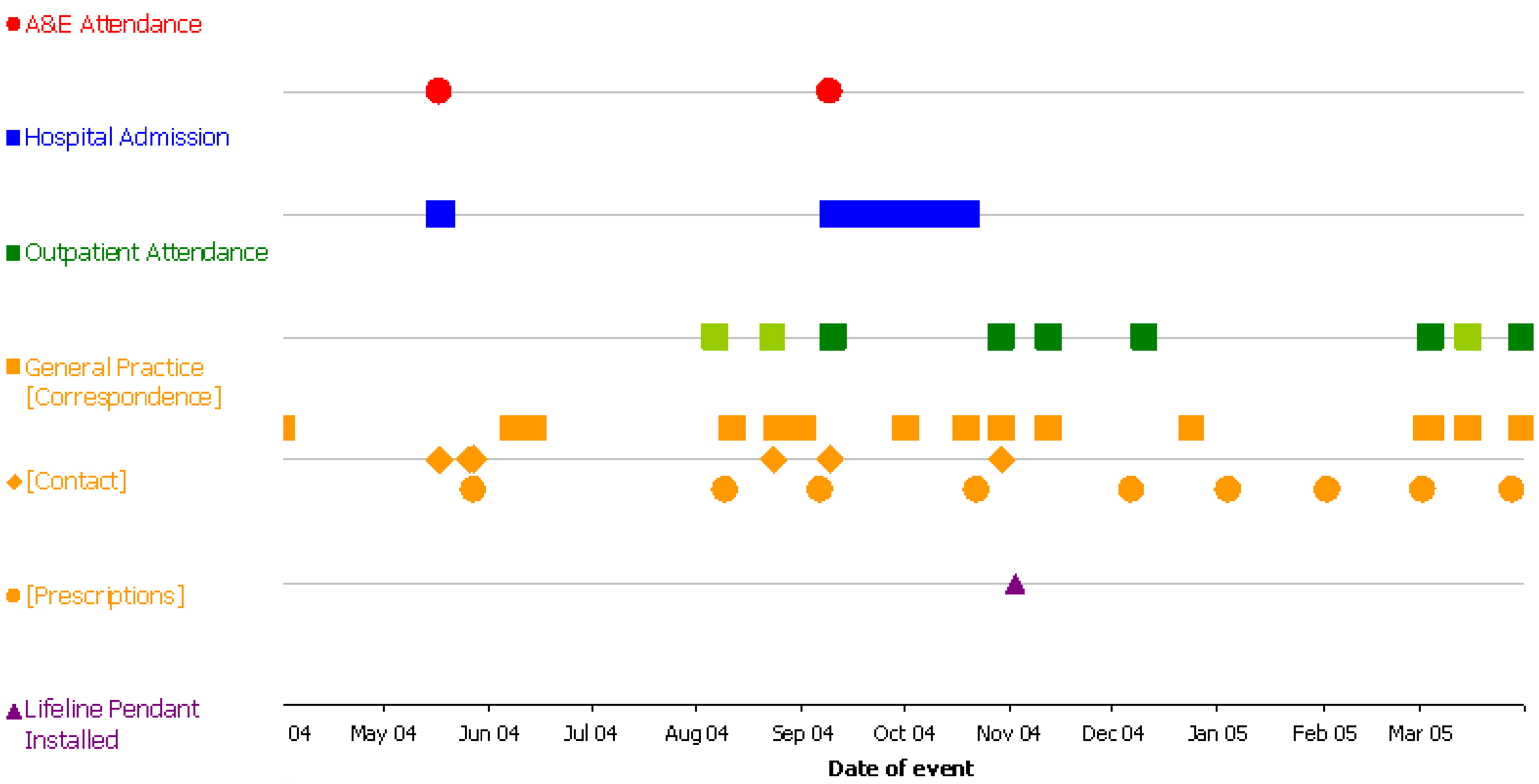
Many routine datasets - examples

- Inpatients, outpatients & day cases
- Child health (heights, weights, immunization records, maternal info etc)
- Births & Deaths
- Cancer incidence (Cancer Registry for Wales - WCISU)
- National screening programs (Breast Test Wales, Cervical Screening + Bowel screening and newborn hearing)
- Education data – National Pupil Database (<18), other ages groups in discussion
- Congenital Abnormalities
- NHS Direct Wales 0845 call centre contacts
- Ambulance Service data (Dispatch & Patient Clinical Record)
- A&E data from Trusts
- GP – full historic extracts (all patients, all conditions) **168** practices at present – more to come
