This page is intentionally blank
# Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Foreword</td>
<td>1</td>
</tr>
<tr>
<td>1. Introduction</td>
<td>2</td>
</tr>
<tr>
<td>2. Mission</td>
<td>2</td>
</tr>
<tr>
<td>3. SAIL Staffing structure</td>
<td>3</td>
</tr>
<tr>
<td>4. Project partners</td>
<td>4</td>
</tr>
<tr>
<td>5. Work Undertaken during 2016 – 17</td>
<td>4</td>
</tr>
<tr>
<td>5.1. Work Package 1 - Data Acquisition and Completeness</td>
<td>5</td>
</tr>
<tr>
<td>5.2. Work Package 2- Data Linkage methods and management</td>
<td>7</td>
</tr>
<tr>
<td>5.3. Work Package 3 - Supporting Health and Social Care Research</td>
<td>9</td>
</tr>
<tr>
<td>6. Performance Indictors for 2016 – 17</td>
<td>15</td>
</tr>
<tr>
<td>7. Key Achievements</td>
<td>16</td>
</tr>
<tr>
<td>8. Impact of SAIL projects in the real world</td>
<td>17</td>
</tr>
<tr>
<td>9. Conclusion</td>
<td>18</td>
</tr>
<tr>
<td>Appendix 1 - Publications by Projects Using SAIL Data</td>
<td>20</td>
</tr>
</tbody>
</table>
Foreword

We are delighted to introduce our Annual Report which highlights the work and achievements of the Secure Anonymised Information Linkage (SAIL) Databank between April 2016 and March 2017. This year has been a particularly significant one for SAIL as it marks the 10\textsuperscript{th} anniversary of the launch of our databank, and the pioneering programmes of work in the field of health informatics, information security and population health data which have accompanied it. SAIL holds a wide range of anonymised health, social care and other related data about the population of Wales, which is made available to approved users via a secure data sharing platform, underpinned by an innovative and privacy protecting Information Governance model. From its inception the Databank has provided a unique resource for population health-related research, and we are proud that during 2016 -17 it has continued to grow and develop. A number of exciting new datasets (including social care records, exercise referral information and narrative data from hospital patient records) have started to become available to SAIL, and during this year a great deal of work has been undertaken to analyse these new datasets and prepare them to be included within the Databank and made available to the research community.

Prior to 2016 we had been working for a number of years to develop a set of innovative data storage and linkage products including the United Kingdom Secure Research Platform (UKSeRP) a high powered, safe and secure e-research platform which allows data owners and researchers to store, access, share, analyse and link data at scale in a governed environment and the National Research Data Appliances (RDAs) which were designed and built under grant funding received from the MRC to create a secure data cataloguing and linkage environment. 2016 – 17 has seen this technology being deployed by a number of national research programmes, including the UK Dementia Platform, UK Biobank and the Avon Longitudinal Study of Parents and Children (ALSPAC) and also within the NHS in Wales.

During this year, relationships with our data providers, the NHS, professional bodies and government departments have continued to grow and develop. This has resulted in a number of significant research and service evaluation projects which have created real public benefit by informing health and social care policy developments and enabling improved care and service delivery. We have also continued to extend our engagement with the public, holding a variety of events to inform the public of our work and to seek feedback on the operation of SAIL. We very much hope you enjoy our 2016 – 17 Annual Report and that you will join with us in celebrating our achievements in this 10\textsuperscript{th} year of the SAIL Databank. 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.
1. **Introduction**

1.1. 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 research 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 multimillion pound investments and has been the springboard for many more. SAIL data is in increasing demand and its influence stretches across the world.

1.2. The SAIL (Secure Anonymised Information Linkage) Databank is managed by the Population Data Science group based at Swansea University Medical School. With over 13 billion anonymised person-based records loaded to date, SAIL represents a valuable and unique national resource for e-health research and evaluation. It is fundamentally a partnership with public sector organisations, primarily the NHS, without whose data it could not operate.

1.3. Based on a state-of-the-art technical infrastructure provided in partnership with High Performance Computing (HPC) Wales, and overlaid with innovative data extraction and transportation technologies, the SAIL databank is an internationally renowned exemplar of how to address the challenges of large scale health data research. Robust Information Governance arrangements and a suite of approved privacy-protecting technologies and approaches ensure that data is made available to researchers in a safe controlled environment.

1.4. SAIL receives funding from Health and Care Research Wales which covers the cost of the core staffing team, and also recovers money on a not for profit basis from individual research projects which use SAIL services. Cost recovery from projects, supplemented by contributions from within the University, fund the technical infrastructure and an additional five staff posts. With an ongoing freeze on funding levels and rising infrastructure and staff costs, resourcing SAIL is becoming an increasing challenge.

2. **Mission**

2.1. To provide a safe and trusted National Data Repository which harnesses population-scaled health and associated data to increase the quality and quantity of research, and support better policy making, practice and citizen health and wellbeing, by
working collaboratively with data guardians, academics, members of the public, practitioners and policymakers from Wales, across the UK and internationally.

