10 YEARS OF SPEARHEADING DATA PRIVACY AND RESEARCH UTILITY

2007 - 2017

www.SAILDatabank.com
Acknowledgements and thanks

We would like to extend our heartfelt gratitude to everyone who has contributed to the Secure Anonymised Information Linkage (SAIL) Databank over the past 10 years. There are far too many people to mention everyone individually, but without their input the SAIL Databank would not be what it is today. We especially acknowledge the support of: Health and Care Research Wales; Swansea University Medical School; NHS Wales Informatics Service; and NHS (GP and secondary care), local authority, Welsh Government and all other data providers. We greatly appreciate the input of the Consumer Panel and other members of the public, and our professional advisors and members of our governance structures. Much of the work over the 10 year period has been without precedent and our staff have repeatedly risen to the challenge to provide practical solutions that synergistically combine to make the SAIL Databank world-class.

Inexpressible thanks are due to our dedicated staff, past and present, who have contributed to the development and operation of the SAIL Databank: Andrew Austin, Arfon Rees, Arron Lacey, Ashley Akbari, Brian Perkins, Caitlin Jones, Caroline Brooks, Charlotte Arkley, Chris Jones, Chris Orton, Christine Mackey, Cynthia McNerney, Dan Thayer, Emma Squires, Ginerva Brown, Huw Collins, Imran Hussain, Jeffrey Peng, Jan Smart, Julie Kennedy, Justin Biddle, Lee Au-Yeung, Luca Ruschetti, Luke Player, Lynsey Cross, Martin Heaven, Mathilda Castagnet, Michael Bole, Mihaela Barbuc, Pam Carrington, Phil Smith, Phil Verplancke, Richard Fry, Richard Noyce, Rohan Dsilva, Ryan Lever, Sarah Rees, Sarah Rodgers, Sharon Hindley, Simon Thompson, Stephanie Lee, Susie Lucas, Ting Wang, Tinnu Sarvatham and others. Grateful thanks are also due to the many data users whose work has benefitted individuals and society, and demonstrated the usefulness of the SAIL databank.

Please cite this report as: Jones KH, Ford DV, Lyons RA (2017) The SAIL Databank: 10 years of spearheading data privacy and research utility, 2007-2017. Swansea University. doi: http://dx.doi.org/10.23889/sail-databank.1001101

Foreword

We are delighted to introduce this report on the first 10 years of the pioneering work of the Secure Anonymised Information Linkage (SAIL) Databank as part of our anniversary celebrations. The SAIL Databank began in 2007, and from small beginnings has flourished into a world-class data safe haven. It has a wealth of anonymised data about the population of Wales, accessible via a secure data sharing platform, all underpinned by an innovative and proportionate Information Governance model. The Databank represents a unique and pivotal resource for population health-related research leading to multiple benefits by informing policy developments and through improved care and service delivery. None of this would have been possible without the goodwill of our many data providers and fruitful partnerships with the NHS, professional bodies, government departments and engagement with the public. Success to date has depended heavily on the expertise of our teams of staff working with us in this complex and rapidly developing field. We look forward to the future with excitement as we continue on this journey towards further improvements and novel developments to benefit health and well-being, along with our existing and new collaborators.

We heartily thank everyone who has played a part in the development of the SAIL Databank so far. We very much hope you enjoy the report and that you will join with us in celebrating the first 10 years of the SAIL Databank.
Introduction

The Secure Anonymised Information Linkage (SAIL) Databank is a national data safe haven of anonymised health and administrative datasets about the population of Wales. It was established by the Health Informatics Group at Swansea University Medical School in 2007 with core funding from Health and Care Research Wales of the Welsh Government, and has been operating for 10 years. The wealth of data within the Databank enables important questions to be answered that could not otherwise be addressed without prohibitive effort, if at all. From the outset, we have built many collaborations and have sought the optimum ways to make data available safely for public benefit. The past 10 years have seen rapid developments in data-intensive research and the SAIL Databank has been key in taking this forward. It has attracted multi-million pound investments and has been the springboard for many more. SAIL data is in increasing demand and its influence stretches across the world. The purpose of this report is to set out challenges that have been encountered and addressed, to celebrate achievements, and to encourage further collaboration with the SAIL Databank.
Breaking new ground

There were many challenges to address in setting out to establish the SAIL Databank; at the time (2006), with the exception of notable examples in Australia and Canada, there were few other countries with data linkage infrastructures from which to learn. Consequently, even dialogues on this topic with government, professional bodies, health and social care organisations, regulators and the public were often ground-breaking activities. So pragmatically, the work of the SAIL Databank began as a pilot in one local authority area in the West Wales region.

We benefitted greatly from partnerships established with the NHS and local authority partners, and with the NHS Wales Informatics Service (NWIS). This allowed us to address the initial challenges, engage with stakeholders in due diligence, and to develop and refine models of working. NWIS provided essential support with access to data, and expertise in developing the separation principle for data anonymisation and the record matching process.

