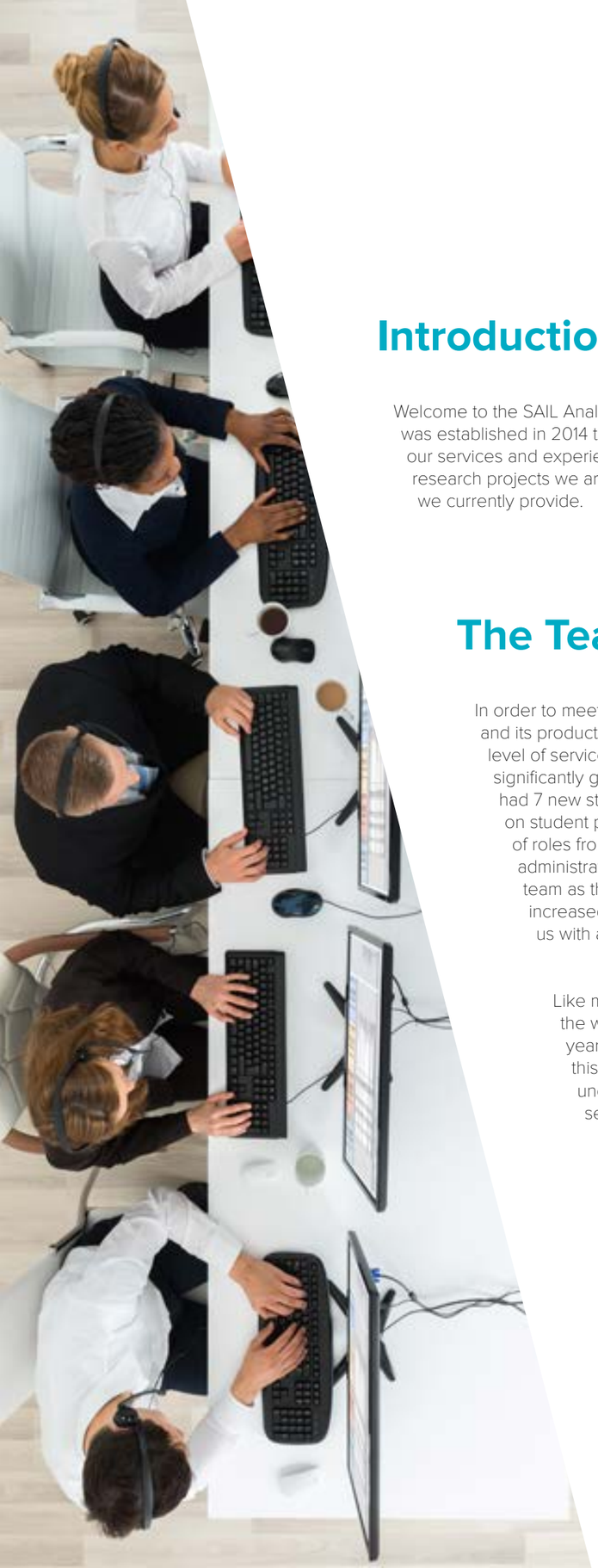
SAIL DATABANK

# Analytical Services Team

Annual Report 2020



# Introduction

Welcome to the SAIL Analytical Services first annual report. The team was established in 2014 to support SAIL Databank users, over time our services and experience has grown. This report highlights some research projects we are supporting and summarises the services that we currently provide.

# The Team

In order to meet the increased demand in SAIL Databank and its products and services offered as well as maintain the level of service expected, the Analytical Services team has significantly grown in size over the last year or so. We have had 7 new starters, 2 of which joined the team originally on student placements. The new recruits cover a variety of roles from developer to research assistant to a team administrator and there are plans to further increase the team as the demand for development led projects has increased recently. The variety of skill sets has provided us with an adaptable team.

Like most organisations, COVID-19 has resulted in the whole team working from home for the last year. The team members have adapted well to this, maintaining a team approach to the work undertaken and with no impact to the level of service we offer.

We are currently investing in the creation of a shared team website which we aim to use to increase the visibility of the team and promote the services we can offer to researchers.




