### 

**Secure Anonymised Information Linkage (SAIL)**

**Information Governance Review Panel (IGRP) Application Form**

**SAIL IGRP Application Form**

The following form has been designed to collect the information needed for the information governance approval process for work involving the SAIL databank. The information you provide will facilitate consideration of your enquiry. Guidance notes on completing this form can be found at: <http://www.saildatabank.com/media/25300/Guidance_Notes_for_SAIL_IGRP_Application.docx>

***SAIL Feasibility Agreement***

*All projects require a SAIL Feasibility Agreement to be completed and signed before proceeding to IGRP. This agreement will have been developed as part of the initial project scoping process with a senior SAIL analyst. Do not continue with this form until you have had your project scoping discussion.*

*Please provide the agreement number: 0661*

**1a. Provide contact details of project lead:**

Name: Professor Mark Rees

Job title: Professor of Neurology Research Swansea University

Organisation: Swansea University / NHS Wales

Address: Institute of Life Sciences 1

Tel: 01792 602203

Fax:

Email: m.i.rees@swansea.ac.uk

**1b. Provide contact details of the lead contact from any other organisation who will be accessing the data:**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Name** | **Job title** | **Organisation** |
| 1 | Arron Lacey | SAIL Research Analyst | SAIL Databank |
| 2 | Dr Owen Pickrell | Clinical Lecturer | Neurology Research Team |
| 3 | Dr Seo-Kyung Chung | Senior Lecturer | Neurology Research Team |

**2. Provide full title of the project:** Linking epilepsy next-generation sequencing datasets with routinely-collected healthcare records.

**3. Provide details on who is commissioning the project:** Swansea Neurology Research Team **/** SAIL / GeDI (Genetic Data Integration into data safe havens – please see Q4)

**4. Provide the aim of the project, including anticipated outcomes:** We aim to link and integrate gene-variant datasets (from next-generation sequencing sources) to corresponding electronic healthcare records within the SAIL databank. The genetic datasets will derive from Welsh patients who have DNA archived in the Swansea Neurology Biobank (SNB). This will be a project to assess the feasibility of linking variant-call format (VCF) files to SQL Databases, and to determine analysis pathways to identify association between genetic variants and clinical outcomes – in particular drug response and co-morbidities. We will work, in parallel, with the GeDI group to help formulate the regulatory framework for this work.

The GeDI project is funded by the MRC through the Wales Genomic Medicine Centre (GMC) being established at Cardiff. The aim of GeDI is to optimise ways of including genetic data in safe havens. Many genetic datasets are released to researchers or are uploaded to websites for access and use in research, but increased awareness of the inherent risks is resulting to a move towards using genetic data in safe havens. The added advantage of this, is that the data can be used in conjunction with health records to lead to new insights. Being able to make better use of genetic data is a high priority for UK governments, and this proposal to incorporate epilepsy genetic data into SAIL for linkage with SAIL datasets presents an ideal opportunity to show how this can be done safely under SAIL governance, and show value for research.

In this project we will study the association of genetic variants with three important areas:

1. *Epilepsy severity*. We will use anti-epileptic drug (AED) polytherapy and hospital/emergency department admissions as a marker of epilepsy severity and compare gene variants in people with epilepsy in these groups.

2. *Co-morbidities*. It is known that there is a link with epilepsy and co-morbidities such as mental illness, migraines and learning difficulties and so we will compare people with epilepsy with and without diagnosis of such. We will compare gene variants in people with epilepsy with and without co-morbidities. We will use GP recorded diagnosis of co-morbidities for this.

3. *Socioeconomic deprivation*. We know from previous work that epilepsy is associated with socioeconomic deprivation. In order to identify any possible genetic reasons for this, we will compare gene variants in the people with epilepsy from more deprived areas (WIMD score) with people with epilepsy from less deprived areas.

*Please include a copy of the protocol/plan for the proposed work with SAIL, including the contact details of any co-applicants when you return your completed form.*

**5. Provide a lay summary of the project:** Whole Exome Sequencing (WES) is a method of looking at an individual's genetic DNA 'code' found within their cells. WES targets the most important part of an individual’s genetic code (approximately 1–2% of the total) comprising of genetic 'instructions' that code for proteins. It is possible to use WES to obtain a list of genetic variants for an individual across all known 25,000 genes. This gene-variant list typically contains around 30,000 – 40,000 genetic variants for each person. The vast majority of these variants will be harmless population-wide changes, but some may cause a genetic predisposition to certain diseases or conditions.

Epilepsy is a common, chronic condition affecting approximately 1% of the UK population (approximately 30,000 people in Wales). It is well established that most epilepsies have a genetic basis, but apart from a small minority of cases, this link is not fully-understood. It is likely that integrating modern genetic sequencing data from techniques such as WES, with large routinely-collected electronic data sets will help unravel the complicated genetics of epilepsy and the consequences of diagnosis.

The Swansea Neurology Team have studied inheritance of epilepsy for 15 years and have identified novel variants attributed to disease. More recent work has focussed on the bioinformatics interpretation of large-WES datasets as the era of ‘omic’ outcomes is becoming mainstream. This presents a unique opportunity to link gene-variants within the WES datasets of Welsh epilepsy patients to their respective Electronic Healthcare Records (EHR) within the SAIL Databank. This will be a pathfinder project to incorporate the 30-40,000 genetic variants per person with their corresponding EHRs.