We met and consulted with many government, regulatory and professional agencies including Welsh Government departments, the Information Commissioner’s Office (ICO) for England and Wales, the National Research Ethics Service (NRES), the British Medical Association (BMA), the Royal College of General Practitioners (RCGP), and organisational Caldicott Guardians. These consultations covered a range of issues, including the overall suitability of the SAIL Databank proposal and its potential to enhance research in the public interest, to ensure data protection, to address policy objectives, and to provide value for money. This informed further improvements and resulted in roll-out of the system across Wales from 2007. We were greatly encouraged by the support and endorsements we received.

Milestones

The first 10 years of the SAIL Databank have seen many developments and new ventures. These have included refinements to our operating model, establishing structures and programmes, and taking part in major research, methodological and infrastructural initiatives. It has been a challenging and highly rewarding journey, shared with our many partners and collaborators. Some key milestones are summarised below, including the celebration of our 10 year anniversary!
Choice of model

There were a number of options to be considered in deciding the optimum operating model for the SAIL Databank. The challenges of working with large-scale person-based data collected routinely in the course of health and other public service delivery led us to settle on a repository model, creating an anonymised national data safe haven. Among the factors that influenced the decision were:

- Regulatory and governance frameworks for the safe re-use of data
- Availability of high performance computing
- Variability in the existing IT systems in health and care settings
- Minimising the workload for data providers
- Needing an NHS-based Trusted Third Party (TTP) to manage identities
- Being able to retain a unique field for each person to link their records across datasets
- Opportunities to monitor data quality and completeness
- Efficiency in being able to use and reuse the data held
- Pragmatism and timeliness whilst still retaining the opportunity to adopt new models in the future.

Ready to roll

Following the successful pilot, we spent the early years of the SAIL Databank concentrating on building up our technology, infrastructure and capacity, and in extending the datasets held by the Databank. Some of the key developments are briefly outlined here.

In 2008, as requests to use SAIL data began to expand, we constituted an independent Information Governance Review Panel (IGRP) to provide us with regular independent advice on proposals to work with the data, rather than use ad hoc review. In 2009, after the planning and procurement phase, we commissioned the first of a programme of specialist independent internal audits for compliance with Information Governance. Being open to professional scrutiny to make improvements in the system and to provide assurance to our data providers, the public, clinical/service professionals, our stakeholders, funders and regulators are important principles to us.

During 2007 to 2009, access to SAIL data could only take place on-site using dedicated terminals under the supervision of SAIL Databank staff. Although this was adequate at first, it wasn’t sufficient to meet the increasing demands for SAIL data. In 2010, we addressed this issue by developing the SAIL Gateway as a remote access technology and analysis platform, accessible from anywhere in the world. These and other developments led to our current operating model.
Operating model

The initial years enabled us to establish a comprehensive operating model for the SAIL Databank with privacy by design at each step of the way. As a result, the Databank has developed a world-class reputation for data sharing platforms incorporating innovative Information Governance solutions to maximise both data protection and data utility.

Data transfer

We obtain formal permission from each participating developing their own Data Guardians and Information Governance structures so that due diligence processes are carried out to their satisfaction, and data sharing agreements are in place, before data are transferred. Ensuring that transfers are secure so that data are not lost, or accessed without authorisation, during the acquisition process is essential for data protection and provider assurance.

The SAIL Databank uses only secure upload/download portals with no data being transferred by means such as portable media, email or postal mail.

Separation principle

Datasets are transferred using the separation principle developed with NWIS so that only the data provider sees the identifiable dataset. The provider divides their dataset into two parts: demographic details (including name, address, date of birth) and content (such as medications, diagnoses, educational attainment). The demographic elements are sent to the NHS Wales Informatics Service (NWIS) who operate as a trusted third party (TTP) to match and anonymise the records. The content elements of the dataset are sent directly to the SAIL Databank.

Reliable record matching

NWIS maintains the Welsh Demographic Service register which acts as a proxy for a Welsh population demographic database. Reliable record matching is important to preserve record integrity and identity in the anonymised datasets in the SAIL Databank. Without this, the validity of subsequent research would be in serious question. The matching algorithm was devised at NWIS and we were able to make use of this for data coming to the Databank. It uses deterministic and probabilistic routines sequentially, and as a result of testing and refinement, high rates of matching accuracy (>95%) are regularly attained.

Anonymisation and encryption

At NWIS, the commonly-recognised identifiers, and the NHS number where present, are removed and replaced with an Anonymous Linking Field (ALF).

This is an encrypted number unique to each person represented in the dataset, with no attributable meaning or currency outside the system. The ALF along with gender and minimal aggregated demographics (week of birth, Lower Super Output Area (LSOA) of residence) are sent to the SAIL Databank for recombination with the content elements of the dataset. This method ensures that the Databank does not receive identifiable data but does receive datasets that can be anonymously linked together using the series of ALFs.

