ETHICAL DATA MANAGEMENT: CHECKLIST FOR GENE/DISEASE SPECIFIC DATABASE CURATORS

Authors
Rosemary Ekong, Mauno Vihinen, Michael Parker

Editor Helen M Robinson

Contents

Introduction and Background

This checklist has been developed to assist curators of gene/disease variant databases in ensuring that their work meets international requirements. Individual database curators have the responsibility to ensure that they always act in a manner that is consistent with their local requirements as these may vary from country to country. Database curators should also ensure that they promote ethical approaches to those in their collaborative networks. Also, where possible, they should contribute to the continuous improvement and refinement of ethical approaches in their national and international work.

This work has a long history. In 2010, guidelines were published (Povey et al. 2010) to help curators of web-based locus-specific variation databases (LSDBs), or gene variant databases, to make information within their databases accessible for use for both clinical and research purposes, while safeguarding the privacy of individuals. In 2014, at an HVP meeting, several curators raised concerns regarding the application of existing guidelines, indicating that they found some were difficult to achieve in practice. In parallel to this, interest in curation of variant information was growing.

In response to this, the Gene/Disease Database Advisory Council oversaw the formation of WG08, to develop "A checklist of actions and processes related to the ethical management of data in a genetic variation database that curators of gene/disease specific databases should consider when establishing and curating their database." WG08 conducted a survey of curators to gain a better understanding of current practice and to determine the extent to which the Povey et al. (2010) guidelines had been implemented.

The Checklist published here draws from the results of this survey, the views of expert database curators from within HVP, members of HVP's various councils and committees as well as inputs from the broader HVP membership. Some of the 'practical' guidelines in Povey et al. (2010) have been retained and information previously published in other articles have also been included (Celli et al. 2012; Mascalzoni et al. 2014).

The Checklist should not stand alone. Gene/disease database curators will need to ensure that they access relevant international, national and regional resources. Other documents that are useful include:

> (1) "Ethical and Privacy Principles in relation to Responsible Sharing of Genomic and Health Related Data" produced by the International Society for Gastrointestinal Hereditary Tumours (InSiGHT) (Appendix 1);
> (2) A broader "Framework for Responsible Sharing of Genomic and Health-related Data" (Version 10, Sept 2014) by the Global Alliance for Genomics and Health (GA4GH) (Knoppers, 2014), also available at
> https://www.ga4gh.org/ga4ghtoolkit/regulatoryandethics/framework-for-responsible-sharing-genomic-and-health-related-data/;
> (3) The "WMA Declaration of Taipei on Ethical Considerations regarding Health Databases and Biobanks" (World Medical Association, 2016), which covers additional ethical principles for the collection, storage and use of identifiable data will be relevant for curators with access to confidential/sensitive data.

(4) The EU General Data Protection Regulation 2016/679 (GDPR) which comes into effect on 25 May 2018

Although this checklist was primarily drafted for curators of web-based LSDBs or gene variant databases, extended consultation has shown that its content may also be relevant to others in related areas including those initiatives that collect information into databases and make these accessible to the public, such as national/ethnic variation databases and biobanks. As such, we encourage the adoption of this checklist by others where it is judged to be applicable. Please share it with colleagues working in the field. The Checklist will be improved by being used. We invite feedback from users – both where is it particularly useful and those aspects that could be improved. Comments can be sent to – info@variome.org

Introducing the Checklist

i   Scope

The purpose of this document is to provide practical steps to enable LSDB and gene variant database curators to collect and share data in a manner that both ensures and promotes acceptable ethical standards. It should be stressed that individual database curators are obliged to be aware of, and act within, the national regulatory frameworks that govern their operations. It is the responsibility of the database curator to seek out this information and act in compliance when sharing variant information both within their own country and across national boundaries. Curators of other research databases, with publicly accessible data, may wish to use this checklist in their work. HVP supports free and open sharing of variant data for purposes of improving health, but it must be done within the spirit of all relevant local, national and international regulatory frameworks.

It is important that when this Checklist is implemented it is modified to meet local requirements. It can also be used as a means for building the case to improve local requirements.

ii   Terms and Definitions

The Checklist uses several terms that need to be defined for clarity and consistency.

- **Coded data**
Coded data refers to data that have undergone pseudonymisation as defined in Article 4(5) of the EU General Data Protection Regulation 2016/679 (GDPR) which comes into effect on 25 May 2018.
    - *Pseudonymisation*
    means the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person". GDPR document is available at http://ec.europa.eu/justice/data-

protection/reform/files/regulation_oj_en.pdf. A coded ID permits re-identification by the submitter.

A simplified definition can be found in the GA4GH Data Sharing Lexicon (GA4GH, 2016) where 'pseudonymisation or coding' is defined as "The act of replacing an identifier with a code for the purpose of avoiding direct identification of the participant, except by persons holding the key linking the code and identifier".