- Pathology results from NHS Trusts (all departments)
- Social Services unified assessment data (older people, mental health, learning disability, children)
- Housing data from Local Authorities (to characterize R-ALFs)
- Swansea pathology tissue bank + biomedical datasets
- Cardiology Images

And more…!

Patient Journey Analysis- Health and Social Care

One Patient's Journey Across all the HIRU datasets, 2004/05



Building the SAIL system

It really is not just about data and databases!

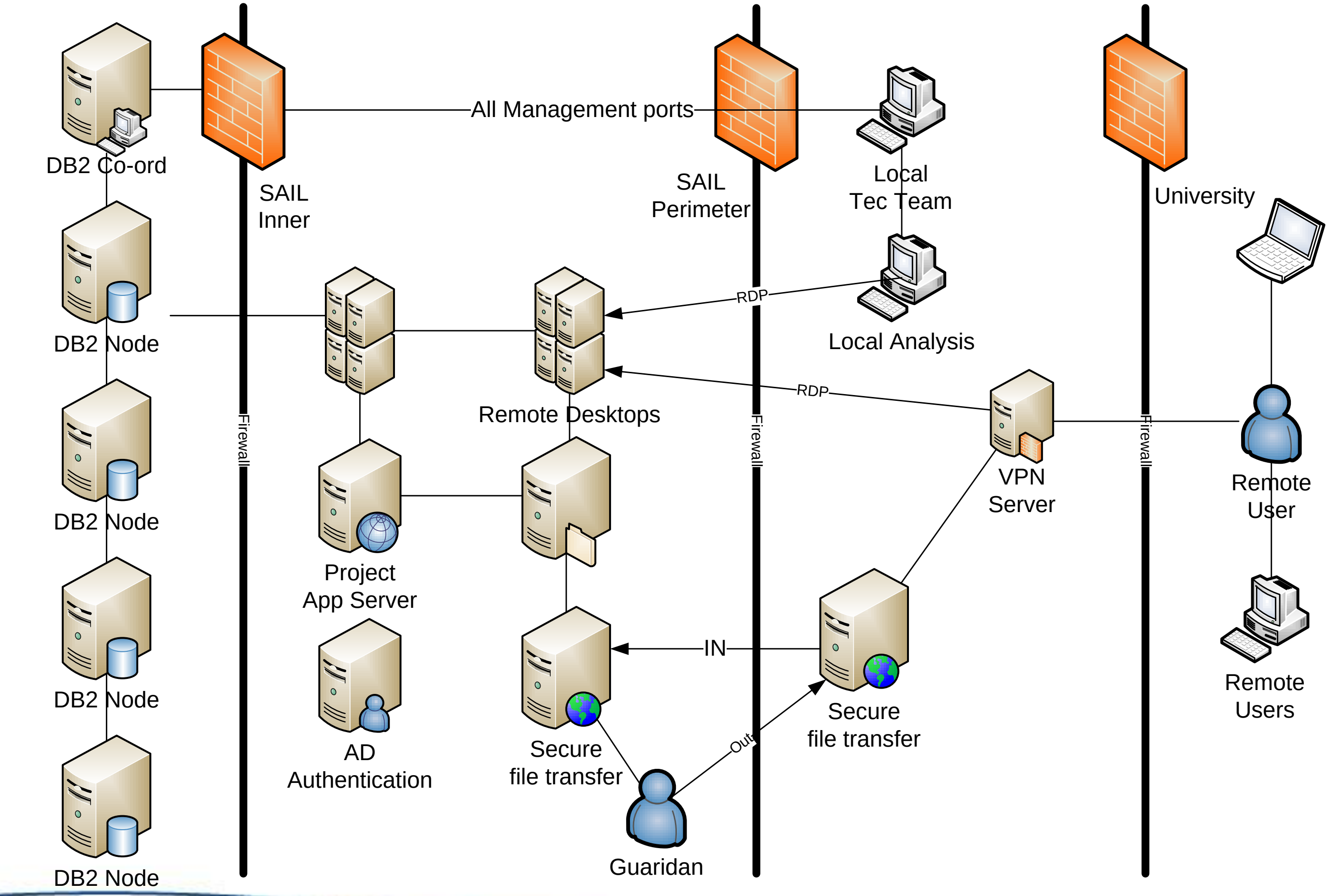
1) Secure data transportation
2) Reliable matching process
3) Anonymisation and encryption
4) Disclosure control
5) Data access controls
6) Scrutiny of data utilisation proposals
7) External verification of compliance with IG