In this study we will look at three important areas in epilepsy and whether there are any links with genetic variation. We will look at the effect of genetic variation on:

1 *Epilepsy severity.* People with more severe epilepsy are typically taking more anti-epileptic drugs and often attend emergency departments and hospitals more often and so we will compare people with epilepsy in these groups.

2 *Co-morbidities.* It is known that there is a link with epilepsy and co-morbidities such as mental illness, migraines and learning difficulties and so we will compare people with epilepsy with and without diagnosis of such conditions.

3 *Deprivation*. It is known that there is a link with epilepsy and deprivation and so we will compare people with epilepsy from more deprived areas with people with epilepsy from less deprived areas.

**6. Provide an outline of the public engagement strategy for the study, or a brief explanation why there is not public engagement:** We do not provide a public engagement strategy as it is not needed for this project. We have permission to use each patients' data for secondary research purposes.

**7. Provide information on the relevant permissions you have obtained or that are being sought:**  *Obtained Being sought Not required*

***Research ethics***[  ] [  ] [  ]

*Please state the name of the committee that is being applied to/ has given approval, as applicable:*

*Research ethics committee:* The genetic data has been obtained as part of two projects with ethical approval: the Swansea Neurology Biobank and the genetic basis of familial epilepsy in Wales

*If you have ticked ‘not required’ please specify the reasons:*

The project will use onlyanonymised data, and therefore research ethics review is not required.

Other:

*Obtained Being sought Not required*

***Independent peer review***[  ] [  ] [  ]

*Please state the name of the peer reviewer that is being applied to/ has given approval, as applicable:*

*Peer reviewer:*

*If you have ticked ‘not required’ please specify the reasons:*

The project will use onlyanonymised data, and therefore independent review is not required.

Other:

***Permission from data-holding*** *Obtained Being sought Not required*

***organisation to use their datasets***[  ] [  ] [  ]

*Please state the name of the data provider that is being applied to/ has given approval, as applicable:*

*Data organisation:* Swansea Neurology Research Team and the Swansea Neurology Biobank

*If you have ticked ‘not required’ please specify the reasons.*

The project uses only SAIL unrestricted core datasets and/or data held by the project.

Other:

**Please note that it is the responsibility of the project lead to ensure that the relevant permissions are obtained.**

**8a. Provide a prospective start date for the work involving SAIL:** 01/09/2017

**8b. Provide anticipated end date of the project:** 01/09/2019

**9a. Provide details of data you require access to for the proposed work with SAIL?**

Please list:

The SAIL datasets you require information from

General Practice

Secondary Care (PEDW)

Wales Demographic data

Child Health Dataset

The information needed from each dataset

GP – Drug prescriptions, epilepsy diagnosis, psychiatric diagnosis

Wales Demographic Data – to obtain Welsh Index of Multiple Deprivation (WIMD) scores. We require RALFs to assertain whether individuals are living alone or not (known risk factor for sudden death in epilepsy)

Secondary Care - All super spells, spells and episode data for each person, inclduing diagnosis and OPCS codes.

Child Health Dataset – Maternal ALF, birth records

Please indicate the time period for which data is requested

1/1/2000 – 31/12/2016

Please indicate the geographic area for which data is requested

We require data for the Individuals with genetic data, these are individuals from South Wales, mainly based around the Swansea and Cardiff areas

Please indicate demographic criteria for the data requested (age, gender, etc.)

There will be a range of ages and both male and females with genetic data

**9b. Will you be providing any other dataset(s) to be incorporated into the SAIL databank?**

Yes [  ] No [  ]

If yes:

Provide the name of the dataset(s): The WES-JME Cohort from the Swansea Neurology Biobank

Provide details of the contents of the dataset(s): VCF files for 50 WES datasets for Juvenile Myoclonic Epilepsy cases. Each person has a VCF file equivalent to a plain text file with 30-40,000 rows of data.

**9c. Provide an outline of your analysis plan including the anticipated outputs:** We will use the principles of gene-variant, pharmacological and phenotypic association studies. We will use statistical methods and machine learning methods (implementing associative rule learning and classification algorithms using R/python) to compare gene variant frequency/ deleterious gene variant density and biological action of genes in the different groups of people with epilepsy. The main focus of outputs will be the principles of genetic information integration in SAIL but we will also report genetic differences in the different groups of people with epilepsy mentioned in question 4. We will not report individual genotypes only genetic differences between the groups in terms of aggregated genetic descriptions. We will also not use very small groups (n<5).

**9d. Are the results/methods developed likely to have other potential applications?**

Yes [  ] No [  ]

If yes, please specify: We aim to explore the best-practice methods for the integration of anonymised-linked genetic information within SAIL and discover how to construct robust and meaningful research as a result of the integration.

**10a. Please indicate your plans for publishing the results of your project, e.g. target journal or intended recipients of report:**

If any results are found suitable for publication, this would likely be in a high impact medical research journal. This will also help with future good-practice / WG policy documents in collaboration with GeDI.

**10b. What are the potentially sensitive issues that need to be taken into account when publicising the findings of the project?**

Please outline the issues and your proposed solutions:

People may be concerned about the potential to identifying individuals from genetic data however as indicated in question 9c we will not report individual genetic data / genotypes and only use aggregated/group genetic markers from which it will be impossible to identify individuals.

**What to do next**

Please return your completed form and supporting documents by email to Cynthia McNerney, Information Governance Coordinator [c.l.mcnerney@swansea.ac.uk](mailto:c.l.mcnerney@swansea.ac.uk) Thank you.