2.2. We are working to achieve this mission by continuing to improve our services in three key areas. These overarching objectives are:

2.2.1. **Data Acquisition and Completeness**
- To develop and increase the flows of routine data into the SAIL databank, in consultation with NHS organisations, public sector bodies and other data-providing organisations.

2.2.2. **Data Linkage Methods and Management**
- To enhance the development of novel methods and techniques to link, manage and utilise large-scale datasets for research in order for Wales to remain at the forefront of health informatics research.
- To provide robust, assured processes for the safeguarding of individual linked data for research.

2.2.3. **Collaborating Effectively to Support Research using Health and other population level administrative data for Research**
- To provide an effective model of collaboration with a wide range of stakeholders for safe and secure access to SAIL databank to support health and social care research.
- To collaborate closely with the NHS Wales stakeholders and other public sector in order to share with them the benefits of the SAIL databank, drive data acquisition and quality, share knowledge and experience, and maximise the impact SAIL has on health and wellbeing in Wales.

2.3. These objectives form the basis of our three main ongoing work areas (work packages) which are described in section 5 below.

3. **SAIL Staffing structure**

3.1. The SAIL core team (the staff funded by Health and Care Research Wales) comprises a small group of highly skilled individuals who maintain and develop the SAIL Databank. An additional five data analysts are funded from and provide support to a range of individual projects. At any one time the SAIL team provide support to a portfolio of around 100 live projects whilst simultaneously managing and processing an average of 75 new enquiries at various stages of the application and project set up process.
Managing this considerable level of demand for SAIL services (which is continuing to increase year on year) requires a high level of efficiency by the SAIL team, and we continue to work to streamline our workflow processes to enable us to cope with demand. We also owe an enormous debt of gratitude to some of our key research group partners. We are co-located with the Administrative Data Resource Centre, the FARR Institute and the National Centre for Population Health and Wellbeing Research (NCPHWR), all of whom work closely with the SAIL Databank team to develop and enrich the SAIL environment, the increase the range of data held and contribute to the body of knowledge around data usage methodologies. Without the contributions made to SAIL by these partner organisations, our ability to continue to develop and improve SAIL would be compromised.

4. Project partners

4.1. Key stakeholders in SAIL’s continued success include the research community whom we serve (both within Wales and internationally), the NHS and other data providing organisations, and the public whose data forms part of the resource which is SAIL. We have a number of ways of interacting with all three stakeholder groups including focus groups, newsletters and other forms of electronic communication. Our 2015 – 16 Annual Report described the establishment of a new joint academic/NHS unit to use data in SAIL for planning and evaluation of interventions to improve the quality of prevention and care services. We are proud to report that this Prudent Healthcare Intelligence Hub (PHIH) is now well established and underpins our increasing level of collaboration with NHS colleagues. Activities relating to the development of our relationship with the NHS and with all our key collaborative partners during the 2016 – 17 period is described in more detail in section 5.3.

5. Work Undertaken during 2016 – 17

In line with our three strategic objectives, we organise our annual work programme for the SAIL Databank team into three work packages.
5.1. **Work Package 1 - Data Acquisition and Completeness**

The purpose of this work package is to continue to increase the range of data available within the SAIL databank, in consultation with the researchers who use SAIL and data providing organisations such as the NHS and public sector bodies. During the 2016 – 17 period we have undertaken a number of very positive and constructive pieces of work in this area, as described below:

5.1.1. **National Data Gazetteer**

This is a single catalogue or “Gazetteer”, a consolidated view of the data assets of Wales not already held within SAIL. SAIL already contains most (but not yet all) of Wales' nationally-scaled health datasets, but a considerable wealth of data remains distributed within individual health organisations. Many of these, despite being held in differing structures and formats, are capable of being analysed individually and simultaneously and are potentially important resources for research, and therefore the Gazetteer is valuable both as a reference source to researchers looking to locate particular datasets, and also to the SAIL team to inform their data acquisition priorities. The SAIL team have continued to make new entries and maintain the Gazetteer. In total, since 2015 81 new datasets have been identified and recorded in the Gazetteer, 56 of which have been added during the 2016 – 17 period.

5.1.2. **Implementation of Research Data Appliances in Wales**

As described in the 2015 -16 Annual Report, funding from the MRC had previously provided capital investment to allow the SAIL technical team to develop a revolutionary new technology: Research Data Appliances (RDAs). These devices provide data holders such as Health Boards or Local Authorities with a secure data cataloguing and linkage technology which helps them understand, link and share their local datasets more easily, both within their organisation and with other partner organisations including SAIL.

Work during the 2015 – 16 period had focussed mainly on development of the technology and achieving support for their use from data owners in Health Boards and the National Wales Informatics Service (NWIS). During this 2016 – 17 period a significant amount of work in partnership with the NCPHWR has led to implementation of RDAs throughout Health Boards in Wales. This has already facilitated the development of an exciting new project Recruitment of Individuals for Trials, Research or EVALuation (RITREVA) which could significantly reduce the time and effort required to identify potential participants for clinical trials. Funding is being sought for this project, which should result in some significant early benefits to the NHS and the clinical trials research community.