Opening access to SAIL

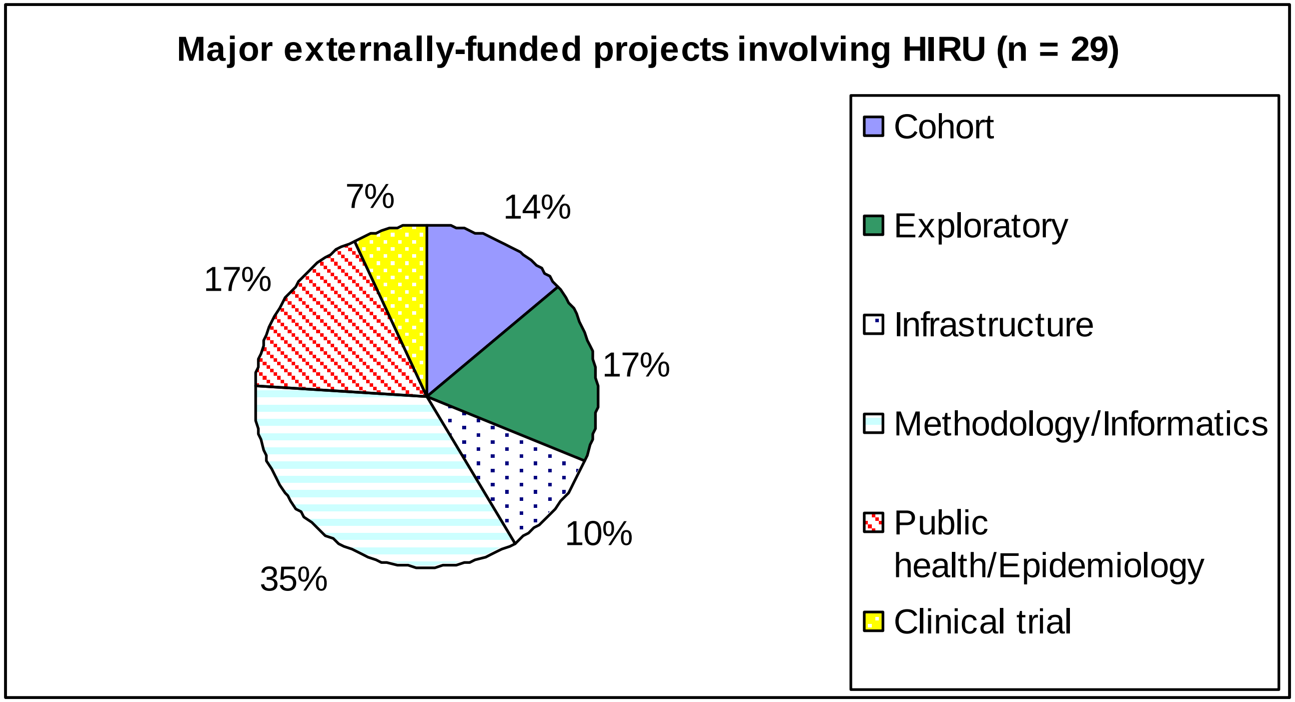
- HIRU wishes to support more research
- Access must be safe and secure (i.e. no data can leave)
- Must comply with our IGRP
- Researchers could be based anywhere (including within NHS and overseas)
- Researchers need different tools, and often need to “feel” the data

Solution: . . . .