Residential Anonymous Linking Fields

As well as ALFs relating to individuals, we developed a system of Residential Anonymous Linking Fields (RALFs). These were created by matching Royal Mail Postal Address Files against the Welsh Demographic Service register at NWIS and producing a unique code for each place of residence. NWIS also provide us with dynamic tables detailing the ALFs associated with each RALF. In this way, households, the individuals associated with them and environmental factors can be studied in combination whilst protecting individual identity and residential location.

Disclosure control

Although the data in the SAIL Databank have been anonymised, we are conscious of the re-identification risks that remain where individuals have sufficient motivation, and so we have developed a suite of measures to mitigate disclosure risks. To begin with, we re-encrypt the ALF to make it double encrypted. This ensures that neither we nor NWIS can reverse engineer the anonymisation process to reveal identities. We are able to limit the extent of, and time period for, data viewed by researchers and to apply methods such as aggregation and suppression of records and variables as necessary. No row level data can leave the Databank unless all relevant regulatory and governance approvals (including informed consent) have been obtained. Our standard operating model is that only results of analysis without cell counts <5 can be released following scrutiny by a SAIL Databank senior analyst acting as a data guardian.

Data access model

SAIL data are accessed via the Gateway\(^2\) platform which uses remote access technology to enable approved researchers to use data within the SAIL Databank virtual environment from their own desktop, wherever they are in the world. The protocols in place allow user authentication and monitoring, and prevent the alteration or removal of SAIL data by users. This has proved to be a highly successful development both in terms of data security and data access. It has influenced the development of data-intensive initiatives in other jurisdictions leading to endorsements of the SAIL Databank.

Data access controls

Data access is controlled and authorised to enable responsible data utilisation. It is imperative that no one can access the SAIL Databank without authorisation, and that there are controls on what is accessed, whilst ensuring the data meets the needs of research studies. Even so, researchers are only allowed to view the data and not to have data extracted from the Databank. The only time this can occur is where a researcher has obtained independent research ethics approval for a study, including permission to use health/administrative records. Individuals wishing to use SAIL data are required to pass a training session for safe conduct with data (‘Safe Researcher’) and to agree to comply with the SAIL data access agreement. They can then be issued with an account and a UbiKey for user authentication.

Data use scrutiny

Apart from in-house scoping and feasibility checking of the prospective uses of SAIL data, all proposals are reviewed by the IGRP. This Panel reviews the proposals to ensure they comply with Information Governance principles and represent an appropriate use of data in the public interest. The Panel includes representatives of professional and regulatory bodies and the general public. Once passed by the IGRP, data access can be provided. Following data analysis, a researcher must submit their results to a SAIL data guardian before they can be released for dissemination. These outputs may consist of tables, charts, coefficients, etc. and they are assessed to ensure risks of disclosure are mitigated.

Information Governance compliance

A programme of specialist independent internal audits for compliance with Information Governance enabled us to implement improvements to our operating processes. We were then able to focus on attaining ISO27001 certification and achieved this for the SAIL Databank in 2015. As a result of on-going efforts, we have a sophisticated suite of Information Assurance measures incorporating physical, technical and procedural controls for the SAIL Databank.

“...secure access to data was straightforward, and being able to use the floating desktop at our own desks is incredibly helpful. Extensive coverage of routinely-collected data in the SAIL Databank really gives Wales a competitive advantage for healthcare research.”

Ruth Poole
Senior Evaluation Scientist, Cedar Healthcare Technology Research Centre, Cardiff University Health Board
Public Involvement and Engagement

We consider Public Involvement and Engagement (PI/E) to be fundamental to the work of the SAIL Databank. Proper PI/E is integral to Information Governance frameworks to ensure that, not only are legal and regulatory requirements followed but, public views and societal interests are taken into account for the socially responsible reuse of data. We have a strong commitment to PI/E and enjoy a varied programme of activities in this area. Initially, we consulted with public interest groups and learned about engaging with the public in data-intensive research.

Members of the public have long been involved on steering groups of research studies, providing valuable input into project direction. Since 2011 we’ve had a thriving and active Consumer Panel comprising members of the public with a variety of interest areas. The Panel advises us on data protection issues in data-intensive research from the perspective of service users and carers. Researchers are encouraged to meet with the Panel to discuss their proposals and receive a public viewpoint and advice. The Panel reviews information to check its suitability for the general public, and advises on routes and methods to engage with people about data-intensive research.

In addition to the ongoing programme of work undertaken by the Consumer Panel, we actively engage with the general public through opportunistic and dedicated channels. Our enthusiastic public engagement team regularly attends events across Wales with different groups and sectors of society to discuss the work of the SAIL Databank. This provides opportunities to describe the work being done, to present a balanced view on data-intensive research, to gauge the public opinion, and to ignite collaborative opportunities.

