The Royal College of Pathologists of Australasia
ABN 52 000 173 231

Durham Hall 207 Albion Street Surry Hills NSW 2010 Australia
Telephone **61 2 8356 5858** Facsimile **61 2 8356 5828**

RCPA
The Royal College of Pathologists of Australasia

THE ROYAL COLLEGE OF PATHOLOGISTS OF AUSTRALASIA (RCPA)

Standards for clinical databases of genetic variants.

1st Edition 2014 (version 1.0)

www.rcpa.edu.au

Online copyright

First published December, 2014, 1st Edition (version 1.0)

Disclaimer

The Royal College of Pathologists of Australasia ("the College") has developed these Standards for clinical databases of genetic variants to assist in ensuring the quality, accuracy, security, and utility of DNA variant databases used for reporting for clinical purposes.

While there are indicators of ‘minimum requirements’ (Standards) and ‘recommendations’ (Commentary), the Standards are a first edition and have not been through a full cycle of use, review and refinement.

Therefore, in this edition, the inclusion of “standards” is provided as an indication of the opinion of the expert authoring group, but should not be regarded as definitive or as widely accepted peer professional opinion. Specifically, these terms do not carry regulatory weight with regard to laboratory accreditation. The use of these standards and guidelines is subject to the health professional’s judgement in each individual case.

The College makes all reasonable efforts to ensure the quality and accuracy of the Standards and to update the Standards regularly. However subject to any warranties, terms or conditions which may be implied by law and which cannot be excluded, the Standards are provided on an "as is" basis. The College does not warrant or represent that the Standards are complete, accurate, error-free, or up to date. The Standards do not constitute medical or professional advice. Users should obtain appropriate medical or professional advice, or, where appropriately qualified, exercise their own professional judgement relevant to their own particular circumstances. Users are responsible for evaluating the suitability, accuracy, currency, completeness and fitness for purpose of the Standards.

Except as set out in this paragraph, the College excludes: (i) all warranties, terms and conditions relating in any way to; and (ii) all liability (including for negligence) in respect of any loss or damage (including direct, special, indirect or consequential loss or damage, loss of revenue, loss of expectation, unavailability of systems, loss of data, personal injury or property damage) arising in any way from or in connection with the Standards and Guidelines or any use thereof. Where any statute implies any term, condition or warranty in connection with the provision or use of the Standards and Guidelines, and that statute prohibits the exclusion of that term, condition or warranty, then such term, condition or warranty is not excluded. To the extent permitted by law, the College's liability under or for breach of any such term, condition or warranty is limited to the resupply or replacement of services or goods.

REVISION HISTORY

Version	Date	Document	Author
0.0	January 2014	Draft Framework	Vanessa Tyrrell
0.1	February 2014	Draft Framework	Vanessa Tyrrell
0.2	March 2014	Draft Standards	Vanessa Tyrrell
0.3	April 09 2014	Draft Standards	Vanessa Tyrrell
0.4	May 05 2014	Draft Standards	Vanessa Tyrrell
0.5	May 08 2014	Draft Standards	Vanessa Tyrrell
0.6	May 29 2014	Draft Standards	Vanessa Tyrrell
0.7	June 12 2014	Draft Standards	Vanessa Tyrrell
0.8	June 16 2014	Draft Standards	Vanessa Tyrrell
0.9	June 30 2014	Draft Standards	Vanessa Tyrrell
0.10	July 09 2014	Draft Standards	Vanessa Tyrrell
0.11	July 31 2014	Draft Standards	Vanessa Tyrrell
0.12	August 11 2014	Draft Standards	Vanessa Tyrrell
0.13	August 18 2014	Draft Standards	Vanessa Tyrrell
0.14	September 20 2014	Draft Standards	Vanessa Tyrrell
0.15	October 2014	Draft Standards	Vanessa Tyrrell
0.16	October 23 2014	Draft Standards	Vanessa Tyrrell
0.17	November 28 2014	Draft Standards	Vanessa Tyrrell
0.18	December 09 2014	Final Draft	Vanessa Tyrrell
1.0	December 18 2014	First edition for release	Vanessa Tyrrell