Looking to the future, funding has just been secured from ESRC and Welsh Government for the Dataflow Development Project. This will pilot the roll out of RDAs into a sample (five) Local Authorities to facilitate the acquisition and onward transmission of data from LAs into SAIL. This work is scheduled to begin during 2017 – 18 and it is anticipated that it will result (for the first time ever) in a regular, standardised flow of local authority data into the SAIL Databank. Being able to routinely link the healthcare records of individuals to information about their social
circumstances (housing, education etc) will enable a new and exciting range of research to be undertaken and will provide social care policy makers with a scientifically robust mechanism for measuring the impact of social care programmes on the health and wellbeing of the population receiving them.

5.1.3. Metadata standards
As the range of datasets within SAIL has continued to expand, researchers are now increasingly requesting access to a wider range of datasets for their projects. This poses a challenge relating to the quality and characteristics of the data since when data from various sources is being compared or combined it is important to understand the quality and reliability of each dataset. Until recently there has been very little available consolidated knowledge about the routine data produced by health and other public services when considered for research. Lack of understanding about the characteristics of data (such as completeness, error rates, formal and informal coding conventions, influences of the historical context of its origination) is a major cause of confusion about applicability of the data, assessment of potential and analytical error, and delay.

If an RDA is used to load and process data (which is now the case with SAIL and will increasingly become the norm for other data providers as RDAs are more widely deployed) information about the data can be captured and recorded as part of the loading process. Since 2015 the SAIL team has been working to turn this information into a standardised metadata catalogue which provides information to researchers about the quality and reliability of the datasets they are using. During 2016 – 17 we have built upon our initial work and can now (using the RDA) collect a range of information about a dataset during the data loading process, including coverage, collection methods, known issues and data administrators.

5.1.4. New dataset acquisition
Continuing to expand the range of data available within SAIL is a core objective for the team, and the addition of relevant non-health data sets is becoming an increasing priority to enable more research into the effects of socio-economic and lifestyle factors (e.g. employment status, housing, access to open space, levels of exercise etc.) on levels of wellbeing within the population.

During the 2016 – 17 period SAIL has acquired a number of new datasets including:

- Outpatient Referral Data Set, Critical Care Data Set, Cancelled Admitted Procedures Data Set, Diagnostic and Therapy Services Waiting Times, Referral to Treatment Times and Ethnicity Data from NWIS.
- Lifelong learning Wales Record – Welsh Government

The work described in our 2015 – 16 Annual Report to acquire Laboratory Information Management System (LIMS), the Cancer Network Information System Cymru (CaNISC) and Radiology data from the RADIS system into SAIL is still ongoing. This has been a difficult piece of work due to the number of organisations involved and the complexity of the permissions required. A Heads of Agreement for these datasets is now in place with Health Board Caldicott guardians and work is ongoing with NWIS.
to put in place the enabling technology to allow the data to begin to flow into SAIL. Acquiring these three key datasets is of significant importance to a number of researchers using SAIL as it would allow diagnostic test results to be included into health records, making these substantially more complete.

5.1.5. **Narrative data**
There is a wealth of information about patient diagnosis, treatment and outcomes stored within narrative data records (typically clinicians discharge letters and similar), and until recently this has been inaccessible to SAIL because of the difficulty of anonymising the data (narrative data may contain names, ages and other identifiable information anywhere in the body of the document), and the further challenge of turning the information within the document into a standardised set of clinical codes and terms which can then be linked and compared to other SAIL data. Unlocking narrative data has been a significant area of focus for us since 2015 and we described two pilot projects we had undertaken in this field in our 2015–16 Annual Report. Since then we have continued to work in this area, with some notable achievements as described below.

As part of the proof of concept work for a major new project called RITREVA (which aims to identify patients for clinical trials more quickly and efficiently), software has been deployed in Abertawe and Bro Morgannwg University Health Board (ABMU) to find suitable patients with Type 2 Diabetes to support a clinical trial. A new method for stripping identifiable data from anywhere within a free text document was developed by the team, and used to successfully identify a significant number of individuals who met the trial exclusion/inclusion criteria. Further pilots of the software in this context are currently ongoing.

The SAIL team have also now developed and been testing an algorithm to match free text to clinical codes and terms using narrative data from Epilepsy clinics. Using this algorithm, we are now able to accurately extract 10 key items of epilepsy information (e.g. clinic date, epilepsy subtype, drug information, investigation information) from real, unstructured epilepsy clinic letters from different consultants. The team has now also received agreement from ABMU to receive Cardiology data, so that we can begin the process of developing the required algorithms and annotations required to extract meaningful information in a standardised form from narrative data about patients with cardiac conditions.

This exciting area of development by the SAIL team was presented at the IPDLN conference in August 2016 and attracted international recognition and acclaim.

5.2. **Work Package 2- Data Linkage methods and management**
When SAIL was originally launched, it was extremely difficult to link and compare health datasets, due to:

- lack of technologies or standards to anonymise data and safeguard privacy,
- the difficulties in linking datasets with different structures and nomenclatures
lack of technologies / standards for storage, linking and sharing of data.