In 2016, we took part in the British Science Festival and presented the work of the Databank to the public in a popular, interactive session via a character called ‘Dave the Data Item’. The session showed how items of data can be linked (like pieces of a jigsaw) within the SAIL Databank to gain new information whilst retaining individual anonymity.

“Because the data are anonymised, preventing individuals’ data being identified, I have faith in the system. There are safeguards in place so that researchers only have access to what is required for their piece of research and there is a rigorous evaluation process before applications to use the databank are approved.”

Dot Williams
Consumer Panel member

“Data security is very important to everyone, but having looked at the stringent security measures taken by SAIL to preserve anonymity, I am confident that the data used will remain confidential and cannot be traced back to any individual person.

So much useful information is already in existence about us as individuals and communities regarding our health, wellbeing, education, development and nutrition that it seems logical that if this information can be used for research which will benefit the public, then it should be.”

Julie Hepburn
Consumer Panel member
Datasets in the SAIL Databank

Wales has a population of approximately 3 million people, and taking into account births, deaths and migrations, the SAIL Databank holds over 15 billion anonymised records about 4.5 million people.

The Databank began with an extract of national hospital records from the Patient Episode Database for Wales (PEDW), and GP and social services data from the pilot local authority area. Since roll-out in 2007, we have been steadily increasing the volume and types of datasets in the Databank, including data going back 10-20 years depending on the supplying system. The SAIL Databank now holds anonymised data from a remarkable ~80% of Welsh GP practices and a wealth of national datasets including: in-patient and out-patient PEDW records, emergency care, ONS births and deaths, screening services, Welsh Cancer Intelligence Surveillance Unit (WCISU), Congenital Anomaly Register and Information Service for Wales (CARIS), All Wales Perinatal Survey (AWPS), Welsh Health Survey, Welsh Index of Multiple Deprivation (WIMD), Welsh Fire and Rescue Service Data and National Pupil Database Wales.

Having begun with health data, we have extended our scope to include administrative datasets to create even greater opportunities to build rich longitudinal cohorts for research. Furthermore, we are able to work with emerging data types such as free-text, image and genetic data as they become accessible. Datasets that have been collected as part of a specific study can be brought into the SAIL Databank for linkage to existing SAIL data provided that the relevant approvals are in place. This can be a great advantage to some studies, as researchers are able to augment their data with the routinely-collected data in the Databank, thereby creating a richer picture. The numbers and types of projects using SAIL data continue to increase and their outcomes are creating a lasting impact.

SAIL Databank projects

Between 2007 and 2017, over 250 projects and 350 researchers have been approved to work with the SAIL Databank. These have included many fruitful collaborations and we are keen to encourage more so that the data are used to best advantage. Studies include research and service/policy evaluations of various designs including: cohorts, case-controls, natural experiments and clinical trials, and address many identified needs. Work has been focused on important areas such as: reducing inequalities, managing chronic conditions, emergency care, health economics and disease burden, screening and preventative medicine, reducing mortality, housing and health, meeting the needs of different age groups, and the promotion of health and well-being. A selection of projects using the SAIL Databank is summarised to illustrate their variety and potential or achieved impacts.

Reducing inequalities

The Welsh Government invests over £120 million each year to support over 60,000 people to live as independently as they can. The Supporting People Programme provides housing-related support to vulnerable people in Wales by offering early help to reduce demand on other services such as health and social services. Initial work showed that GP and A&E visits were reduced among people who took up the programme and recommended a fuller evaluation using linked administrative data in conjunction with the SAIL Databank. The findings of this study will inform policy-making to improve the quality of life and ability to live independently of some of the most vulnerable groups in Wales, including people threatened with homelessness.

Managing chronic conditions

Ankylosing spondylitis (AS) is the second most common inflammatory arthritis after rheumatoid arthritis. Since the condition starts early on in the life course, education, employment and family prospects can be limited. The effects of the condition can vary widely from one person to another which makes it difficult to predict who will become most affected by the condition. By linking data within the SAIL Databank this study found that people who had severe disease flares had poorer outcomes in terms of functional ability, disease severity, mental health and employment. It estimated the cost of AS to be around £20,000 per patient, per year, mainly attributable to functional impairment and loss of work. Early detection of those most likely to be more affected by AS means that individuals can be identified as candidates for expensive medications, and allows more accurate cost-benefit analyses to take place. This study produced the most comprehensive assessment of the true cost of AS to date.

Emergency care

Falls in elderly people are an important issue internationally. It is estimated that around 30% of people living at home aged 65 years or older fall every year in the UK. A six-year study using SAIL data to estimate the costs and benefits of new procedures for paramedics in assessing older people following a fall, with an option to leave them at home and referral to a community falls service. The findings indicated that ambulance services could introduce the new procedure safely and at low cost, and expect modest reductions in further 999 calls.

