Genomic Data Integration (GeDI)

Professor Kerina Jones

Rationale for GeDI

Huge interest and value in genomic data research and precision medicine
- Debate - anonymization of genetic data - identifiable/de-identified/anonymized
  - GDPR Article 9: special category data, along with health, political views, race, etc.
  - Unique ≠ identifiable
  - How different is genomic data anyway?
    - Genomic data (largely) persists
    - Sensitive/discrimination
    - Partially unknown
    - Impact on kin - heritable conditions, paternity
  - But not exclusively...
  - Hearts and minds swayed by opinion
  - Need for clearer data governance
  - To date: majority of genomic data is collected for research

Are genomic data special?

- Laissez faire vs ‘genetic exceptionalism’
  - Open genomic data advocates
  - Special rules & regulations
- Survey of genetic data uses:
  - Forensics
  - DTCs
  - Precision medicine
  - Health research/secondary use
  - Genealogical use
- But - maybe the jury is still out...
- Public - DNA information different to other medical information: Yes 52%, No/unsure 48%
- Largest differences in opinion between genetic health professionals and the public
Studies using genetic data and health records

<table>
<thead>
<tr>
<th>Health conditions</th>
<th>Health records</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alzheimer’s disease</td>
<td>Cardiovascular disease</td>
</tr>
<tr>
<td>High cholesterol</td>
<td>Cancer</td>
</tr>
<tr>
<td>Alcohol dependence</td>
<td>PTSD</td>
</tr>
</tbody>
</table>

Types of genetic data

- Presence/absence of trait
- Polygenic risk scores
- Genetic activity scores
- SNPs
- VCFs

Sources of genetic data

- UK Biobank
- Generation Scotland
- China Kadoorie Biobank
- eMERGE Network
- OncoShare project
- Mayo Genome Consortium

Sources of routine data

- EHRs
- Disease registries
- Death registries
- Health insurance systems
- Deprivation indices

Sources of genetic data

- UK Biobank
- Generation Scotland
- China Kadoorie Biobank
- eMERGE Network
- OncoShare project
- Mayo Genome Consortium

Releases of data to researchers with IRB approval and data use agreement

- Sax Institute: 260,000 people over whole genomes and derived data
- SAIL: Psychosis cohort with psychosis polygenic risk scores and VCFs
- IC/ES: Cohort: ~2,200 children with neurodevelopment disorders
- BC Generations: 30,000 people aged 35-69
- DPUK: Dementia research: >200,000 people
- DPUK: Polygenic risk scores and CNVs
- DPUK: Data release but if linked then access is only via safe haven

Mini-case studies: genetic and health data in data safe havens

<table>
<thead>
<tr>
<th>Institute</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sax Institute</td>
<td>260,000 people over whole genomes and derived data</td>
</tr>
<tr>
<td>SAIL</td>
<td>Welsh cohort with psychosis polygenic risk scores and VCFs</td>
</tr>
<tr>
<td>IC/ES</td>
<td>Cohort: ~2,200 children with neurodevelopment disorder</td>
</tr>
<tr>
<td>BC Generations</td>
<td>30,000 people aged 35-69</td>
</tr>
</tbody>
</table>

Models of genomic data access

- Held by/released to researchers: mainly via biobanking or research studies
- Accessed within a data safe haven: mainly via data centres
- Publicly available on a website
  - Personal Genome Project
- Can be with/without linked data
- Can be identifiable/de-identified/anonymised
- How data are collected, risk appetite, access restrictions, approvals, purpose, technical capacity, social acceptability...
Willingness for the reuse of genomic data collected for research

Suggested wording - starting to see results!

For all research studies using personal data if want to option to share for linkage
Plan from the outset

For the Participant Information Sheet -
‘The data you provide to us is important and may be useful to other research studies. So that your data can be used anonymously for further research in the public interest, an NHS organisation will replace your identifying details with a unique anonymous code. This will enable your data to be linked to routinely-collected data, including your health records. The data can then be used for research in anonymous form in a secure environment, such as the Secure Anonymised Information Linkage Databank (SAIL).’

For the consent form -
‘I understand that my personal data will be de-identified by an NHS organisation so that it can be used in anonymous form for further research in the public interest.’

