H3ABioNet Health Informatics Work Package

ClinDB project

Data governance guidelines

Contents

1. Key Tenets of Data Governance for Bioinformatics

Four main areas that should be considered in assessing governance of data

1.1. Consent and ethics

Consent by individuals for data use:

1.1.1. What information was offered to participants?

- Check participant information documents
- Do participants know how their data will be stored, used, and destroyed?
- Do participants know who is responsible for their data?
- Do participants know how they will/won't benefit from the study
- Do participants know how to withdraw from the study should they choose to do so?
- Do participants know who to contact with concerns?
- What consent was given (TIERED consent)? - Primary use, secondary use, return of findings and incidental findings.

1.1.2. Ethics of data use, IRB approvals

- Does the protocol that was submitted to ethics explain how data will be used, so that it is clear that the ethics board fully understands the intended data use? Even though ethics may have been granted, sometimes IRBs are not aware of the risks associated with data, and may have approved a protocol that includes unethical data use.

Key considerations for ethical data use:

i. Maximise benefits/minimize harms

- Beneficence: The good, and benefits, that can come from using these data
- Potential harms: how serious are the risks that the participants are exposed to. How do the potential benefits from the data use balance out against the risks? i.e. A piece of research that has no tangible benefit, and is being done for the sake of curiosity only, would not be sufficiently beneficial to put individuals at risk of disclosure of personal data or confidentiality breach.

ii. Do good science

- Is the study design appropriate – will the study yield real results?
- Will the data be used (within the bounds of the consent) to yield maximum benefits?
- Data fidelity: is the dataset of good enough quality to provide real results?

- Is there evidence of the IRB approval? Is it up-to-date/not expired?

1.2. Data access controls

1.2.1. Structural control

- Where are the data stored?
- What are the physical barriers to access? Firewalls, login access, gateway servers etc
- How are the data structured to prevent confidentiality breach?
  - Separation of identifying information from sensitive information
  - internal linking IDs are not shared elsewhere, to prevent linking of separate datasets by endusers (unique random study numbers generated for each dataset)
  - For de-identified datasets, obfuscation of data where possible to avoid re-identification (e.g. year of birth instead of date of birth; shifting all dates by an undisclosed integer to retain epidemiological relevance but make re-identification much harder)
  - When datasets are transported, they are encrypted, password protected, identifying and sensitive information are transported separately. Only secure online transfer platforms are used (add bioinformatics examples), and passwords are shared by separate media e.g. sms, or by phone call.

1.2.2. Procedural control

- Set up of DBACS – Access committees who oversee dataset requests are constituted according to a defined structure. Often must contain a bioethicist, as well as other defined roles, number required etc.

- SOPs will describe an appropriate data access request pipeline: all requests are centralised and processed in a standardised way.

1.3. Legislation

National legislation, will vary from country to country. Types of legislation might include:

- National Health Act
  Describes the responsibilities of health care providers in dealing with the health data for those in their clinical care
- Protection of minors and vulnerable people
  Special considerations for any studies involving minors, and vulnerable people e.g. those who are mentally unable to make decisions for themselves.
- Protection of Personal Information Act
  Specifically around the use, reuse, repurposing and sharing of personal information
- Public Access to Information Act
  Gives participants the right to know how their data have been used. Requires tracking and logging of all instances of data use.

1.4. Sustainability

FAIR data, including:

- Documentation – avoid reliance on particular individuals. Document where the data are, how they were generated, codebook, supporting docs, etc
- Complete metadata.
- Backups – ensure data are appropriately backed up to a secure location, have a disaster-recovery plan
- Version control – if the data are altered over time ensure a version control plan
- Interoperability – ensure data can be shared cleanly, where appropriate e.g. HL7 or FHIR specification for clinical records
- Make a succession plan for if people leave, funders change, etc

2. Checklist for data governance

[we can build this with more commentary for each point if needed. I am thinking about adding a point about the analysis being truthful and fidelity of the data]

- The study participant knows what I am using their health data for, and is ok with it
- The ethics board knows what health data I am using, and how; and is ok with it

- This data use complies with legislation (POPI, Healthcare Act, in SA)
- All datasets and drives are password protected and encrypted
- All participant identifiers are stored and transferred in a separate file to any clinical data
- All anonymised data cannot be re-identified
- I never use email to transfer identifiable health information, and I always send passwords separately (not by email).
- If I were a study participant I would be happy with the way my personal health data are being used.