Health economics and disease burden

Asthma UK estimates that there are around 5.4 million people in the UK currently receiving treatment for asthma. Three people die every day due to asthma attacks and research shows that two thirds of asthma deaths are preventable. This study sought to provide the first UK-wide estimates of the epidemiology, healthcare utilisation and costs of asthma, and used national health surveys, and routine administrative, health and social care datasets. For Wales, the study used SAIL data. It was shown that asthma results in at least 6.3 million primary care consultations, 93,000 hospital in-patient episodes, 1,800 intensive-care unit episodes and 36,800 disability living allowance claims, and costs of at least £1 billion. The profile of asthma created for the UK will be useful to national and regional policymakers and to health commissioners both in the UK and internationally since it can be used as a template for mapping asthma in other nations.
Screening and preventative medicine

According to Cancer Research UK, cervical cancer is the second most common cancer in women under the age of 35. In the UK, 2,900 women a year are diagnosed with cervical cancer, which equates to about 8 women every day. Girls aged 12 to 13 years are offered an inoculation against the Human Papilloma Virus (HPV), as part of the NHS childhood vaccination programme, to protect against cervical cancer. Further information was needed on the adult uptake of cervical screening in relation to vaccination status. By analysing the data in the SAIL Databank relating to 31,000 women, it was found that cervical abnormalities were more likely among those who had not been vaccinated. Also, women who lived in socially deprived areas were less likely to have been vaccinated or to attend cervical screening. This has led to recommendations to promote further engagement of younger age groups with health services in these areas.

Reducing mortality

Every year over 800,000 people die by suicide worldwide, accounting for 1–2% of deaths. Prevention of suicide is a global public health challenge. Suicide is a tragedy and causes distress for family, friends, professionals and the wider community. Although the factors that contribute to suicide are many and complex, it is potentially preventable. The Suicide Information Database for Wales holds linkable data on prior health, nature of previous contacts with services and wider social circumstances of all those who have completed suicide within the population of Wales within the SAIL Databank. The rich information gained is being used to make recommendations to reduce suicide in Wales and internationally.

Housing and health

Poor quality housing adversely affects residents’ health but there is a lack of high quality evidence to support improvements. This NIHR Public Health Research programme funded study used the SAIL Databank to investigate the health impact of bringing council housing up to a national quality standard. A package of housing quality standard improvements. The project team is working with public health and housing colleagues to disseminate results to inform national and local authority housing policy.

Children

The Welsh Electronic Cohort for Children (WECC) includes anonymised records of the >800,000 children living in Wales between 1990 and 2008 compiled from multiple datasets in the SAIL Databank. In addition to health-related data, WECC includes education records from 1994 onwards. This complex cohort enables numerous research questions to be addressed. For example, WECC has shown that: very low birth-weight, prematurity, being small for gestational age at term, and birth by Caesarean section all increase the likelihood of respiratory admissions up to age 5 years; prematurity is associated with educational underachievement at age 7 years; and, children who move house frequently or at particular periods have an increased risk of poor health and educational underachievement. WECC represents a rich resource for child health research.

Young people

There is growing concern about the increasing numbers of young people suffering from anxiety and depression and that some of the medications used are not fully tested or licensed for use in these age groups. The study used SAIL data on over 300,000 children and young people aged 6 to 18 years and found that, between 2003 and 2011, the rate of young people newly presenting with anxiety symptoms more than tripled with a significant increase in prescribing for the 15 to 18-year-old group. The study team is now working with Welsh Government, the Welsh Medicines Resource Centre and Health Boards to improve practice. The findings are also informing on-going programmes to redesign child and adolescent mental health services in Wales and further afield.

Older people

There are approximately 700,000 people in the UK with dementia. According to Alzheimer UK, 90% of people with dementia experience behavioural and psychological symptoms such as agitation, aggression, delusions and hallucinations. As a result they might be prescribed antipsychotic drugs. However, there is a growing concern about the inappropriate use of antipsychotics in the elderly. Using data in the SAIL Databank, this study showed an increased risk of adverse medical outcomes, in particular venous thromboembolism, stroke and hip fracture in older people with dementia who are exposed to antipsychotic medication. These results support recommendations for reducing the use of antipsychotic drugs for people with dementia, which is a key element of the dementia plan for Wales and a national priority in England.