- **Database curator**

Database curator, as used here, refers to a person or persons who is/are "…responsible for assuring the quality, integrity, and access arrangements of data and metadata in a manner that is consistent with applicable law, institutional policy, and individual permissions". The activities of a database curator include data extraction, integration, presentation, publication, management, monitoring and reducing redundant information, thereby resulting in up-to-date information that is as accurate as possible. They also oversee who has access to the information and under what circumstances. In HVP terminology, the database curator is responsible for ensuring that the database policy is kept current and useful. Some organizations, like GA4GH, use the term 'data steward' to equate with database curator (see for example   - *GA4GH Data Sharing Lexicon* (2016)) but this term is not widely used.

- **Personal data or Identifiable data**

The Checklist adopts the definition in Article 4(1) of the EU General Data Protection Regulation 2016/679 that defines 'personal data' as "any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person"; and in alignment with the *GA4GH Data Sharing Lexicon* (2016), personal/identifiable data refers to "data that alone or in combination with other data may reasonably be expected to identify an individual".

- **Personal data breach**

This is defined in Article 4(12) of the EU General Data Protection Regulation 2016/679 as "a breach of security leading to the accidental or unlawful destruction, loss, alteration, unauthorised disclosure of, or access to, personal data transmitted, stored or otherwise processed".

The Checklist

The Checklist in brief

The Checklist follows thirteen simple elements that are set out in the table below. Each is then explained in more detail in the following section.

1.	Define the purpose of your database.
2.	Define the database policy.
3.	Offer attribution to submitters.
4.	Establish an oversight committee.
5.	Data collected with valid consent and coded ID (responsibility of submitters).
6.	Curate unpublished data to protect patient privacy whilst remaining useful.
7.	Requests for non-public data should be forwarded to the submitter.
8.	Requests to keep submitted data non-public can be honoured.
9.	Requests for submitter's details should be forwarded to the submitter.
10.	Giving your opinion may be considered if you have a team qualified and knowledgeable about the disease.
11.	Information can be shared with genome browsers.
12.	Personal data breach should be dealt with immediately.
13.	Data erasure should be honoured where details have been provided.

The Checklist – full information

1. **Define the purpose of your database**
   a. Include the scope and type of information in database.

2. In accordance with the agreed purpose of your database, **define the database policy** governing data collection, display, access, variant classification and corrections. The policy should also include clauses on data correction and withdrawal/erasure. Examples of database policies are in:
   - Vihinen et al. 2012, (Appendix 2 in this document) and
   - Knoppers, 2014 (GA4GH Proposed Policy Template).

   This should also include information suitable for patients with limited technical knowledge wishing to submit their data.

3. **Attribution**

   a. In recognition for the submission of unpublished data, submitter names should be listed along with their data (unless not permitted by database regulations).
   b. Where local database regulations do not permit listing submitter names, consider offering co-authorship to submitters where their data are used in publications authored by curators, and the submitters have made significant scientific contributions. What constitutes 'significant contribution' can be found in the document published by the ICMJE (International Committee of Medical Journal Editors, 2017).
   c. Acknowledge submitters in publications if 3a and 3b are not possible.

4. **Establish an Oversight Committee (OC):** In the curation of unpublished data practical ethical questions can arise. Therefore, it is necessary to have an independent body (the Oversight Committee) where the curator can direct questions to be discussed and addressed, including requests for large scale access. This is essential where unpublished data is accepted into databases but may not be necessary where all data come exclusively from publications. The OC should be independent of the curation and funding of the database that they oversee. It should be noted that an Oversight Committee differs from Ethics Committees (e.g. Institutional Review Boards (IRBs), Independent Ethics Committees (IECs) and Research Ethics Committees (RECs)) that are charged with ensuring high standards in the ethical conduct of research involving human subjects.

   a. Purpose of OC

      i. To act as an independent forum for the consideration of practical ethical issues arising in the day-to-day work of the database.

      ii. To consider any other matters relating to sharing of unpublished data submitted to the database, in line with local regulations/requirements and recommendations in the field.

   b. Guidance on composition of OC

      i. Members should be independent of the curation and funding of the database they have oversight of, but knowledgeable about the condition and represent the different groups involved, e.g. clinicians, researchers, database curators and lay persons from patient groups.

      ii. At least one member of the OC should have ethics training.

5. **Data collection**

    a. **Published data:** Privacy concerns discussed in this document do not apply to data obtained from publications.

    b. **Consented unpublished data**

        i. Inform submitters of their responsibility to ensure that valid consent has been obtained and that only coded patient IDs are submitted. Coded IDs allow submitters to respond to queries from the curator or to update new information about a specific case.

        ii. Note that completely anonymising patient IDs makes it virtually impossible to update valuable information that subsequently becomes available, either by the submitter or curator.

        iii. For submissions that are not linked to any publicly available source, e.g. data from diagnostic labs (health service labs and commercial sources), clinics/clinicians and sometimes from patients:

    c. Ensure coded IDs are used.