Over the ten years of its existence, SAIL has developed a suite of innovative solutions in all these areas, often leading the way at national and international level. However as more and different data becomes available, new challenges to continue to become apparent. The purpose of this area of work is to continue to develop novel methods and techniques to link, manage and utilise large-scale datasets and to provide robust, assured processes for the safeguarding of individual linked data for research.

During 2016 – 17 we have continued to make significant progress in this area:

5.2.1. **New national linkage and anonymisation service**
A key challenge in working with anonymised data is to identify other anonymised records belonging to the same individual in order to link them together. Our 2015 – 16 report described how, in partnership with colleagues in Curtin University, Western Australia, we were developing a new and improved solution for the anonymization and linking of datasets which offered improved record matching by using a technology called Bloom Filtering. During the 2016 – 17 period we have completed and tested this work and hope to roll out the technology during 2018 – 19.

5.2.2. **Transition to UKSeRP**
Our 2015 – 16 Annual Report described our work to create a suite of technology called the UK Secure eResearch Platform (UKSeRP), funded by the MRC and the Farr Institute. This work was originally undertaken to meet an increasingly widespread need within the research community to combine and link large-scale data owned by a variety of owners in a way which allowed those that are responsible for the data (usually not ourselves) to accumulate, manage and share the data in a trusted, highly secure, IG-compliant manner.

During 2016 – 17 the UKSeRP technologies have completed a full testing programme and have been released for use in a live environment. The SAIL gateway has now fully moved onto a UKSeRP infrastructure. We have completed a three-year project with UK Biobank to implement their UKSeRP platform which is now fully operational. Dementia Platform UK also use UKSeRP technology to support their platform. We expect to conclude contract negotiations with Avon Longitudinal Study of Parents and Children (ALSPAC) for use of a UKSERP by March 2018.

5.2.3. **Further development of the SAIL Information Governance Review Panel**
Our 2015 – 16 Annual Report described how a sustained increased over several years in numbers of researchers applying to SAIL for access to data had led to projects taking longer to be reviewed by our Information Governance Review Panel (IGRP). We addressed this by implementing a new and improved operating model for the IGRP with published timescales for getting a response from the panel. Membership of the IGRP was increased to lessen the burden on reviewers, and a Chair was appointed to streamline some aspects of the process.
During 2016 – 17 we implemented a ‘Stop the Clock’ mechanism to measure the compliance of each IGRP application with the timescales set as part of the new operating model. For the last twelve months the IGRP has consistently achieved its target of reaching a decision on applications within 42 days or less. This significantly surpasses the speed of any other similar data linkage system in the world that we are aware of.

5.2.4. Geospatial analyses in privacy constrained environments
During 2015 – 16 a significant amount of work was undertaken to develop Residential Anonymous Linking Field (RALF) technology in collaboration with colleagues from the Geographic Information System team at Swansea University. This allows researchers to identify and group the anonymised records of people who live in the same household (important for research examining the effect of home environments on individuals, for example the health impact on children of growing up with a smoker in the household).

This technology is now fully implemented within SAIL and is in regular use by researchers. Work to refine and improve the technology based on researcher feedback is ongoing.

5.2.5. National and international accreditation and certification
Compliance with best practice in all aspects of data security is key to maintaining the reputation of SAIL and its ability to engage with and negotiate data sharing agreements with owners of complex and sensitive public sector data.

ISO27001 accreditation was awarded to SAIL on the 16th October 2015 and to UKSeRP on the 26th November 2015. During 2016 – 17 both SAIL and UKSeRP received approval from NHS Digital as environments safe to hold NHS England data, and HES and related datasets for a number of individual projects have already been transferred under this agreement.

5.3. Work Package 3 - Supporting Health and Social Care Research
This area of work is focussed upon meeting the needs of current and future users of SAIL. Its purpose is twofold; firstly to provide the most effective services possible to researchers who already use SAIL, and secondly to extend the use of SAIL to new researchers and research organisations to support health and social care research.

A summary of how we have worked during 2016 – 17 to support and extend our user base is set out below.

5.3.1. Developing the SAIL user community
SAIL has an established relationship with the research community within Wales, the UK and internationally with over 370 users from 78 organisations worldwide (see figure 1 below). On average during 2016 – 17 the SAIL team were engaged in providing support to a portfolio of around 100 live projects whilst simultaneously managing and processing an average of 75 new enquiries at various stages of the application and project set up process. We received an average of five new enquiries every week from researchers interested in using SAIL data.
We have an established programme to help us stay connected to and support our existing user community. We hold remotely accessible bi-monthly forums and events for SAIL users, where researchers share information about their projects, or more technical elements of their work such as SQL coding, definition of conditions, using specialist elements within SAIL data such as Residential Anonymous Linking Fields (RALFs). Data analysts are encouraged to share their codes and definitions with other members of the SAIL user community on a specially created wiki to make best use of innovation and to encourage development of consistent research methods.

5.3.2. User training and capacity building

A key challenge within the informatics / data analysis community within Wales and more widely across the UK is the severe shortage of people suitable to undertake population data research. There are very few people with the required combination of technical and analytic skills combined with knowledge and understanding of health or other public sector services, and the datasets they generate. This limits the amount of research which can be carried out and potentially stops the full potential for benefit from SAIL being realised.