Promotion of health and well-being

The Health and Attainment of Pupils involved in a Primary Education Network (HAPPEN) is made up of professionals working in health, education, and research. Good health and wellbeing of children is important to maximise achievement, employment opportunities, and their subsequent health and wellbeing as adults. HAPPEN focuses on children aged 9-11 years making use of data collected on body mass index (BMI), fitness, nutrition, physical activity, sleep, concentration, and the children’s recommendations on improving health in their area. These data are then used in conjunction with SAIL data to assess health and educational outcomes. The first batch of data analysed was on 2,500 school children in the Swansea area and reports were fed back to local schools, health professionals and charities. The number of schools joining the network continues to grow, and the partnership approach provides a more unified and evidence-based way to promote child health and well-being.
International influence

The robust, proportionate and innovative Information Governance and infrastructural models used in the SAIL Databank are widely acknowledged to be world-class. They have influenced numerous data-intensive systems in Great Britain and internationally, and they have been strongly endorsed in reports of such developments. Some examples are given here by way of illustration. The 2016 report of the Irish Data Access, Storage, Sharing and Linkage (DASSL) system, praises many aspects of the SAIL Databank model including: the separation principle using a TTP, our public engagement workstream, providing safe access to data in accordance with privacy-protecting policies and procedures, and the wealth of datasets held.

The Council of Canadian Academies’ report (2015) identified the SAIL Databank as a “best practice entity” in multiple areas such as: evaluating privacy concerns when enabling data access, ensuring data use is appropriate, Information Governance and respect for privacy, as well as the separation principle and our public engagement work. The SAIL Databank has been highly recommended in two recent Organisation for Economic Co-operation and Development (OECD) reports. Their 2015 report highlights the value of the Gateway, the robust Information Governance model, the SAIL Databank application and proposal approval process, the support provided to researchers and analysts, and the important role of a Consumer Panel in educating the broader community in the drive to be open and transparent. The report published in 2016 was similarly complimentary and recommended that:

“SAIL should be seen as a highly valuable resource, and an example for other countries to follow.”

Senior members of the SAIL Databank team are regularly approached to act as advisors on infrastructural developments, Information Governance models and effective methods for the reuse of population data in data-intensive initiatives. These include the Australian Population Health Research Network and the Canadian Ministry of Health and Long-Term Care, both of which are in countries with highly developed data sharing infrastructures, especially at state/provincial levels.

Peer-reviewed outputs

The outcomes of many SAIL Databank-related studies have resulted in a growing list of peer-reviewed outputs, with over 160 produced to date. Publication numbers between 2007 and 2016 are illustrated below.

Outputs cover results of cohort studies, case-control projects, clinical trials, surveys, natural experiments, case studies, economic, service and policy developments, methodological advances, Information Governance innovations and infrastructure developments, amongst others. They include work involving the NHS and other service professionals, local and national government researchers, academics from many disciplines, industry collaborators, and public involvement. It is encouraging to note that outputs originate, or include contributions from, a broad range of collaborators within Wales, other parts of the UK and beyond. This demonstrates the value and wide applicability of the rich data within the Databank to answer important questions. The updated list of publications can be found on the SAIL Databank website: https://saildatabank.com/saildata/sail-publications

“SAIL provides a rich and trusted data resource, exceptional in many respects, supporting research for patient and public benefit: a vital facility for interdisciplinary team science.”

Professor Carol Dezateux
FMedSci, Lifecourse Epidemiology and Biostatistics, University College London

“We in the Health Research Board, Dublin, have been glad to work with our thoroughly professional and innovative colleagues from SAIL in progressing the DASSL model – a model for the safe linking and sharing of health research and related data, in Ireland.”

Rosalyn Moran
Cross-Organisation Strategic Project Manager, Health Research Board, Ireland

“The SAIL databank provides access to a rich array of linkable data that is invaluable for research and policy development. It has been a powerful model for databanks that have been/are being built in Australian jurisdictions.”

Dr Merran Smith
Chief Executive, Population Health Research Network, Australia

“We dare to suggest that the SAIL database is a model that has been well worth the effort...”

Michael Hillmer
Ministry of Health and Long-Term Care, Ontario, Canada
The success we have enjoyed with the SAIL Databank has led to new developments and partnerships with multi-million pound investments based on SAIL data, infrastructure platforms, innovative Information Governance and lessons learned during the 10 year journey. We refer to the SAIL Databank and these developments collectively under the umbrella term of Population Data Science®, being the science around data about people. These include:

UK MS Register
From 2009, the MS Society of Great Britain and Northern Ireland has funded our work on the UK MS Register. This innovative register is the first of its kind to combine data provided directly by people with MS with neurology clinic and routine datasets. It has an active research programme, the findings of which are clarifying the picture of living with MS in the UK and leading to policy and practice improvements¹⁻²⁻³.

www.ukmsregister.org

The Farr Institute for Health Informatics Research
From 2012, the Medical Research Council (MRC) and a consortium of other funders invested in four e-health research centres across the UK. The work of the Farr Institute focuses on delivering high-quality data-intensive research to advance the health and care of patients and the public. We lead one of these centres with a collaborative research programme based on the SAIL data and model.

www.farrinstitute.org

Administrative Data Research Network
In 2013, the Economic and Social Research Council (ESRC) established four Administrative Data Research Centres (ADRCs), to enable new research using administrative data for public benefit. There is an ADRC in each of the UK jurisdictions and we lead ADRC-Wales. In collaboration with our partners, we are building on our experience with existing SAIL data and working with a range of administrative datasets from government departments and other sources.

https://adrn.ac.uk

Welsh Research Infrastructure
The SAIL Databank has always collaborated with, and had a formal input into, the Welsh research infrastructure funded by Health and Care Wales (Welsh Government), and this relationship has been further strengthened following the latest commissioning refresh in 2014. Various themed research centres and units, regional clinical trials units and infrastructure support services have been created or extended. This is enabling many valuable co-working opportunities to flourish and serve the needs of the Welsh population⁴⁻⁵.