# COVID-19 Impact

**The COVID-19 pandemic has become a major health research event which, due to focus from Welsh Government and other organisations, has resulted in increased collaborations with different organisations as well as increased opportunities.**

**1** There as been an increased frequency of data updates of certain key datasets, e.g. Welsh Demographic Service Dataset (WDSD) has moved from monthly to weekly updates, Welsh Longitudinal GP (WLGP) has recently moved from quarterly to monthly.

**2** The overall breadth of data now offered has increased with new datasets being added on an ongoing basis, many of these being refreshed on a daily basis, e.g. COVID Test Results & COVID Vaccination data.

**3** Collaborations with organisations studying and collecting data related to COVID, e.g. ZOE App data.

**4** The overall volume of work has significantly increased across all aspects of the team services: Scoping, IGRP reviews, Project provisioning, File Out Reviews, Data Quality checking & Metadata curation.

**5** Since March 2020 we have had 27 new datasets introduced into the SAIL Databank

# Research and Development

The SAIL Analytical Services team is multidisciplinary, with a wide range of skills relevant to research, and is delivering health data research outputs in various capacities. Team members are leading and contributing to projects in several areas, including methodological and clinical research topics. Examples of research programmes and projects and led or supported by the team are listed below: (N.B. this is not all projects that the team are supporting)

## Projects: A Summary of Involvement

### EUROLinkCAT
This is a Europe-wide project aimed at establishing a linked cohort of children with congenital anomalies. SAIL is the lead organisation for Wales. Our team transforms the SAIL data using reusable research methodologies so that it is ready for research. https://www.eurolinkcat.eu/

### ConcePTION
This study involves 20 European nations, and is examining the safety of medicines use pre-conception, and during pregnancy and breastfeeding. Building on the knowledge and methodologies developed in EUROLinkCAT. SAIL is the lead organisation for Wales. https://www.imi-conception.eu/

### BREATHE
This programme is delivering a research hub facilitating research into respiratory health. Our team is leading the collation of metadata and data provision.
https://popdatasci.swan.ac.uk/centres-of-excellence/breathe/

### UK-REACH
The SAIL team have a significant supporting role in the delivery of this project, which is examining the impact of COVID-19 on healthcare workers, depending on their ethnicity. https://uk-reach.org/main/

### ADP
The SAIL team are supporting the Adolescent Mental Health Data Platform which is a programme to develop a specialist research environment focusing on the mental health of children and young people.
https://popdatasci.swan.ac.uk/centres-of-excellence/adolescent-mental-health-data-platform/

### The Family Justice Data Partnership
The SAIL team are providing analytical support to this programme of work, examining outcomes for children who have been involved in family proceedings.
https://popdatasci.swan.ac.uk/centres-of-excellence/family-justice-data-partnership/



## Projects cont.

### ADR
The SAIL team are supporting the Administrative Data Research Wales research themes to maximise the utility of anonymous and secure data to shape public service delivery, which will ultimately improve the lives of people in Wales.
https://popdatasci.swan.ac.uk/centres-of-excellence/administrative-data-research-wales/

### RECOVERY
The SAIL team are providing fortnightly data updates of the individuals recruited to the RECOVERY (Randomised Evaluation of COVD-19 Therapy) trial "recruited over 28,000 patients with COVID-19 from 176 hospitals around the UK and is the largest study in the world to test treatments for people admitted to hospital with COVID 19. It is identifying which treatments are effective and which are not." https://www.recoverytrial.net/

### UK Biobank
The SAIL team are providing monthly data updates of the individuals recruited to the UK Biobank.
https://www.ukbiobank.ac.uk/

### Millennium Study Cohort
The SAIL team have provided an update of the individuals recruited to the Millennium Study Cohort.
https://cls.ucl.ac.uk/cls-studies/millennium-cohort-study/

### GenOMICC – COVID-19
The SAIL team will be providing monthly data updates of the individuals recruited to the GenOMICC COVID-19 study. https://covid.genomicc.org/