Improving access: SAIL Gateway



Externally funded research - types of studies



At July 2009

Not just routine data - study specific datasets

- Wide range of **study** datasets being added to SAIL for (anonymous) linkage (e.g. SAFER Trials, FSBI, Health Survey Wales)
  - For long term follow up
  - Trial data augmentation (for primary and secondary outcomes)
  - Creation of retrospective and prospective cohorts (‘e’ and hybrid)
- Imaging and pathology data

Backroom challenges . . .

A lot of work, despite a big computer!

- Data prep and cleaning
- Concept disambiguation and translation
- Subject expertise with data skills
- Information Governance / security
- Capacity and sustainability

Biggest Single Challenge (RLS) - analysts

Making silk purses from sow's ears!

– 'just give me the data!'

- Many routine datasets have several limitations
- Designed for administrative, not clinical or public health purposes
- FCE v people
- Incomplete or missing data
- Erroneous data
- Recruiting and training analysts to summarise data across many datasets is difficult and time consuming – no training courses
- Issues highlighted by creation of WECC

Concepts to codes – an ongoing effort

An example . . .

To identify individuals with a diagnosis of asthma from (only) GP data, it involved specifying, agreeing and querying for:

14 asthma diagnosis read codes

65 asthma administration read codes

692 asthma medication read codes

Wales Electronic Cohort for Children (WECC)

An all-Wales public health and paediatrics collaboration

Ronan Lyons, David Fone, John Gallacher, David Ford, Caroline Brooks, Sinead Brophy, Frank Dunstan, Mike Gravenor, Kerina Jones, Sailesh Kotecha, Gareth Morgan, Shantini Paranjothy, Sarah Rodgers, Rhys Williams, Gareth John.

Wales Electronic Cohort for Children (WECC)

- Anonymised data from 800,000+ children
- Platform for translating information into child health population policy: answer two questions (of many possible)
- What factors determine the future health service need for individuals that are vulnerable at birth, and inform the development of interventions to reduce health inequalities in these groups?
- What is the influence of the social and physical environment on childhood obesity?

Anonymised datasets to be included in WECC

- NHSAR – population register
- NCCHD - Community Child Health
- ONS Public Health births and deaths
- AWPS – Perinatal Survey
- PEDW – inpatients
- OPMD – outpatients
- CARIS – congenital anomalies
- GP data – depending on availability
- FSM entitlement from NPD
- Environmental data (RALF and small area derived)

Green space and obesity

Access to more green space is associated with lower odds of increasing BMI in children:

- OR: 0.87 (0.79-0.97)

Neighbourhood greenness and 2-year changes in body mass index of children and youth. Bell JF et al, Am J Prev Med 2008;35:547-553.

Research using GIS and health data

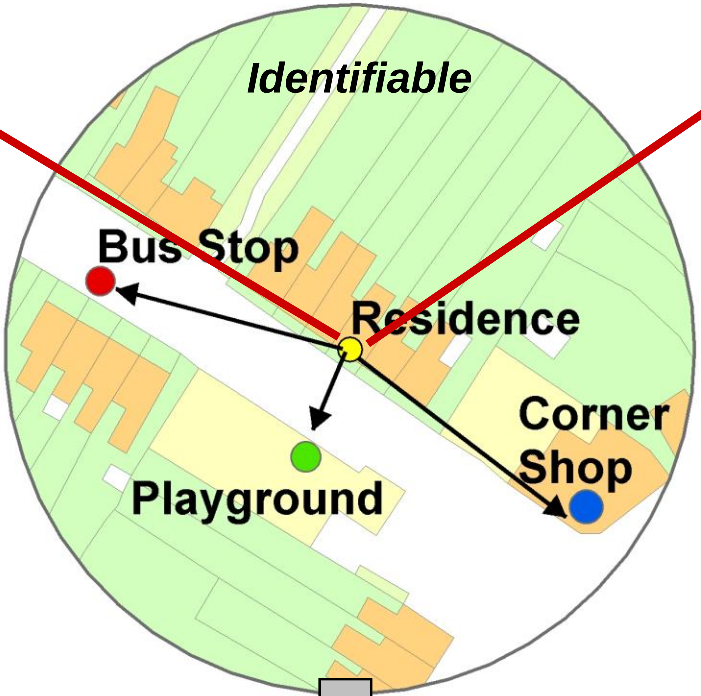
Objective : Create RALFs for every residence in Wales