National Research Data Appliance
In 2014 we developed the National Research Data Appliance (NRDA). This is a suite of novel, modular, data concentrating technologies for use by the NHS and other data owning organisations to enable them to put their data to more effective use. It includes automated probabilistic matching, anonymisation, record linkage, dataset management, metadata capture, data quality assessment, security and access control, data transportation and sharing. It can also provide optional modules such as Natural Language Processing (NLP) for free-text analysis.

“Over the past decade our partnership with SAIL, and its role in ADRC-Wales, has provided ground-breaking research on social and health issues using innovative, safe and secure technology that has directly informed Welsh Government programmes. It’s a crucial part of our research infrastructure here in Wales.”

Glyn Jones
Chief Statistician, Welsh Government

UK Secure eResearch Platform
In 2014 we created the UK Secure eResearch Platform (UKSeRP). This is an ISO27001 certified independent and customisable technology and analysis platform that allows multiple, complex datasets to be managed, analysed and shared safely. It is underpinned by NRDA technology and can be configured to meet the particular needs of research programme owners, subject to data provider permissions and legislative and regulatory approvals. Requirements are determined through due diligence processes and are formalised in contract to ensure Information Governance responsibilities are met. UKSeRP can host data for multiple independent research programmes whilst keeping the respective data separate within a secure environment. It can also accommodate emerging data types, such as free-text, imaging, genetic and geographic information, thus enabling the creation of rich cohorts for research⁶⁻⁷. UKSeRP is being used by long-standing major research programmes such as UK Biobank, and newer ones such as Dementias Platform UK.

Dementias Platform UK
The MRC-funded Dementias Platform UK (DPUK) was established in 2014, to be a world-leading resource for person-focused dementia research, by fast-tracking scientific understanding, treatments and the prevention of disease. It is based on multi-centre collaborations and data-intensive population studies to reveal new insights into dementia. DPUK works with us for access to the rich array of data in the SAIL Databank, and we provide DPUK with a UKSeRP secure infrastructure and technology platform to host DPUK data.

Prudent Healthcare Intelligence Hub
The NHS in Wales has to meet the perennial challenge of achieving better outcomes for patients while faced with an increasing demand on healthcare services and shrinking budgets and resources. To tackle this problem, in 2016 the Farr Institute created the pan-Wales Prudent Healthcare Intelligence Hub to deliver better outcomes for the NHS through the use of anonymised health data held in the SAIL Databank and to train NHS staff in data science. It supports the NHS in gaining an improved understanding of their populations, and in evaluating and planning services. Prudent Healthcare projects are owned by NHS organisations and supported by the Farr Institute and Swansea University.

HealthWise Wales
This initiative, launched in 2016, aims to develop in-depth knowledge on the health of the nation by collecting information directly from individuals about their well-being and lifestyle, and studying this anonymously in conjunction with health records in the SAIL Databank. It aims to recruit 260,000 people in Wales, aged 16 and older, over a five-year period. The information gained will be used to help the NHS plan for the future, including support for health needs and evaluation of services and policies. The study will also help with recruitment to research studies as members of the public who sign up will also have the opportunity to be contacted to take part in health and social care research projects relevant to their needs and interests.