Modelling: risk-control-utility

Controls within a data safe haven – options and combinations determined on a project-by-project basis
1. Remote access as per usual (SAIL) including results release following scrutiny
2. Data use contingent on data provider permission
3. Access by analyst employed by SAIL/DSH only
4. Access within safe room only
5. Conditions for any data outs contingent on specific approvals
6. Suppression, aggregation, masking, differential privacy, etc., retaining maximum utility
7. Federated access i.e. genetic data not brought into DSH

Red – Amber – Green type model:
Green: business as usual
Amber: plus additional safeguards
Red: modify research question or no go

Metadata/Data

Presence/absence of a trait
Monogenic risk score
Polygenic risk scores
Link
SNPs
VCF files
Partial gene sequence
Full single gene sequence
Multiple gene sequences
Full genome sequence

Risk factors:
Sensitive/stigmatising conditions
 Rare conditions
 Data granularity
 Data extent
 Linkage to other datasets
 Other factors

* +/- annotations
Summary

- Evidence to inform data governance for working with genetic data is essential
- Data protection legislation categorises genetic data as for other health data
- Currently the majority of genetic data is with research studies
- Our work indicates public support providing data are handled properly - data safe havens preferred
- General move from data release to data in safe haven
- Current state of play in research studies - missed opportunities (not just for genomic data)
- Importance will further increase as genetic data becomes more mainstream in EHRs

Further work

- How/what/who to store: beyond VCFs further processing is likely to be needed before can analyse with health records
- Pipelines: governance and technical models for secure data transfer, management and linkage
- Analytics: how to prepare data and work with phenotypic and genomic data, software and skill sets
- Pathfinders: exemplar projects to test and consolidate processes
- Stakeholders: working with data controllers/providers, researchers, etc.
- Working as part of Genomics Partnership Wales: genomics for precision medicine strategy implementation; Wales Gene Park; Welsh research infrastructure; biobanking community, etc.
- Further public and stakeholder engagement: on genomic data research and precision medicine - just opened the box
- Advising: on the ethical, legal and societal issues in integrating genomic data with health records
- Education and capacity building: sharing what we’ve learned

SAIL interim policy

Genomic data (derived from the DNA sequence of an individual) can be more sensitive than general health data, bringing with them specific issues when data is released. Furthermore, due to their inherent uniqueness, genomic data derivatives may pose additional risks when linked to other datasets in SAIL. To mitigate these risks, we have set out the following rules in relation to use of genomic data derivatives in SAIL:

1. Project-based (non-routinely collected) genomic data may be deposited into SAIL providing all relevant approvals and permissions have been obtained, and an appropriate consent model has been used.
2. Routinely-collected genomic data are brought into SAIL as part of standard data feeds.
3. A member of the IGRP with genomics knowledge will sit on the committee when considering requests for access to genomic data.
4. If deemed particularly sensitive, some types of genomic data will require additional protective measures before release of results. This will be discussed with researchers during data scrutiny.
5. All the standard SAIL B SAIL Gateway operating policies and procedures will apply to these data in addition to the specific points described.

Finalise with reference to GA4GH: Framework for Responsible Sharing of Genomic and Health-Related Data
The GeDi study is funded by the MRC as part of the Wales Genomic Medicine Centre. For further information, please contact Kerina Jones: k.h.jones@swansea.ac.uk.

Are genomic data special?

Our survey of the range of current uses of genetic information and of the likely future scenarios of these uses illustrates that this is an area that needs careful ethical analysis, and that careful thought to present primary, sensitive and complex combinations of information is necessary and of potential benefit to patients in addressing the challenges of the scenario.

The volume of potential genetic information, the speed with which this is increasing and the potential for complex combinations of this information with other data have (as we have seen) led some writers to argue that we have to have special measures in place to deal with this type of information. However, we do not consider that such genetic information is either necessary or helpful in describing and protecting the ethical values involved. The scope of the clinical uses of social ethics affect the rights and corresponding responsibilities of individuals. The need for clear legal and regulatory protection of privacy and personal information, and the challenge of balancing between the protection of the rights of individuals and the promotion of the public good. The rapid progress of genomics does seem to demand — sometimes in a direct way — these new ethical issues, that it does not in any way change their character: the protection of privacy, and a measured understanding of what it entails, are the core issues of contemporary society. There is no special status called "genetic privacy."

Public workshops

- To gain views on socially acceptable approaches for accessing genetic + health data in anonymised form
- Focused on how (not if): open access, release to researchers, within a data safe haven
- Genetic data collected for research
- Series of 8 public workshops (N = 116)
- Men 54 : women 62
  - 6th formers
  - SU staff and student group
  - Business and NHS group
  - Consumer Panel
  - Science festival
  - Grand round
  - University of the 3rd age (U3A)