Contents

Definitions

Custodian	
Normative (as applied to commentaries and appendices)	Prescriptive or mandatory and the material carries the same weight as the Standards to which it is attached.
Informative (as applied to commentaries and appendices)	The material is presented to assist in the application or interpretation of the Standards to which it is attached.
Sequence variant	The entry (which would normally be the logical record) within a DNA variant database
Cloud	
Clinical	
Clinical Grade Sequencing Data	DNA sequencing performed in an accredited facility to defined quality standards, undertaken to inform a particular clinical indication
Curation	The activity of managing and promoting the use of data from its point of creation, to ensure it is fit for contemporary purpose, and available for discovery and re-use[1].
Database	A computer structure that houses a collection of related data
Database Management System (DBMS)	Determines the data model, storage, maintenance and retrieval of data, security and other functions necessary to use the database.
DNA variant database	The structured collation of records of variations in DNA sequence or structure identified in patients or subjects versus a specified reference sequence.
Genetic Variant	An alteration in the DNA sequence compared to a reference sequence, the significance of which is often unclear.
Healthcare database	A specific class of database, the primary use case of which is to store data for use in healthcare environments for the clinical management of patients. Such databases typically hold personally identifiable and protected health information. Examples include health information management systems and Electronic Health Records (EHR)
Identity	The whole of the characteristics of a document or a record that uniquely identify it and distinguish it from any other document or

[1] Lord, Philip, and Alison Macdonald. e-Science Curation Report: Data curation for e-Science in the UK: an audit to establish requirements for future curation and provision. Digital Archiving Consultancy Limited, 2003.

	record. With integrity, a component of authenticity[2].
Integrity	The quality of being complete and unaltered in all essential respects. With identity, a component of authenticity[3].
Knowledge data base	A collection of information about the data stored in the database; expressing what we know about a particular piece of data (e.g.: variant is the data and the information is what we know about pathogenicity, inheritance patterns, population distribution, etc,)
Deidentified data	Data reasonably disconnected from the identity of a person according to the requirements of the Privacy Amendment Act 2012 or equivalent
NPAAC	National Pathology Accreditation Advisory Council (give ref or URL)
Orthogonal search	The combination of two or more searches whereby the method of search was independent of the previous method/s of search
Ontology	(using a shared vocabulary to denote the types, properties, and interrelationships of concepts within a domain)
Patient identifier	A string (alphanumeric, numeric or alphabetic) that identifies a patient unambiguously to the data submitter but maintains anonymity of the Patient to all other users of the database
Personal Information	Definition of "personal information" in the Australian context can be found in the Privacy Amendment Act 2012, and includes all information *or opinion* about an individual who is identified or reasonably identifiable, whether such information/opinion is true or not.
Preservation	An activity within archiving in which specific items of data are maintained over time so that they can still be accessed and understood through changes in technology[4].

[2] InterPARES 2 Terminology Database. http://www.interpares.org/ip2/ip2_terminology_db.cfm

[3] InterPARES 2 Terminology Database. http://www.interpares.org/ip2/ip2_terminology_db.cfm

[4] Lord, Philip, and Alison Macdonald. e-Science Curation Report: Data curation for e-Science in the UK: an audit to establish requirements for future curation and provision. Digital Archiving Consultancy Limited, 2003.

This page is intentionally blank

Scope

***Purpose* – This document presents a broad set of standards for sequence variant databases used for clinical purposes. It complements existing NPAAC reference materials for laboratory accreditation, and**

- Is applicable to all databases of genetic variants used for clinical purposes
- Provides a benchmark for the structure of such databases, including standards for ontologies and minimum content requirements.
- Provides standards for the tools which manipulate the data in such databases.
- Assists laboratory professionals in identifying databases of appropriate quality for clinical purposes.
- Set standards for data sharing for clinical purposes: including bi-directional data transfer, interfacing, and other collaborative methods within the boundaries of existing privacy laws

***Benefits* – Patient care will be improved by**

- encouraging the collation of curated information about DNA variants identified in patient care
- facilitating the accurate interpretation of analytical results
- enabling the sharing of curated data between laboratories, thereby developing a broader repository of data to inform clinical interpretation
- improving the efficiency of interpretation and timely reporting to clinicians.

***Exclusions* – This document does not detail requirements regarding:**

- The phenotypic information that is to be stored. Clinical laboratories generally lack control over the quality and volume of clinical information about a patient provided to them. This document notes the phenotype data fields which should be entered if it is available, and recommends the implementation of standard terms to describe phenotypes i.e. a defined ontology.

