Plan Overview

A Data Management Plan created using DMPonline

Title: Molecular pathology cancer epidemiology studies to inform public health strategies for prevention, early detection and precision medicine

Creator: Jonine Figueroa

Affiliation: University of Edinburgh

Funder: Medical Research Council (MRC)

Template: MRC Template

Project abstract:

The overarching goal of this programme is to develop new tools and algorithms that will improve National Health Services (NHS) effectiveness and productivity, to prevent, detect and treat cancers earlier, for improved mortality and morbidity outcomes. The future of cancer surveillance, diagnosis and treatment decisions, will be increasingly emphasising molecular genetics of blood, benign tissues and tumours. Building on my experience in interdisciplinary research, my programme would draw on Scottish national health records datasets and epidemiologic studies in the UK and abroad, to investigate how molecular genetics, pathology and radiology imaging data might improve earlier detection and precision medicine in a productive and cost-effective manner. Specific aims are:

- Assess how to improve diagnostic pathways and precision medicine treatment for molecular subtypes of breast cancer. Breast cancer is the only tumour where molecular marker expression data is routinely captured in UK cancer registries, specifically: oestrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor (HER2). We will use already ethically approved population electronic medical records data for over 70,000 breast cancer cases diagnosed since 1997 in Scotland, to determine diagnostic pathways, incidence and mortality trends for these molecularly defined subsets of tumours, with future scalability to the rest of the UK.
- Using detailed molecular genetics and epidemiologic risk factor data from consented cohorts UK Biobank and Generation Scotland, create a digital pathology and molecular genetic tumour dataset (Total~1400 breast cancer cases), which in collaboration with international consortia—would provide the most robust risk prediction models for breast cancer that could then be implemented at population scale.
- Identify new imaging biomarkers prior to any diagnosis of cancer, providing novel insights into future cancer risk, which could lead to innovative tools and markers for improved surveillance and early detection of disease in high risk women. To achieve this objective, we would use unique tissue biobanks for marker discovery combined with radiology imaging data, to assess known tissue markers of risk, as well as perform machine learning and deep-learning studies to identify new imaging markers. These new imaging markers could be assessed in population datasets to determine their potential to detect cancer earlier or refine treatment protocols for improved survival outcomes.

Although initial studies would be on breast cancer, the tools developed could be applied to many different tumour types where molecular data have been or are in the process of being collected at the University of Edinburgh (colon, ovarian, and brain cancer), which are common or have unmet need. The long-term vision is that these unique datasets with tissue

phenotypes will facilitate the UK’s vision to maximise NHS data as a longitudinal population-wide cohort, comprising molecular pathology health-relevant data. Cumulatively, this programme would create new informatic tools and identify molecular and imaging markers, that would make NHS services more efficient, increasing economic productivity and improving survival outcomes in the UK and worldwide.

ID: 28661

Last modified: 13-01-2021

Copyright information:

Molecular pathology cancer epidemiology studies to inform public health strategies for prevention, early detection and precision medicine

0. Proposal name

0. Enter the proposal name

Molecular pathology cancer epidemiology studies to inform public health strategies for prevention, early detection and precision medicine

1. Description of Data.

1.1 Type of Study

Epidemiologic studies with data on demographic, genetic, morbidity and mortality outcomes. Images and molecular genetic data will be collected from tumour and normal tissues from human specimens that will be linked to historical data from Scottish health records and radiology imaging as well as more comprehensively assessed cohort studies of subjects recruited in Scotland (UK Biobank and Generation Scotland).

1.2 Types of Data

- Risk factor questionnaire data
- Medical records data (e.g. International Classification of Disease coding, treatment, and symptoms)
- Radiology images
- Genotypic data (host germline data)
- Tumour and normal histology images
- Quantitative molecular genetic data from tumour and normal tissues

1.3 Format and scale of the data

SQL databases for

- FASTQ for mRNA or DNA sequencing data
- GEN files for genotype data
- Digital pathology images will be in Hamamatsu format NDPI, and Aperio SVS format, Ventana and other vendors TIF
- Quantitative CSV, text file of mRNA expression
- DICOM radiology images

Open source standard formats will be used wherever possible. Otherwise standard file formats in the field of research will be used.

