SCOTEY: The SCottish Oncology & Tissue EpidemiologY study to inform public health, policy, prevention and precision medicine

A Data Management Plan created using DMPonline

Creator: Jonine Figueroa

Affiliation: University of Edinburgh

Funder: Wellcome Trust

Template: Wellcome Trust Template

Project abstract:

Cancer is a complex disease whose aetiology and treatment differ by molecular markers. Through a Wellcome trust seed award, we have shown that in Scotland incidence and survival outcomes clearly differ over time by molecular subtypes of breast cancer and by different subgroups. Harnessing the tumour tissue data and high quality electronic records data resources, we aim to expand this work and build a retrospective cohort study of cancer in Scotland including over 80,000 breast cancers dignose since 1997 to determine: How have changes in risk factors, whose associations differ by molecular subtype, impacted breast cancer incidence rates in Scotland? Which molecular subtypes have worse survival outcomes, and do we see differences by certain sub-populations (e.g. age, deprivation, screen-detected)? How have different treatments and adherence to treatments impacted survival outcomes over time and have these gains been seen across all populations or only certain subgroups (e.g. age, deprivation)? Are there additional biomarkers that predict better outcomes using a) tissue specimen images b) molecular genetic profiling of tumour tissues? These data will inform population-based prevention, screening and treatment of breast cancer and also serve as an exemplar for other

cancer sites.

ID: 48108

Last modified: 12-11-2019

Copyright information:

SCOTEY: The SCottish Oncology & Tissue EpidemiologY study to inform public health, policy, prevention and precision medicine

Data and software outputs

The data and software outputs your research will generate

Study

I am proposing a molecular epidemiologic study 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 genetic data from embedded cohort studies recruited in Scotland (UK Biobank and Generation Scotland).

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; quantitative data from image analysis studies of pathology specimens.

Format and scale of the data

Open source standard formats will be used wherever possible. Otherwise standards in the field will be used. Records data on up to 70,000 subjects:

- SQL databases for medical records data
- 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 for Nanostring

Software

- R software, including Rmarkdown which is open access statistical analysis for https://github.com/rstudio/rmarkdown
- QuPath open access digital pathology software which is open access https://qupath.github.io
- Omero digital pathology database software which is open access https://www.openmicroscopy.org/omero/
- STATA for certain statistical analyses

Any analysis or algorithms developed using the above will be shared to ensure transparancy, reproducibility, replication and extension for other research efforts

When you intend to share your data and software

Data and software underpinning research articles will be available to other researchers at the time of preprint submission prior to publication in journals, providing this is consistent with:

- any ethics approvals and consents that cover the data (meta data will only be used for electronic health records data in line with ethics and privicy guidelines)
- reasonable limitations required for the appropriate management and exploitation of IP.

Where your data and software will be made available

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, including depositing on github.

How your data and software will be accessible to others

At the end of the project Pure, the University's Current Research Information System (CRIS), will be used to store individual metadata records. This system feeds Edinburgh Research Explorer, 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.

Whether limits to data and software sharing are required

Deposited data will be made available in accordance with the Wellcome 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. Finally, analysis of collected data will be published in peer reviewed journals and will be presented at national and international meetings.

How datasets and software will be preserved

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). Data preservation strategy and standards Electronic data will be moved at regular intervals to locally hosted resilient network archiving facilities, where it will be securely held for a minimum ten years.

Research materials

What materials your research will produce and how these will be made available

Historical health data will be obtained by NHS Scotland Information Services Division (ISD), electronic Data Research and Innovation Service (eDRIS). CSV files of coded structed data will be provided. Pipelines for acquiring 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 acquired using Nanostring nCounter values. mRNA and DNA sequencing would be acquired through either Lexogen Quant-seq RNA sequencing technology. See description of how data will be made available for datasharing.

Intellectual property

What IP your research will generate

We envisaged that in order for the ‘project outputs’ to be adopted successfully as [e.g. a clinical intervention,] a competitive advantage will be required. The IP strategy will appropriately consider what aspects of the project outputs are appropriate for formal ‘technology transfer’ and those where non-proprietary wide dissemination e.g. via publication will provide the greatest impact and uptake. The types of IP that may be created include:

- Novel biomarkers of prognostic relevance and new methodologies to assess the molecular characteristics of cancers using little input pathology material for more rapid assessment.
- Innovative informatics software tools for visualising and summarising population based data using molecular characteristics of groups of patients for more accurate assessment of the state of incidence, morbidity and mortality outcomes
- New image processing and (AI)-based tools for assessment of pathology tissues to

identify novel biomarkers and treatment targets.

Academic personal development will be supported through EI's Enterprise Team, which provides regular training and workshops in commercialisation and entrepreneurship for the University's academics, such as the 'Staff Business Bootcamp'; as well as one-to-one mentoring through the team to create a bespoke support package.

How IP will be protected

As a charitable organisation, key to developing a commercialisation strategy (where appropriate) is to ascertain the optimal route to market (e.g. Industrial partner/licensing, company formation) to transfer the technology smoothly into a partner capable of exploitation to generate economic, patient and societal benefit. All licensing of intellectual property arising from UoE incorporates good practice provisions expected of a public charitable body including (i) 'use or loose' safeguards, (ii) retaining research rights for both UoE and other not-for-profit institutions, (iii) ensuring publication isn't delayed and/or prevented, and (iv) consideration of the University's essential medicines policy. This activity would be supported by EI's Business Development Team, who would meet regularly with the academic lead to track progress in order to develop the project's commercial pathway.

How IP will be used to achieve health benefits

To provide the best evidence for improved tools for clinical decision making, multiple large datasets are needed for validation, calibration, to design possible routes toimplementation. Leveraging my track record of outstanding cross-disciplinary research and achievements in conducting multiple epidemiologic studies in the UK, Europe, the USand Africa, I will leverage the world's best researchers in molecular genetics, epidemiology, pathology, computer science, informatics and statistics-for improved risk prediction,early diagnosis and treatment of cancer. Such alliances would enhance the competitiveness of the UK knowledge and health bases and strongly influence the international cancerresearch agenda for improved patient outcomes world-wide.

Required resources

You should consider what resources you may need to deliver your plan and outline where dedicated resources are required.

To achieve the above we have budgeted for the following resources:

- Full-time database manager including managing meta-data for digital pathology analysis
- Histopathologist costs for processing and curating digital pathology image database
- Data storage/management costs through University of Edinburgh

- Training for software and analysis tools for team members
- Data access costs for health records and pathology specimen retrieval, processing and scanning
- Open access costs for publications and websites