In order to address this skills gap, a new postgraduate masters course (MSc in Data Science) was developed and launched by Swansea University in 2014, and a series of Data Analysis modules were added to the existing MSc in Health Informatics in 2015. Both courses liaise closely with the SAIL Databank team. In 2016 funding was obtained by Professor Ronan Lyons from the Health and Care Research Wales Senior Fellowship award to train NHS analysts on MSc Health Data Science. Six NHS analysts are currently on the MSc Health Data Science course.
5.3.3. Collaboration with Health and Care Research Wales Centres and Units
SAIL has well established relationships with most of the Centres and Units within the Health and Care Research Wales network, and supports an ongoing portfolio of research activities with them. We are currently involved in a study led by PRIME Centre Wales called Transient Ischaemic Attack 999 Emergency Referral (TIER) and a strategic project led by Ann John’s team called Children and Young People with Adolescent Mental Health (AMH) problems. Two projects from the National Centre for Population Health and Wellbeing Research (NCPHWR) are also about to conclude; ‘ACTIVE (Active Communities Through Individual Vouchers - Evaluation)’ and ‘Surveillance using routine data (Pharmacoepidemiological surveillance for children with chronic conditions using data mining)’

During 2016 – 17 we have been engaged in supporting a number of projects other Health and Care Research Wales Centres and Units. Some examples of these include ‘Multivariate logistic regression analysis of factors relating to lung cancer survival and treatment’, ‘Linking Bacterial Genome and Patient Records Pilot’, ‘Living Well Living Longer Evaluation’ ‘Epidemiology of dental procedures carried out under general anaesthetic in children’ and the ‘ARRISA Asthma project’.

Our ongoing collaboration with the National Centre for Population Health and Wellbeing Research continues to produce a portfolio of successful projects including ‘Creating a profile of population-based stage at diagnosis by cancer type and geography in Wales’, ‘Unravelling polypharmacy: Mining the complex interaction patterns between medications for enhanced patient care’.

We are also collaborating with members of the Diabetes Research Unit Cymru on a number of projects, investigating type 1 diabetes and pregnancy, the impact of type 1 diabetes on educational outcomes, and type 1 diabetes prodrome.

5.3.4. Collaboration with the NHS
Our 2015 – 16 Annual Report described the successful project to establish the Prudent Healthcare Intelligence Hub (PHIH), to facilitate collaboration between NHS organisations and the SAIL Databank. Throughout 2016 – 17 PHIH has continued to grow and has developed formal collaborations with a number of organisations including Abertawe Bro Morgannwg University (ABMU) Health Board, Welsh Cancer Intelligence and Surveillance Unit (WCISU), and A Regional Collaboration for Health (ARCH). The Hub is now supporting 29 NHS projects across several organisations including:

- ABMU - a series of pilot projects using ABMU pathology data on Primary Biliary Cholangitis (PBC), and Stroke Prevention through targeted interventions for unknown AF population and the unmet need for anticoagulation
- Aneurin Bevan and Cwm Taf UHBs - Living Well Living Longer
- Cwm Taf UHB, Utilisation of variation in hospital stays to measure the effectiveness of the management of conditions in Primary Care
- Public Health Wales, WCISU (An evaluation of the 2015 Lung Cancer Initiative Campaign),
• ARCH (A regional wellbeing and health needs assessment to inform the ARCH case for change, being undertaken by SAIL and colleagues within Hywel Dda and ABMU Health Boards).

This activity has also facilitated the enrolment of NHS professionals on the MSc Health Data Science. Since the start of the project, 6 NHS staff from Health Boards across Wales have taken part in the MSc course which contributes to building capacity in the NHS workforce. Professor Ronan Lyons is also working closely with several leading clinicians and other NHS staff within Health Boards, WCICSU and Public Health Wales.

5.3.5. Collaboration with Social Care Researchers

SAIL is co-located and works closely with the Administrative Data Research Centre Wales (ADRC-Wales) who are engaged with the Welsh Government Knowledge and Analytical Services (KAS) team and the Chief Statistician for Wales on a number of social care themed research projects. During 2016 – 17 the ADRC and Welsh Government have been engaged in a social research programme (using SAIL and other data sources) to evaluate the effectiveness of key social care intervention programmes including:

- Fuel Poverty: A WG funded programme to provide energy saving/efficiency measures to those living on low incomes who are at risk of being unable to afford sufficient fuel for their living requirements.
- Supporting People: A WG funded programme providing housing related support to vulnerable people to support independent living
- Flying Start: A WG funded programme to support families living in poverty with infant children through education/healthcare support

It is anticipated that this extremely valuable collaboration with the ADRC and Welsh Government will continue into the future, with more research and service evaluation projects being undertaken on an ongoing basis.

5.3.6. Collaboration with Third Sector Organisations

MS Register: SAIL continues to host the MS Register, a nationally acclaimed patient portal designed to collect and hold information about patients with Multiple Sclerosis. The information links together patient provided data from structured questionnaires with health records. The MS Register is the result of a long standing collaboration with the UK MS Society, and during 2016 – 17 the collaboration has continued with an ongoing programme to maintain and develop the patient portal.