2. Data collection / generation

2.1 Methodologies for data collection / generation

Historical health data will be obtained by NHS Scotland Information Services Division (ISD), electronic Data Research and Innovation Service (eDRIS). CSV files of coded structured data will be provided. Pipelines for aquiring DICOM images from radiology are in process and we expect by Summer 2019 to obtain access to mammagraphy images in DICOM format from eDRIS. Hematoxylin and eosin stained sections (H&E), is the principal stains in histology to assess pathology and tumour presence. Images of H&E's from cancer cases identified through the Scottish Cancer Registry dataset will be scanned using either a Hamamatsu or Leica digital slide

scanner.

Germline genetic data, biochemistry and other risk factor data for UK Biobank are structured and coded.

mRNA expression data would be aquired using Nanostring nCounter values. mRNA and DNA sequencing would be aquired through either Lexogen Quant-seq or Affymatrix RNA sequencing technologies.

2.2 Data quality and standards

Data will be collected, managed and stored in compliance with the requirements of the Joint Code of Practice for Research, MRC's Good Research Practice, University of Edinburgh's Research Data Management policy and local protocol documents. These guidelines and documents will be used to assure that the design, execution and results of the experiments are captured and documented using best practice. The University of Edinburgh has a superb training programme on data collection and management (RDMS MOOC and MANTRA). Researchers working on this project will enrol and complete one or the other training programme, in order to assure the data is collected and managed using the highest standards and the best practices. Training of personnel will be documented and annual reviews of competence conducted. Recording of data will be regularly peer reviewed by the project PI and other academic peers, to ensure research data quality is maintained.

3. Data management, documentation and curation

3.1 Managing, storing and curating data

An integrated data repository will be developed to assure samples and their data can be easily linked and tracked for quality management and integrety purposes. We will develop tools to easily interrogate the database and status of tissue samples that are being aquired, availability, sectioned and H&E stained, digital image of H&E aquired, tumour presence and %, number of sections for molecular profiling and DNA, mRNA expression or sequencing data. Collected data will be held on password protected (via two factor authentication) network storage servers owned by The University of Edinburgh (DataStore).

https://www.ed.ac.uk/information-services/research-support/research-data-service/working-with-data/data-storage. DataStore is professionally managed by IT Infrastructure Team. It has fully redundant/resilient infrastructure with "daily snapshots" every weekday morning and daily backups to off-site tape every evening, and a "disaster recovery" copy to second site overnight. Data stored in DataStore will only be accessed from computers registered on the UoE network using two-factor authentication (EASE – the UoE's authentication service), or via using a Virtual Private Network (VPN) connected to the UoE when working from home at any time during the duration of the project lifecycle.). Servers are located in secure buildings accessible only via swipe card and in rooms protected by alarms and CCTV operated by the University IT Infrastructure Team.

Data will be retained for a minimum of ten years in locally archived private storage with public metadata record upon completion of project.

3.2 Metadata standards and data documentation

Data will be documented/described including all methodology. Data generated during the project will be accompanied by standardised, structured metadata record explaining the purpose, origin, creator(s), access conditions and terms of use of the data. Metadata of a minimum Dublin Core standard will be produced.

All research outputs will be linked to PI's entry in Edinburgh Research Explorer, Edinburgh preclinical imaging web page and recorded in the University's PURE system, accessible through the University Research Outputs Portal (www.research.ed.ac.uk/portal/), via digital object identifiers (DOI).

3.3 Data preservation strategy and standards

Electronic data will be moved at regular intervals to locally hosted resilient networked archiving facilities, where it will be securely held for a minimum ten years.

4. Data security and confidentiality of potentially disclosive personal information

4.1 Formal information/data security standards

As our project involves the use of genetic data and sequencing information, we will require strict access and monitoring of servers with data only to those individuals trained to gain approved access. A list of approved investigators with ethical and data management training will be given access. To minimise identification, biological samples and genetic datasets will only contain psuedonymised ids to minimise the identification of any persons in the study.
The linked dataset will be held on the University of Edinburgh Safe Haven. The unique study ID along with any personal details and contact info will be stored separately from the identifiable data. In this study we use the year of birth and sex, as quality check variables; to ensure we are linking the data correctly.
Identifiable data will not be made available to the researchers.

4.2 Main risks to data security

Identification are the main risks for this study and hence why we will minimise any use of these data to anyone except clinical care doctors, and the research study nurse. Storage of genetic data will be encrypted and only available through secure monitored access by individuals with accounts monitored by University of Edinburgh.