### HealthWise Wales
The SAIL team have a significant supporting role in the delivery of this project, including linking the study data to SAIL datasets, provisioning data and reviewing file out requests.
https://www.healthwisewales.gov.wales/

### SIMPLIFIED registry trial
The SAIL team are providing biannual data updates of the individuals recruited to the SIMPLIFIED registry trial.
https://cctu.org.uk/portfolio/core/trials-open-to-recruitment/the-simplified-registry-trial

### TIME (Treatment in Morning vs Evening)
The SAIL team are providing biannual data updates of the individuals recruited to the TIME study.
https://www.isrctn.com/ISRCTN18157641

# Projects cont.

The SAIL team also contribute to other studies examining various clinical and methodological topics; recent examples include acute eye care, dystonia, prostate cancer, older people's health and inflammatory bowel disorder. The team contribute in different ways according to the needs of each project; contributions include study design, data cleaning and preparation, descriptive and inferential statistical analysis and writing research reports.

**Outputs from projects led by SAIL, or where the team made a significant contribution, include:**

A methodology paper describing the measurement of follow-up time in routinely-collected health datasets (DOI: 10.1371/journal.pone.0228545. eCollection 2020)

A methodology paper describing developing a standardised approach to the aggregation of inpatient episodes into person-based spells in all specialties and psychiatric specialties (https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-019-0953-2)

A research collaboration focusing on warfarin research - "An observational study of INR control according to NICE criteria in patients with non-valvular atrial fibrillation-The SAIL Warfarin Out of Range Descriptors Study (SWORDS)" https://pubmed.ncbi.nlm.nih.gov/31774502/ & "Bleeding events associated with poor INR control. An observational study of patients prescribed warfarin for non-valvular atrial fibrillation in the Welsh population. The SAIL AF Bleeding Risk Evaluation (SABRE) study." DOI:10.13140/RG.2.2.26165.63205

A research letter describing prevalence of hidradenitis suppurativa in Wales (DOI: 10.1111/bjd.19210 https://onlinelibrary.wiley.com/doi/10.1111/bjd.19210)

A report estimating incidence and prevalence of inflammatory bowel disorder in Wales (summary here https://www.crohnsandcolitis.org.uk/news/study-shows-over-50-more-people-in-wales-have-crohns-or-colitis-than-previo)

Papers ready for submission to peer-reviewed journals, examining a specific form of prostate cancer, and dual diagnosis of mental disorder and substance misuse in children and young people.



# Development

While capacity for development was somewhat reduced by meeting the needs of the COVID-19 response, our team progressed a number of important developments to support researchers.

**In 2020, we officially launched the Concept Library, our flagship tool for managing, publishing, and sharing definitions used in research. This has been recognized as a leading tool in the domain, and adopted as a solution by both the Adolescent Mental Health Data Platform and HDR-UK. In 2021 a UK-wide solution built on the Concept Library will be released by the latter organization. We are key collaborators in this work, leading a multi-site development team.**

**We progressed research-ready datasets, through standardising and automating processing of a number of datasets, both SAIL core datasets (e.g. GP data) and derived data (standardized coverage measures). We also developed automated processing to support daily GP data feeds used for COVID-19 priority work.**

**We developed an R version of our code that creates a matched control group.**

**Ongoing development and improvment of our processes of data provision and dataset QA, to support greatly increased velocity due to COVID-19.**



# Team Services

**The SAIL Analytical Services team offer a wide range of services and support, both internally within SAIL and externally to research projects. Whilst the majority of the year has been focused on responding to the COVID situation, we took the opportunity to improve our processes to ensure a more pro-active response to the impact on our services...**