Cloud Infrastructure for Microbial Bioinformatics
This infrastructural development (CLIMB), launched in 2016, aims to deliver bioinformatics capability to microbiologists working in UK academia. With improvements in next generation sequencing technologies, generating genomic data sets has become easier. However, many academics don’t have access to the necessary resources for bioinformatics analysis. CLIMB deploys a world-leading cyber-infrastructure for microbial bioinformatics providing cloud-based compute, storage, and analysis tools to support microbiological research. Being able to study the human microbiome in conjunction with routinely-collected data within the SAIL Databank will provide unique new insights for research.
From small beginnings in 2007 the work of the SAIL Databank and associated initiatives in Population Data Science has expanded greatly. From a staff of less than 10 personnel at the outset, we currently have approximately 100 multi-disciplinary staff housed within our dedicated state-of-the-art Data Science Building (DSB), apart from partners and collaborators in other parts of Wales, the UK and across the world. The 6-floor DSB was made possible by a multi-million pound investment by the MRC, ESRC and Welsh Government to bring together the Farr Institute, ADRC Wales, the SAIL Databank, closely associated research groups, the NHS and industry. It is equipped with physical access controls throughout, zoned security levels, CCTV monitoring and a purpose-built safe setting as a micro-data lab to contribute to robust Information Assurance. Having a dedicated building greatly assisted us in successfully gaining ISO27001 certification for the SAIL Databank, as we were able to define the physical security systems and configuration. The DSB team have varied backgrounds, including: epidemiology, statistics, health service research, public health science, multiple clinical specialties, geography, demography, social science, management, economics, law, information governance, public engagement, genetics, bioscience, computer science, mathematics and physics. Altogether the DSB is a vibrant environment for research and development with space for further growth in the future.
The first ten years of the SAIL Databank have coincided with a period of rapid developments in data-intensive research and initiatives. For us it has been a time of building partnerships and collaborations, learning and innovation, addressing considerable challenges, and passing many milestones in establishing the SAIL Databank as the world-class data safe haven and rich databank that it is today. On reflection, we’ve learned that pretty much everything takes longer, more money, more staff, more energy, a wider range of expertise, and is more complicated in practice than in theory! But as a result, the SAIL Databank is widely recognised for its breadth and depth of datasets, for secure data sharing platforms and for its innovative Information Governance, which together combine to create an unsurpassed, highly-reputable, unique and pivotal resource for population health research.

The work is by no means static or complete. A current area of expansion is in the use of emerging data types, notably, genetic and free-text datasets as well as imaging, accelerometers (for ‘actigraphy’ data on physical activity) and smartphone data. A programme of research is underway to elucidate the particular Information Governance issues pertinent to these data types, and to develop the appropriate technical and methodological tools to enable their safe, efficient and acceptable use in conjunction with SAIL data.

The innovative UK MS Register has been designed to combine information provided directly by people with MS with data from neurology clinics and routinely-collected health and administrative data. Even so, there are additional data types that could be used with the appropriate methodologies and data governance models in place. There is an enormous amount of unused free-text data about patients’ conditions, treatment plans and care circumstances buried in clinical notes. The MS Register team is working to address this challenge by using Natural Language Processing (NLP) technology. This makes sense of unstructured data in clinical notes and transforms it into the SNOMED terminology which can then be analysed by researchers. Further work is underway to incorporate important MRI data and to work with actigraphy data provided by individuals on their activity levels. The MS Register is already making a major contribution to revolutionising our understanding of MS in the UK, and success in elucidating the information in largely untapped data types would open up new and exciting insights to benefit people with MS.

“Looking to the future

The first ten years of the SAIL databank has seen it develop into a world class platform for the use of linked data for research and evaluation. Wales is now in the enviable position of having an enormous opportunity to use the SAIL databank in the coming decade to understand and improve the health and wellbeing of the population of Wales in a systematic way.”

Dr William King
Specialist Registrar in Public Health, Public Health Wales

“If you want to know what data linkage in the UK might look like in future decades, take a look at where SAIL is currently.”

Dr Dermot O’Reilly
Director of ADRC-Northern Ireland, Queen’s University Belfast
Closing remarks

Since the SAIL Databank was established in 2007 there have been major developments in the data landscape, with new data paradigms, technologies and governance frameworks, and the pace of change continues to accelerate. We would like the SAIL Databank to remain at the leading edge of data-intensive research and we strive towards this purpose. This report has included much about the benefits of using anonymised person-based data for research. The literature abounds with such examples and this is a great motivation to inspire further developments in this field. The constraints and restrictions which should be applied to the use of anonymised person-based data for research is a similarly well documented and published area of study. However, far less has been published about the harm caused when data are not used, and this is also an important issue to consider in promoting the safe use of data. An international case study was recently conducted on the harms due to the non-use of health-related data. It showed that proving unequivocally in law that an instance of harm was caused by the non-use of data can be difficult. But it is, without doubt, a real problem with far-reaching consequences, being implicated in the deaths of thousands of individuals and potentially costing billions of pounds to societies worldwide. In terms of the analogy of the blind man in the dark room looking for the black cat, the non-use of data is a large, agile, polymorphic, lethal, black cat that must be tamed so data can be used more efficiently for research to benefit the public and save resources15.

“As SAIL has set the global benchmark for innovation, security and quality of population health research based on linked data records. The benefits of SAIL will be seen for generations to come.”

Professor Belinda Gabbe
Head of Pre-hospital, Emergency and Trauma Research, Monash University

As this report shows, the SAIL Databank has come a long way in ten years, and has overcome significant challenges to achieve the international recognition it currently enjoys. We have already formed strong links with many research organisations, and we hope to forge relationships with new partners as we move forward. There will be challenges ahead, but we look to the future with excitement as we continue to work on improvements and novel developments to enable the best use of data to benefit population health and well-being.

References