UK BioBank: UK SeRP technology is used to host the UK Biobank, a major national and international health resource, and a registered charity which has the aim of improving the prevention, diagnosis and treatment of a wide range of serious and life-threatening illnesses – including cancer, heart diseases, stroke, diabetes, arthritis, osteoporosis, eye disorders, depression and forms of dementia. The move to a
UKSeRP platform in 2016 marks the latest phase in a long-standing collaboration between SAIL and UK BioBank, who have been conducting joint health data related research projects for a number of years.

Funding via third sector sources: SAIL also engages indirectly with a number of third sector organisations who are a key source of funding for researchers, who then collaborate with us to use SAIL data. Examples of projects of this type from the 2016 – 17 period include ‘Creating a profile of population-based stage at diagnosis by cancer type and geography in Wales’ which has been funded by Cancer Research UK, and ‘Use of administrative data in developing direct measures of social capital: An exploratory study.’ which utilises data supplied by the Welsh Council for Voluntary Action (WCVA).

5.3.7. Collaborations with Commercial and Industrial Organisations

SAIL data is of considerable interest to the commercial sector, particularly pharmaceutical companies. Whilst SAIL has always been open to collaborations with industry partners, a number of factors have previously made it challenging to actually initiate a project. These factors have included:

- a lack of understanding by commercial organisations about the purposes for which SAIL exists, particularly the underpinning principals that any access to the data must be for research with potential to create public benefit and
- anxieties upon the part of our data providers and the public about commercial use of personal health data, particularly in the period following the negative press coverage of the care.data scheme in England.

During the 2016 – 17 period we have succeeded to a large extent in addressing those issues by clearer statement of the terms under which commercial companies can engage with SAIL (i.e. that the project must be genuine research with the potential to create public benefit, not simple market research). We have also introduced a policy that any company with a potential commercial interest in the content of SAIL data is not allowed direct access, but instead must commission any research from SAIL analysts or other similar independent third party analysts. This has removed the risk that a company may use the data for purposes which are not in the public interest, and has allayed to a significant extent the concerns of the IGRP members reviewing commercial project applications. As a result we have been able to work on a number of projects with commercial partners during the 2016 – 17 period, including:

- Janssen-Cilag Ltd - An ongoing collaboration between the SAIL Databank and Janssen-Cilag Ltd. Investigating type 2 diabetes care and outcomes in Wales, focusing on questions such as how prescribing practice in Wales and the relationship between diabetes and deprivation. Initial results were reported to the Bevan Commission to inform their recommendations to government about health policy in 2015, and the company has commissioned further follow up work in the same research area. This follow up project is complete and awaiting publication.
Clinithink Ltd / IBM – We have formed ongoing collaborations with both these companies, and continue to work with them on a number of projects to investigate the potential applications of these natural language processing software products.

Dementias Platform UK - Dementias Platform UK is a collaboration between academic and commercial partners, working together to speed up dementias research. The DPUK Data Portal (a data storage, analysis and linkage platform which provides the research resource for the community) is hosted on an instance of UP SeRP technology, developed and implemented by the SAIL team during 2016.

We are currently discussing collaborations with a number of other commercial partners including a portfolio of research projects using SAIL data with Bristol-Myers Squibb Pharmaceuticals Ltd and Boehringer Ingelheim Ltd.

5.3.8. Improving public involvement and engagement with SAIL

Whilst members of the public do not use SAIL data or services directly, we have always been keen to increase public awareness of and support for the SAIL Databank. During 2016 – 17 we have undertaken a number of public engagement activities:

SAIL continues to host a long-standing Consumer Panel which consists of members of the public who actively engage with and provide advice to researchers who utilise data held in the SAIL databank. It currently has 16 members with on-going recruitment. The panel meet on a quarterly basis for a full day meeting; researchers are able to book a session with the panel during this day. Some of the panel’s duties include;

- Acting as advisors on issues in research
- Advising on how best to engage with the public
- Offering guidance on how to recruit people to study steering groups
- Providing views on data protection issues
- Discussing proposals for research
- Reviewing information designed for a lay audience
- Acting as advocates for data linkage research

In 2016, we took part in the British Science Festival and presented the work of the SAIL Databank to the general 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.

Representatives of the SAIL Team attended the Carmarthenshire 50+ Forum annual event, which in 2016 attracted in excess of 400 residents from across Carmarthenshire to explore a range of the health and wellbeing initiatives currently active in the county. SAIL presented an interactive exhibit making use of the well-
known game, Connect 4, offering participants the opportunity to learn more about the data linkage process and the research being conducted.

An interactive exhibit on the SAIL Databank formed part of the ‘Techniquest After Hours Does Cancer Research’ a unique event hosted by Cancer Research UK. The aim of the event was to introduce Welsh adults to the research being conducted in Wales around cancer. Over 400 people attended the event from all across South Wales, with many visiting the SAIL exhibit to find out more about the research which is being conducted using the Welsh population’s data, and how they are contributing to this cutting edge research.