# ...as a result we've also managed to deliver:

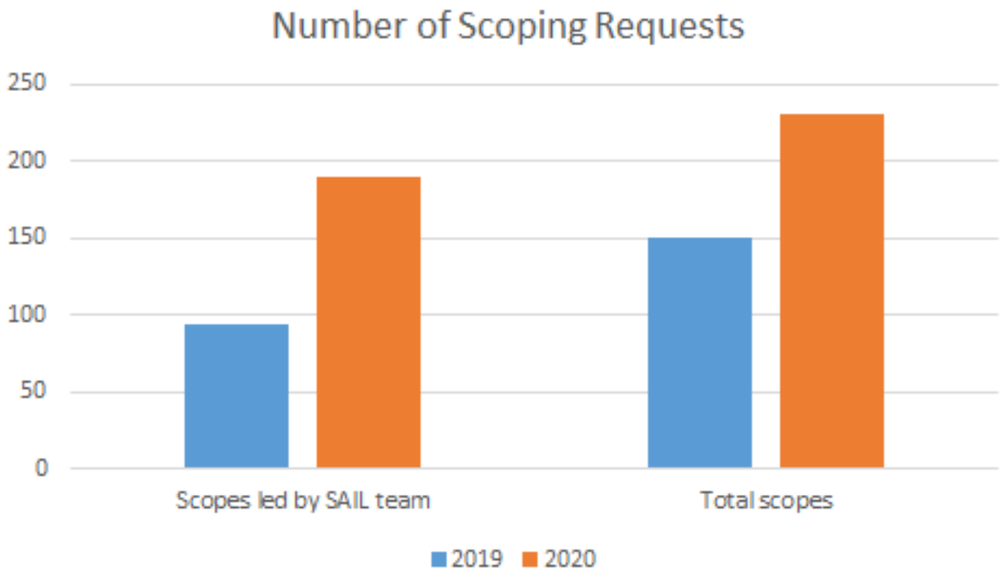
## Enquiries and Scoping

The first contact that researchers have with SAIL is generally when they approach SAIL with a project which requires scoping. Some researchers contact us with a fully developed project, which has a clear research question and method and a definitive list of datasets that are required; other projects may not have reached this stage, and require support and advice from the SAIL team to help them to refine their projects.

The SAIL team work with researchers to scope their projects, with the aim of producing a custom document which clearly sets out their requirements, alongside the associated costs. The process of scoping projects varies according to the needs of each research team, but may include providing guidance about the SAIL application process and the technical and governance aspects of working with SAIL, assessing project feasibility, providing advice about appropriate datasets and variables, assistance with developing research questions and methodological approaches, and assistance with grant applications.

Projects may require different levels of support; some bring their own team of analysts and require only provision of data, while others have no analytical support, or have no experience of working with routine data, and require a SAIL analyst to become a key part of the project team. The SAIL team is multidisciplinary, with skills in a wide range of fields relating to data science research, so we are able to tailor the support we offer to researchers according to their specific requirements. For some projects this may mean that most or all stages of the project, from study design, data preparation and analysis, to producing drafts for submission to peer-reviewed journals, are led by SAIL.

This year has seen demand for scoping by the SAIL team double, compared with the previous year.