5. Data sharing and access

5.1 Suitability for sharing

The ISD and eDRIS datasets can not be shared as these are entrusted public health records, but summary level data will be made to investigators upon request (held in DataVault). For data for consented cohort subjects from UK Biobank, Generation Scotland, KConFab, and Komen Tissue bank-these data will be suitable for sharing and any new data generated from these subjects (e.g. imaging features) will be shared via DataShare and raw data returned to the organisations.

5.2 Discovery by potential users of the research data

To enable data sharing and ensure long-term discoverability and accessibility, the imaging data, together with relevant and appropriate metadata, will be offered to Edinburgh DataShare. Edinburgh DataShare will, on acceptance of the data, supply a DOI and suggested citation to be used by anyone citing this data in the future. It will also undertake to ensure that the data remains discoverable, accessible, and reusable for as long as practically possible.
At the end of the project Pure (www.pure.ed.ac.uk), the University's Current Research Information System (CRIS), will be used to store individual metadata records. This system feeds Edinburgh Research Explorer (www.research.ed.ac.uk/portal/), which provides an overview of the research activity of staff members at the University of Edinburgh. Terms of Access will be defined in the metadata for those researchers and academics who will make a request upon approval of the PI and a declaration of usage of data (controlled access). Publications deriving from the project will report the information on where and how data will be accessed.
Deposited data will be made available in accordance with the RCUK Open Access policy, except if data is considered sensitive to the research and not suitable for open access or if specific embargo periods have been agreed. We will follow the MRC policy on research data-sharing.
Finally, analysis of collected data will be published in peer reviewed journals and will be presented at national and international meetings.

5.3 Governance of access

The project PI will be responsible for the governance of the research data access during the project period. The PI will be assisted by the University of Edinburgh Research Data Support team who will advise on best practice for the maintenance and governance of data.

5.4 The study team's exclusive use of the data

We will share our research data in a timely fashion and in compliance with the MRC’s requirements. A limited period of exclusive use of data for primary research is reasonable giving the nature and value of data. The Edinburgh Research and Innovation team will be consulted if advice is needed on intellectual property (IP) rights. Data will be made publicly available at time of corresponding paper publication or earlier as appropriate and following IP issues having been addressed.

5.5 Restrictions or delays to sharing, with planned actions to limit such restrictions

The University of Edinburgh encourages publication of research results in a timely manner. We do not foresee restrictions for the majority of the collected data and will endeavour to provide immediate access to research outcomes where appropriate. Legal advice will be sought if this position changes for any reason.

5.6 Regulation of responsibilities of users

Data sharing and transfer within the University of Edinburgh will be conducted using data sharing services available to internal and external users, such as, DataSync and direct access to data in folders granted by the project PI. Staff are expected to adhere to the Institute's data management policy that defines users' responsibilities. Research outcomes will be regularly stored and backed up as per University of Edinburgh guidelines using robust storage platforms, namely DataStore. In case of incidental deletion of files, this system allows for tracking and retrieving of older files within a timely manner. External users will be expected to adhere to requirements set out in a data sharing agreement.

6. Responsibilities

6. Responsibilities

PI responsibilities: ensure data is generated according to protocols, ensure appropriate training of post-doctoral researchers in data collection and storage, ensure overall quality of data is of high standard and decide when data is ready to be shared with the wider community.
PI responsibilities together with Institute's Data Manager: ensure that best practice in data management is followed and maintained.
IT team responsibilities: ensure the data is properly secured and backed up on regular basis, in line with PI and Institute's Data Manager expectations.

7. Relevant policies

7. Relevant institutional, departmental or study policies on data sharing and data security

Policy	URL or reference
Data Management Policy and Procedures	https://www.ed.ac.uk/information-services/research-support/research-data-service https://www.ed.ac.uk/information-services/about/policies-and-regulations/research-data-policy
Data Security Policy	https://www.ed.ac.uk/records-management/policy/data-protection https://www.ed.ac.uk/information-services/about/policies-and-regulations/security-policies/security-policy
Data Sharing Policy	https://mrc.ukri.org/documents/pdf/mrc-data-sharing-policy/ https://www.ed.ac.uk/information-services/research-support/research-data-service/sharing-preserving-data/data-repository/service-policies/preservation-policy https://www.ed.ac.uk/information-services/research-support/research-data-service/sharing-preserving-data/data-repository/service-policies/data-metadata-policy
Institutional Information Policy	
Other	
Other	

8. Author and contact details

8. Author of this Data Management Plan (Name) and, if different to that of the Principal Investigator, their telephone & email contact details