6. **Curation of consented unpublished data**
    a. **Unpublished data:** Received as a query or submitted for inclusion in the database. For example, this data may come from a clinician, genetic counsellor, diagnostic labs or a patient.

        i. If the data is from a query, inform the enquirer that the variant will be included in the database.

        ii. Assign a coded ID to each entry, if there is none already.

        iii. Keep sensitive personal data non-public. This refers to information that is of a private nature that could be used in a discriminatory manner.

        iv. In linking entries to details of the submitter, curators should abide by relevant applicable regulations.

        v. Make variant nomenclature to adhere to HGVS standards (http://www.HGVS.org/varnomen). Keep the original variant description in a separate column. If a variant is ambiguous or does not match the reference sequence, consult with the submitter. Do not place incorrect variant descriptions on the database. If the problem with a variant description cannot be solved, exclude the case.

vi. Publicly viewable data (from submitted unpublished data)

  a) Summarise publicly viewable data to ensure clarity on family relationships.

  b) Curate submitted data to ensure personal details do not identify individuals.

  c) Phenotype information is important for clinical diagnosis. Where phenotype information is available, and efforts have gone into protecting the identity of the individual, details on the phenotype should be displayed.

vii. Non-public data

  a) This section of the database is reserved for confidential information that curators will need to refer to.

  b) Requests to share non-public data should be forwarded to the submitter (see point 7).

7. **Permitting the use of non-public data for scientific/clinical purposes**

   a. **Request from clinician or diagnostic lab**: Curators may receive requests to share non- public information from bona fide clinicians/diagnostic labs who need the information for patient care/diagnostic report. An example may be a new variant with the associated clinical data, segregation information and pathogenicity, which a submitter has requested that these should not be made public until after their impending publication (see point 8).

      i. Forward request to the submitter.

   b. **Request from researcher**

      i. Forward request to the submitter.

8. **Request to keep submitted data non-public:** Some submitters request that data be kept non-public until they are published. Make the submitter aware that:

   a. DNA diagnosis is improved by sharing data on genes, variants and phenotypes; and publicly sharing data offers optimal care to patients and their families.

b. Publishing the variant in the database does not result in the rejection of a subsequent manuscript that mentions the data.

c. Note that searches in publicly available variant databases may return a message indicating a non-public record with a variant at that position is in the database, with the suggestion to contact the curator to receive more details. The following options may be adopted:

    i. Enter data but make the entire entry 'non-public'. Note point 8c above; or

    ii. Enter data but make variant public and associated information 'non-public'. This option should be discussed with the submitter.

d. Any request for information should be forwarded to the submitter.

9. **Request for submitter's details:** Some LSDBs do not link submitter details to unpublished data.

a. Any request for submitter details should be forwarded to the submitter allowing them to respond directly with the requester.

10. **Giving your opinion:** As a curator you will be considered as an 'expert' and will be asked your opinion on the consequences of an identified variant, or other aspects of the disease.

a. If you have a team (clinical and scientific) that is qualified and knowledgeable about the disease, an opinion on the potential consequence of a variant may be given, especially when you (as the curator) have assigned "concluded pathogenicity" to variants listed in your database.

b. If you do not have a team and you do not have in-depth knowledge about the disease, refrain from giving any opinion.

11. **Sharing information with genome browsers:** This increases visibility for your database and should be encouraged.

12. **Managing personal data breach:** Curators may handle sensitive personal data. Where this happens, steps should be taken to ensure personal data breaches are dealt with immediately incidents are identified.
    a. Prior to data curation, curators should identify who to notify in the event of a personal data breach and obtain further procedures on reporting.
    b. On discovering an incident involving personal data breach, file a report of the incident with the appropriate contact.
    c. Follow procedures provided by your establishment and/or infrastructure provider.

13. **Data erasure:** Complete erasure of publicly available data, presented in web-based databases, is impossible as the data may be distributed to cyberspace. However, where database curators have knowingly shared data, e.g. with genome browsers, and submitters have requested data erasure, those recipients should be informed of the request for erasure. The request for data erasure may be due to the withdrawal of consent. Further details on the specific grounds for request of data erasure can be found in the EU General Data Protection Regulation 2016/679.

   When a request for erasure is received from a data submitter:
   a. Ensure you are provided with details that permit the identification of the data to be erased, e.g. the coded ID and a specific factor such as the genetic variant, to ensure the correct data is erased.
   b. Identify and delete the data from your database.
   c. Where the data in question has been shared, e.g. with genome browsers, contact the recipient to inform them of the request for erasure.