- Central list held securely within HSW
  - Assign each ALF to a RALF,
  - Track movements of ALFs between RALFs over time (never knowing who or where)
  - Attach GIS measures to addresses – later RALFed
- Conduct research:
  - Likelihood to walk and take exercise
  - Household exposures (e.g. air quality, heating, etc)
  - Model infection disease spreads through household contacts
  - Link survey data with environmental exposure and health outcomes

Environmental metrics and links to health and social data – fictional example

David Williams
8 Main St, Swansea

Male
age 8
Asthma
--
Free School Meals



Margaret Williams
8 Main St, Swansea

Female
age 35
Diabetes
Smoker

Social Database

Anonymous

ALF	RALF	Medical 1	Medical 2	Social 1	Environment1
11223387	5448893	Asthma	--	FSM	5.85
11238889	5448893	Diabetes	Smoker	--	5.85

Same RALF: cohabiting

GP Database

Same environmental metric
e.g. dwelling density,
land use mix,
metres to bus stop or GP surgery

Unusual example

Free School Breakfast Initiative – Data Augmentation

- NPRI/MRC funded
- Relationship between breakfast and nutritional intake and educational outcomes
- 5,758 pupils from cohort in original study
- Linked through SAIL with area deprivation scores (NHSAR link) and individual socioeconomic data (FSM) and educational data (NPD)
- Moore L, Benton D, Lyons R, Murphy S, Tapper K

Not just Wales

- SAIL mechanism can work with data from anywhere
- Collaborate with HRSS and SHIP – UK E-Health Dataset Consortium
- ‘UK’ studies using or proposing to use SAIL:
  - ALSPAC GP linkage
  - MS Register (E, W, NI)
  - AS Register (Swansea/Aberdeen/England) – discussions
  - Sheffield (Ambulance, ED, inpatients, deaths)
- International studies
  - Western Australia DLs
  - South Australia/Northern Territories (SA/NT Datalink)

Acknowledgements: SAIL progresses through continuing collaboration

NHS Organisations

- Health Solutions Wales (HSW)
- Informing Health Care
- Public Health Wales NHS Trust and its components
- Welsh Assembly Government: Information Services Division

Research Organisations/Groups

- NISCHR Registered Research Groups (RRGs)
- Other research groups in Wales and beyond
- WISERD, DECIPHer, UKBiobank, …

And particularly . . .

- NHS bodies, Local Authorities and other government departments (data providers)

Relevant methodology papers

- Ford DV, Jones KH, Verplancke J-P, Lyons RA, John G, Brown G, Brooks C, Bodger O, Couch T and Leake K. **The SAIL Databank: building a national architecture for e-health research and evaluation.** *BMC Health Services Research 2009,* 9:157 (4 September 2009)
- Lyons RA, Jones KH, John G, Brooks CJ, Verplancke J-P, Ford DV, Brown G and Leake K. **The SAIL databank: linking multiple health and social care datasets.** *BMC Medical Informatics and Decision Making* 2009, 9:3 (16 January 2009)
- Rodgers SE, Lyons RA, Dsilva R, Jones KH, Brooks CJ, Ford DV, John G and Verplancke J-P. **Residential Anonymous Linking Fields (RALFs): A Novel Information Infrastructure to Study the Interaction between the Environment and Individuals' Health.** *Journal of Public Health*, 2009, pp. 1-7.