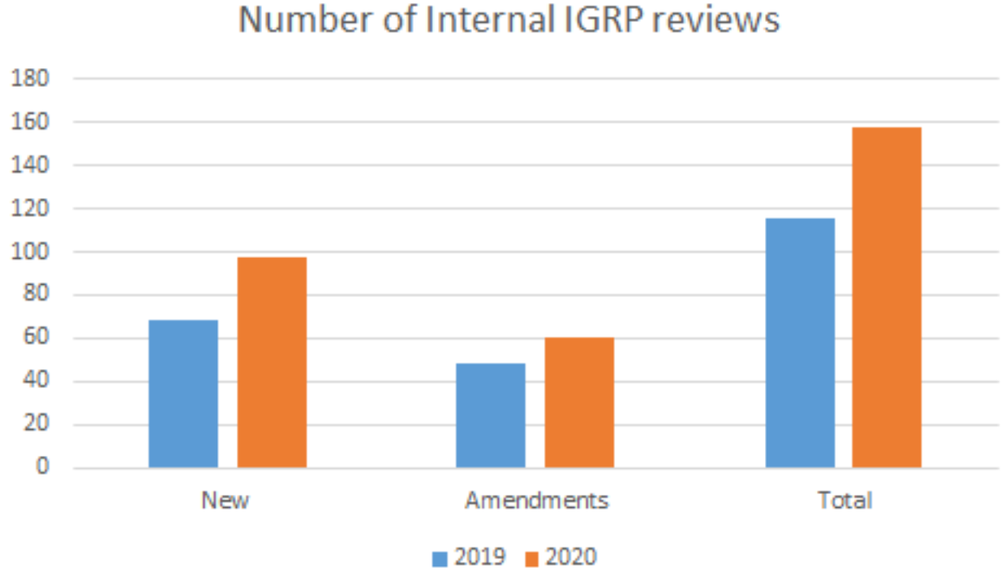



## Internal IGRP

When researchers want to use the SAIL data for their project they have to make an application to gain IGRP (Information Governance Review Panel) approval. We provide an internal service to review the application before it is passed to the Information Governance Review Panel. Our role in the review is to "consider the feasibility, practicality and resource implications of the proposal" and to ensure that the application is completed to a high quality to ease the review burden for our independent external panel.

We aim to review and provide a decision &/or feedback within 2 weeks of receipt of an application. Once our internal review is complete the application is passed to the external IGRP. The full approval process usually takes within 3 months which is considered fast in the realm of population health data science.

The number of new applications and project amendments increased 35% in 2020. Due to this increase in demand, and the urgency needed by the COVID-19 response projects, we are training three new reviewers to increase the available pool to five internal IGRP reviewers.



## Metadata

Metadata provides information about the datasets SAIL Databank is responsible for, increasing data utility. This gives potential SAIL users the ability to assess the feasibility of using SAIL Databank for their project. For existing users, this provides a resource that can help provide detail about their project data. This metadata also feeds national work such as National Core Studies programme. This metadata is powerful because it has potential to support a variety of SAIL Databank core services, such as scoping, governance and provisioning processes.

Metadata is maintained on a SAIL-specific platform and on the HDR-wide innovation Gateway. Currently, there are over 50 SAIL datasets published on these platforms, with more in the pipeline. This metadata is rich because of the networks and experience embedded in SAIL Databank, and because of this SAIL leads the quality rankings on the Innovation Gateway. Currently a project is underway to deliver an internet-facing version of the SAIL metadata catalogue.

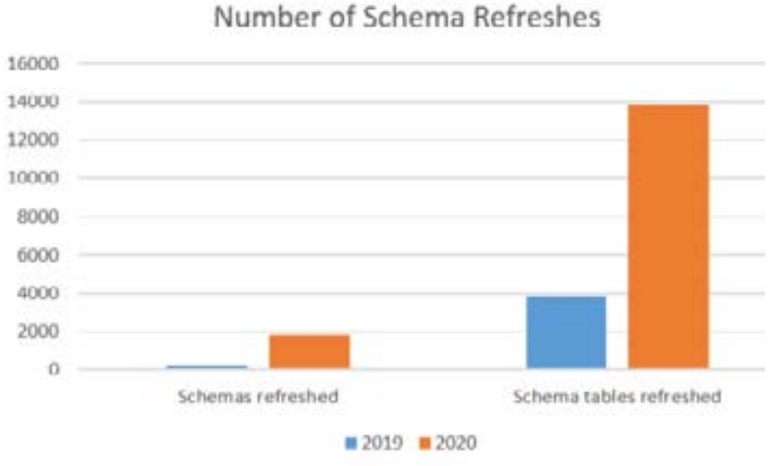
At the moment, dataset and column level descriptions are provided. Future plans to increase the quality of this metadata includes adding value level descriptions, adding governance and data provisioning specific metadata properties, and including additional summary statistics for temporal, spatial, and demographic characteristics.

## Data Quality Reviews

As data is received from data providers, we provide an internal service to review the data before it is released into the Databank. These reviews ensure that received data meets certain criteria and does not provide any elements that could be deemed as disclosive thus maintaining the integrity and reputation of the SAIL Databank.

The teams capacity to run these quality checks has increased over the last year allowing for 93% of received datasets to be reviewed and 75% released into the Databank within 2 working days. This increasingly allows projects to have access to the latest available data at point of provisioning while at the same time maintaining the expected data quality levels. This has been especially important in 2020 due to significant increase in data requiring to be quality assessed. The volume of data received from various data sources in the last year has seen a staggering increase. We started to receive daily feeds for several key COVID related datasets which was never contemplated in 2019, around 1400 daily refreshes (3200 tables) were processed. Of our other datasets there was a 75% increase in the number of refreshes in 2020 compared to 2019, resulting in a 170% in the number of tables refreshed.