Female researchers and analysts from the SAIL team conducted a master class session; ‘the Spread of Disease’, as part of the 2017 Swansea University Mathematics Masterclasses for Girls programme, aimed at Year Ten female school pupils. The session focused on how mathematics is used to measure the frequency and spread of diseases by using probabilities associated with the risk of infection.

A new and improved SAIL website was launched in March 2017. This offers current and potential users of SAIL a clearer overview of the Database, datasets available and the application process. It also offers a variety of information about SAIL to members of the public, including the aims and mission of SAIL, how data is collected and de-identified and examples of the types of research undertaken.

6. **Performance Indictors for 2016 – 17**

In addition to the three work streams described above, the SAIL team has been working during 2016 – 17 towards achievement of the three Key Performance Indicator Targets agreed with Health and Care Research Wales as a condition of our funding. We have until 31st March 2018 to achieve these targets in full.

6.1. **Whole Project Key Performance Indicator 1:** Increase the research income generated through attracting high quality research to SAIL by 10% to a total figure of £30,800,796

We have made good progress to date and expect to achieve our target by March 2018. The figures and graph below provide a breakdown of progress.

- Research income generated in Quarter 1 (Apr – Jun 2015): £1,929,393
- Research income generated in Quarter 2 (Jul – Sep 2015): £1,620,007
- Research income generated in Quarter 4 (Jan – Mar 2016): £2,096,228.00
- Research income generated in Quarter 5 (Apr – Jun 2016): £2,750,833.40
- Research income generated in Quarter 6 (Jul – Sep 2016): £5,053,357
- Research income generated in Quarter 7 (Oct – Dec 2016): £8,300,587
- Research income generated in Quarter 8 (Jan – Mar 2017): £3,742,302
- Total research income generated to date: £28,104,855
6.2. **Whole Project Key Performance Indicator 2**: Enhance the portfolio of datasets available to SAIL users by increasing the numbers of datasets available on SAIL. Specific target is to engage with five new Data Providers and bring in 10 new datasets to SAIL by April 2018.

To date we have added 8 new datasets. These are Outpatient Referral Data Set, Critical Care Data Set, Cancelled Admitted Procedures Data Set, Diagnostic and Therapy Services Waiting Times, Referral to Treatment Times from NWIS, Active Adult Survey, School Sport Survey from Sports Wales and the Lifelong learning Wales Record from Welsh Government.

We have signed collaboration agreements with four new Data providers (Sports Wales, the Brecon Group, ME Epidemiology Genomic Alliance (a partnership of academic and commercial researchers and The Office for National Statistics). We anticipate that these new partner organisations will bring in a number of further datasets to SAIL as we begin to undertake projects with them into the future.

6.3. **Whole Project Key Performance Indicator 3**: NHS staff made up 10% of the SAIL user base in April 2015. Specific target is to increase this to 20% by April 2018.

To date we have increased the percentage of NHS users of SAIL to 16% and we expect to achieve our target within the next 12 months.

7. **Key Achievements**

As described above the SAIL team continue to make excellent progress in all areas of our work programme. The following highlights are considered to be particular achievements.
7.1. Our continued engagement with and expansion of the SAIL user base, with new collaborations with the NHS, social care researchers and the commercial sector being particular highlights of the 2016 – 17 period.

7.2. Continuing to realise the potential of SAIL as a resource to leverage additional research funding into Wales. We are on course to achieve our figure of £30,800,796 awarded to projects using SAIL data, which will represent a 10% increase from the level of research income attracted during the 2012-15 period. Obviously SAIL only receives a tiny fraction of this funding, with the majority going to the research teams undertaking the projects, but the total funding secured demonstrates the benefit SAIL provides to the research community and the levels of research it supports.

7.3. SAIL projects resulted in a number of key publications during the 2016 – 17 period, which contributed to the body of research knowledge and helped to inform health and social care policy makers. Brief details of these publications are given in Appendix 1 attached.

8. Impact of SAIL projects in the real world

It often takes many years for the results of research to be translated into a tangible impact on public services. SAIL will continue to assess all projects it supports to ensure that there is a realistic prospect of an eventual public benefit. In the meantime, we present several examples of projects which have used SAIL data and gone on to have a tangible ‘real world’ impact during 2016 - 17:

8.1. The Suicide Information Database-Wales or SID-Cymru anonymously links electronic routinely collected information on all those who complete suicide within the population of Wales via the SAIL Databank. Data from SID-Cymru has informed ‘Talk to Me 2’, the latest suicide prevention strategy for Wales, which was launched in June 2015 and was implemented as active policy during 2016.

8.2. The ‘Exploratory study using data linkage to examine use of the Primary Care Palliative Care Register’ was a primary care led study. This project found that patients who had a terminal condition other than cancer were significantly under-represented on the Register and were potentially missing out on the chance to receive their choice of palliative care – particularly the chance to die at home if they wished to do so. The project led to several GP practices modifying the way they record data to make sure that a wider range of palliative care patients are captured on the Register and can arrange and achieve end of life plans in accordance with their wishes.

8.3. Dementia Platform UK. The MRC’s Dementias Platform UK (DPUK), part of a £53 million public/private investment by the MRC, technology and pharmaceutical companies is now hosted upon a UK SeRP platform within Swansea University Data Science Building. This collaborative infrastructure is seen as world leading in this field
and a number of European and North American cohorts have expressed interest in joining the DPUK initiative.