Bibliography

Celli, Jacopo, Raymond Dalgleish, Mauno Vihinen, Peter E M Taschner, and Johan T. den Dunnen. 2012. "Curating Gene Variant Databases (LSDBs): Toward a Universal Standard." Human Mutation 33 35(2): 291–97. doi:10.1002/humu.21626.

EU General Data Protection Regulation 2016/679 (GDPR). 2016. http://ec.europa.eu/justice/data-protection/reform/files/regulation_oj_en.pdf

Global Alliance for Genomics and Health. 2016. GA4GH Data Sharing Lexicon. https://www.ga4gh.org/docs/ga4ghtoolkit/regulatoryandethics/GA4GH_Data_Sharing_Lexicon_Mar15.pdf.

International Committee of Medical Journal Editors. 2017. Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals. http://www.icmje.org/icmje-recommendations.pdf.

Knoppers, Bartha Maria. 2014. Framework for responsible sharing of genomic and health-related data. The HUGO Journal 8(1):3.

Mascalzoni, Deborah, Edward S Dove, Yaffa Rubinstein, Hugh J S Dawkins, Anna Kole, Pauline McCormack, Simon Woods, et al. 2015. "International Charter of Principles for Sharing Bio-Specimens and Data." European Journal of Human Genetics :23(6):721-8. doi: 10.1038/ejhg.2014.197.

Povey, Sue, Aida I. Al Aqeel, Anne Cambon-Thomsen, Raymond Dalgleish, Johan T. den Dunnen, Helen V. Firth, Marc S. Greenblatt, et al. 2010. "Practical Guidelines Addressing Ethical Issues Pertaining to the Curation of Human Locus-Specific Variation Databases (LSDBs)." Human 4Mutation 31 (11): 1179-84.
doi:10.1002/humu.21339.

Vihinen, Mauno, Johan T. den Dunnen, Raymond Dalgleish, and Richard G. H. Cotton. 2012. "Guidelines for Establishing Locus Specific Databases." Human Mutation 33 (2): 298–305. doi:10.1002/humu.21646.

World Medical Association. 2016. WMA Declaration of Taipei on Ethical Considerations regarding Health Databases and Biobanks. https://www.wma.net/policies-post/wma-declaration-of-taipei-on-ethical-considerations-regarding-health-databases-and-biobanks/

Appendix 1 – InSiGHT - *Ethical and Privacy Principles in relation to Responsible Sharing of Genomic and Health-Related Data (2018)*

Ethical and Privacy Principles in relation to Responsible Sharing of Genomic and Health-Related Data – example from InSiGHT – how a framework approach can be helpful. A flexible approach assists database curators and their collaborators to address changing local and international changes in regulations, as well as changing community expectations. While policies in this area benefit from regular discussion and review for continuous improvement, this should not be an onerous task; a framework approach helps to keep the focus on improving outcomes. https://www.insight-group.org/content/uploads/2018/06/EthicsFramework.pdf

Appendix 2 - Example of database policy

Example of database policy from ORAI1base (Variation registry for Severe combined immunodeficiency) at
http://structure.bmc.lu.se/idbase/ORAI1base/?content=_db_policy/IDbases.

DATABASE POLICY

The ImmunoDeficiency Variation Databases (IDbases) and other variation databases maintained at the Protein Structure and Bioinformatics Group (PSB), Lund University, are maintained and provided as a public service for academic community.

Individuals submitting data to and using the variation databases managed by the PSB should be aware of the following:

1. *The PSB has a uniform policy of free and unrestricted access for academic community to all of the data records their databases contain. Scientists worldwide can access these records to plan experiments or publish any analysis or critique. Appropriate credit is given by citing the database. Instructions for citing are provided in each individual database.*

2. *The databases are intellectual property of the PSB. Details are available for Copyright and Liability.*

3. *Corrections of errors and update of the records by authors are welcome and erroneous records may be removed from the next database release.*

4. *Submitters are advised that the information displayed on the Web sites maintained by the PSB is fully disclosed to the public. It is the responsibility of the submitters to ascertain that they have the right to submit the data. This applies also the appropriate consent from the patient and/or family.*

5. *Beyond limited editorial control and some internal integrity checks, the quality and accuracy of the record are the responsibility of the submitting author, not of the database. The databases will work with submitters and users of the database to achieve the best quality resource possible.*

6. *Data in the PSB mutation databases may be shared with central repositories according to published Human Genome Variation Society guidelines.*

7. *The information provided on this site is designed to support, not replace, the relationship that exists between a patient/site visitor and his/her existing physician.*

8. *We keep the confidentiality of the data relating to individual patients and visitors to the web site, including their identity. No data is collected that would allow identification of the patients for whom information is stored and distributed in the database. We do not share any information about database visitors with third parties. As database curators and owners we undertake to honour or exceed the legal requirements of medical/health information privacy that apply in Sweden.*

9. *The database does not host any advertisements.*