N.B. A copy of the DQ reports run for datasets deemed SAIL Core or SAIL Restricted can be found on the R drive within the gateway under SAIL Reference Library/SAIL DQ Reports.





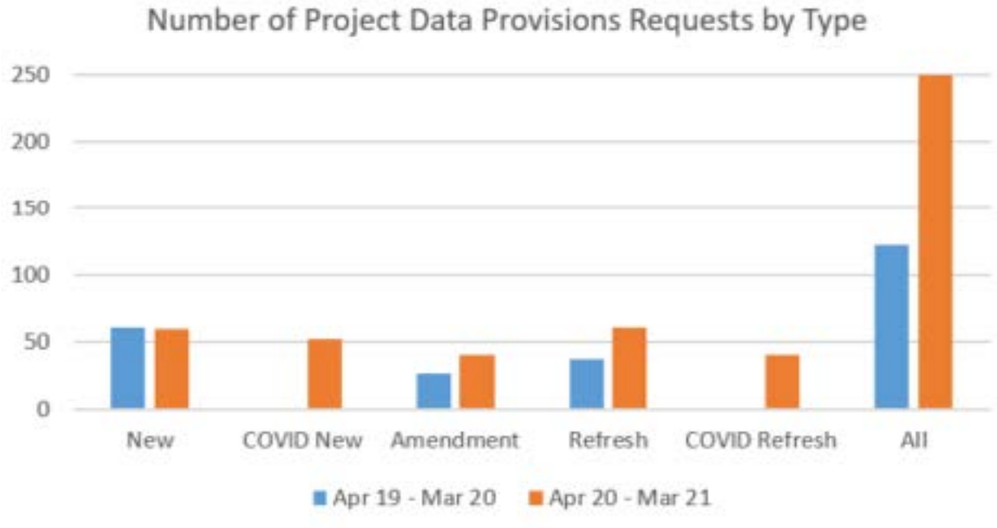

## Data Provisioning

Once a project is approved, we provide an internal service to construct the project data views before they are made available to the researchers. This process ensures that the researchers only receive the data needed and approved within their IGRP application, and adhering to the data providers requirements.

Last year we changed our data provisioning processes from using a web interface to sending API requests from R scripts. This has significantly improved our ability to respond to the increased demand and support project view data refreshes. Also, we have defined standard project view quality checks which are performed against all project views.

We aim to provision project views within 30 days of application approval, where data is available and approved. In first quarter of 2021, 98% (51/52) of the project data provisions completed were done within 30 days of approval &/or data available. The teams capacity to create project views has increased over the last year ensuring that we were responsive to the urgent COVID-19 data provision (from two to eight, four of which are on call daily to support the main COVID-19 response project 0911). This means that projects requiring access to the "COVID-19 Symptom Tracker Dataset" receive access within 2 days of approval and all new COVID-19 projects have been able to start working on data within 18 days.

2020/21 has seen a 100% increase in the number of project provisions requested compared to 2019/20 with around 75% of these new requests being COVID related.