8.4. **UK Biobank.** UKSeRP is now the underpinning infrastructure used to support multi-modal linked UK Biobank data to support the working of a distributed group of research professionals accessing this national data resource.

8.5. **Millennium Cohort Study (MCS).** UKSeRP is being used to provide access to linked data from MCS sweeps 1-5, physical activity (accelerometer) and oral fluids data, hospital admission data from England, Wales and Scotland for a Wellcome Trust funded study involving the Swansea and London Farr nodes “Using health record linkage in the UK MCS to investigate childhood obesity, asthma, infections and immunisations in early life”.

8.6. **GeDI Project.** A collaboration between the SAIL Databank and The 100,000 Genome programme led by Genomics England (GEL) to identify issues pertaining to the integration of genetic data into data safe havens, and its linkage to anonymised, routinely-collected health and social care data for use in research. The findings are being used to inform the working relationship between the Welsh Gene Park, the Wales GMC and the SAIL databank, and also have general relevance to other research groups. The guidance on linking genetic data with SAIL datasets, and general principles for working with genetic data in safe havens produced by the project will form part of a report from the Wales GMC to the MRC.

9. **Conclusion**

9.1. Since its establishment in 2007, the SAIL Databank has witnessed a sustained period of rapid development in the use of informatics and big data as a research resource, with new paradigms, technologies and governance frameworks emerging, often as a result of work led by the SAIL team. This has resulted in international recognition for SAIL as an organisation at the leading edge of data-intensive research and the SAIL team are committed to maintain this position into the future.

9.2. As this report shows, the 2016 – 17 period has continued to be one of rapid development, particularly in terms of emerging data types, notably, genetic and free-text datasets as well as imaging, accelerometers (for ‘actigraphy’ data on physical activity) and smartphone data. We continue to work with partner organisations to identify the information governance and other challenges pertinent to genetic, imaging and mobile phone data, and to identify appropriate technical and methodological tools to enable their safe, efficient and acceptable use as resources to the research community. Previous work on usage of free text data has begun to bear fruit during 2016 - 17, with successful pilot projects to use narrative data from clinicians’ letters being undertaken, and the MS Register continues to make a major contribution to revolutionising understanding of MS in the UK.
9.3. 2016 – 17 has also been notable for the success we have made in forging relationships with the NHS and with research groups undertaking projects at the intersection between health and social care data. The success of the Prudent Healthcare Intelligence Hub and the increasing levels of support being provided to the research programmes undertaken by the Administrative Data Research Centre and the Welsh Government Knowledge and Analytical Services team are examples of this, as is our developing programme of work with organisations in the private sector.

9.4. We have continued to face challenges in obtaining some key datasets, particularly Laboratory Information Management System (LIMS), the Cancer Network Information System Cymru (CaNISC) and Radiology data from the RADIS system. Another challenging area of work has been rollout of the Research Data Appliances into the NHS IT infrastructure. Whilst progress has been made on both these areas during the last year, we had hoped to have achieved more by this stage.

9.5. Reflecting on the success and challenges of 2016 – 17, our key conclusion would be that pretty much every development in the field of data science takes longer, more resources, more energy (and also patience), a wider range of expertise, and is more complicated in practice than in theory! However if the first ten years of the SAIL Databank have taught us anything, it must be that the work of bringing together data for research will never be finished, and there will always be challenges with each area of development we undertake. We believe that the SAIL team are uniquely well qualified to meet these challenges and 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.
Appendix 1 - Publications by Projects Using SAIL Data

<table>
<thead>
<tr>
<th>Title</th>
<th>Year of publication</th>
<th>Journal</th>
<th>Journal impact factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fone, D., Morgan, J., Fry, R., Rodgers, S., Orford, S., Farewell, D., ... &amp; Brennan, I. (2016). Alcohol outlet density and harm to population health: literature review.</td>
<td>2016</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>Title</td>
<td>Year of publication</td>
<td>Journal</td>
<td>Journal impact factor</td>
</tr>
<tr>
<td>-------</td>
<td>---------------------</td>
<td>---------</td>
<td>-----------------------</td>
</tr>
<tr>
<td>Title</td>
<td>Year of publication</td>
<td>Journal</td>
<td>Journal impact factor</td>
</tr>
<tr>
<td>----------------------------------------------------------------------</td>
<td>---------------------</td>
<td>--------------------------------------------------</td>
<td>-----------------------</td>
</tr>
<tr>
<td>Title</td>
<td>Year of publication</td>
<td>Journal</td>
<td>Journal impact factor</td>
</tr>
<tr>
<td>----------------------------------------------------------------------</td>
<td>---------------------</td>
<td>--------------------------</td>
<td>-----------------------</td>
</tr>
<tr>
<td>Title</td>
<td>Year of publication</td>
<td>Journal</td>
<td>Journal impact factor</td>
</tr>
<tr>
<td>-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>---------------------</td>
<td>------------------------------</td>
<td>-----------------------</td>
</tr>
</tbody>
</table>