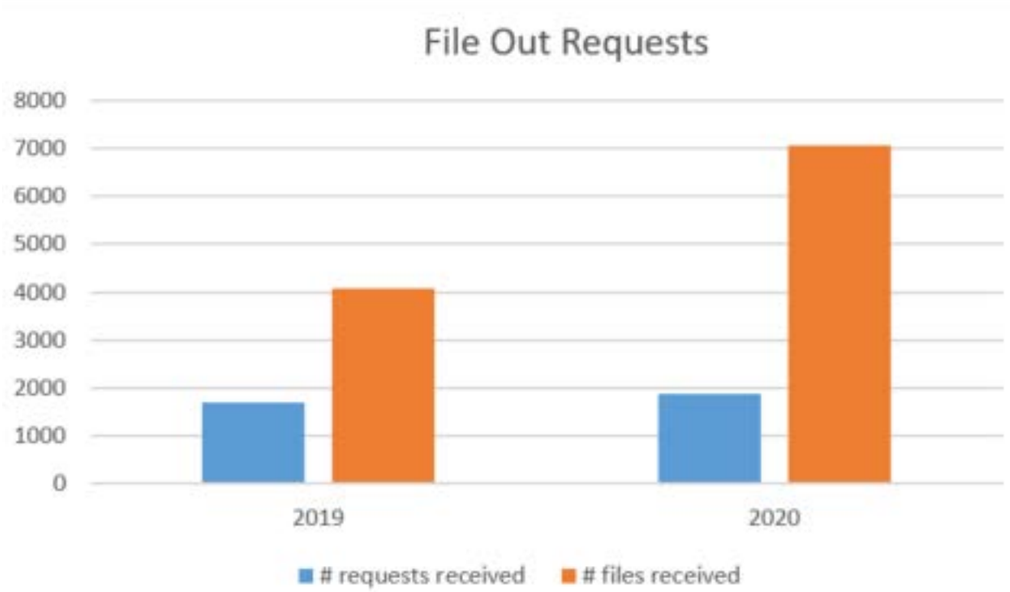
## File Out Reviews

Over the last year, the amount of files being requested to be released from SAIL has increased significantly.  This is partly down to COVID-19 research but also the increase in projects using SAIL for their research in general.  For the period of June 2020 to December 2020 the team received 1145 requests and reviewed over 5000 files, each month exceeding our SLA of reviewing 90% of requests within 2 working days of receipt. In comparison, for the same period in 2019, we received 1011 requests reviewing approximately 2000 files

From December 2019, as part of our DEA (Digital Economy Act) accreditation we defined and published a SAIL outputs policy as well as a User Guide to assist researchers in requesting their data. We adapted to a dual review system which involved developing a review checklist, a training process, increasing the number of reviewers within the team, and having admin support in managing the system.

Since April 2020, where Covid-19 related projects provide justification, we have been able to prioritise their requests and complete the reviews within an expedited timeframe of 6 working hours or less.

As part of our ongoing service improvement process we will be looking next to develop utilities which will check for common issues encountered in SAIL File out requests with the aim to increase file out review timescale and reduce the risk associated with human reviews.






## Project Audits

The project audits assure compliance with SAIL governance and with external accreditation such as the Digital Economy Act. In total, 21 projects were audited as part of the annual SAIL projects audit programme. These were selected at random from a report of active projects at the end of 2019. The projects are assessed based on rigorous criteria covering the project lifecycle, and completion of each audit requires input from across the SAIL team.

Some key actions were identified and have been implemented:

1. Creation and publication of a SAIL good practice guide that will show researchers how to work in a way that aligns with SAIL governance and policies.

2. Development of automated project view quality checks which will help ensure that the data provisioned is in line with IGRP approvals.

Findings are reported to the SAIL Operations Group and implementation of actions is monitored by the internal audit committee.

## And finally ...

This concludes our summary of our teams' services and research endeavours. For more details about SAIL see www.SAILDatabank.com and if you have any specific questions relating to this report then please contact us via help.SAILDatabank.com

# Secure
# Anonymised
# Linkable

Containing billions of person-based records, SAIL Databank is a rich and trusted population databank. It improves lives by providing researchers with secure, linkable and anonymised data that can be accessed and analysed from anywhere in the world.

Funded by

Ymchwil Iechyd a Gofal Cymru
Health and Care Research Wales

Ariennir gan Lywodraeth Cymru
Funded by Welsh Government

UKRI Economic and Social Research Council

Supported by

Swansea University
Prifysgol Abertawe

Certified & accredited

bsi. ISO/IEC 27001 Information Security Management
IS632731

UK Statistics Authority
Awdurdod Ystadegau'r DU

In partnership with

GIG CYMRU NHS WALES
Gwasanaeth Gwybodeg
Informatics Service

Powered by

SeRP
ISO certified and UK Statistics Authority Accredited

www.SAILDatabank.com