- The ownership or physical location of databases.

- The implementation of these standards.

Introduction

This document has been developed by the Royal College of Pathologists of Australasia (RCPA) in collaboration with the Human Variome Project (HVP), and the Human Genetics Society of Australasia (HGSA). It presents a set of standards to be used in conjunction with other reference materials (listed below) to promote the quality, accuracy, security, and utility of DNA variant databases used for clinical purposes.

This document is not designed for databases which

- are used for other purposes such as research or public health repositories, or
- electronic health records which include open identification of patients and act as record of their management.

The fundamental principle underpinning this document is that DNA sequence variant databases intended for use in clinical diagnostic testing should be developed, curated, and maintained as safe, secure, and accurate repositories of genomic data.

These Standards have been developed with reference to current and proposed Australian regulations and standards from the International Organisation for Standardisation (ISO), including *AS ISO 15189 Medical laboratories – Requirements for quality and competence.* The standards should be used for guidance where accessing and or assessing databases outside the Australian accreditation framework.

In addition to these standards, existing NPAAC Requirements apply to all Laboratories seeking accreditation for medical testing in Australia, and must be applied in conjunction with jurisdictional and other regulatory requirements.

In each section of the document, points deemed important for practice are identified as ‘Standards’ or ‘Commentaries’.

- A Standard is the minimum requirement for a procedure, method, staffing resource or facility that is required before a Laboratory or other accreditable entity can attain accreditation – Standards are printed in bold type and prefaced with an ‘S’ (e.g. **S2.2**). The word ‘**must**’ in each Standard within this document indicates a mandatory requirement for practice. Commentary is provided to give clarification to the Standards as well as to provide examples and guidance on interpretation.

- Commentaries are prefaced with a ‘C’ (e.g. C1.2) and are placed where they add the most value. Commentaries may be normative or informative depending on both the content and the context of whether they are associated with a Standard or not. Note that when comments are expanding on a Standard or referring to other legislation, they assume the same status and importance as the Standards to which they are attached. As a general rule, where a Commentary contains the word ‘**must**’ then that Commentary is considered to be **normative**.

Please note that any Appendices attached to this document may be either normative or informative in nature and should be considered to be an integral part of this document.

Background

It has become routine practice to compare a DNA variant identified during clinical testing with the description and interpretation of variants recorded in databases, and using this information to guide clinical interpretation of the patient's variant. Although numerous DNA variant databases already exist, there are few that meet the accuracy and reliability required for clinical diagnostics[5]. Current databases are of variable quality and may contain errors in variant calls, non-standardised nomenclature, incomplete pathogenicity associations and limited phenotypic information linked to genomic data. These all represent limitations and risks to the quality of pathology reporting and to patient care.

The increasing use of genomic technologies such as massively parallel sequencing is producing increasing volumes of data that need to be recorded, interpreted, and shared. This provides an additional risk of propagating errors, so that an incorrect or incomplete database entry is used to interpret other database entries or reports, which are in turn compromised. With the growing interdependence of databases for clinical reporting, the integrity of these databases becomes a critical issue.

The Standards development project

There are numerous initiatives directed at the integration of genomic technologies into mainstream clinical diagnostics, however there are no specific standards or equivalent mechanisms to assure the quality or guide the accreditation of DNA variant databases. An Australian project led by the Royal College of Pathologists of Australasia (RCPA) in collaboration with the Human Genetics Society of Australasia (HGSA), and the Human Variome Project (HVP), developed these standards for DNA variant databases intended for clinical use. This project was supported by an unrestricted grant from the Australian Department of Health's Quality Use of Pathology Program (QUPP).

The standards are a broad reaching set of standards that are sympathetic to the rapidly changing landscape of clinical genomics, and that can be applied to assess extant databases and to guide the development of new databases. The fundamental goal of the document is to provide a quality framework for the oversight of DNA variant databases. These standards complement existing laboratory standards and accreditation requirements, act as a guide to identify a quality database, assist the development of new databases, and in improving existing databases that have been developed in non-clinical environments.

Maintaining the quality, accuracy, and clinical relevance of DNA variant databases will reduce the risk of misinterpretation and inappropriate reporting of variants, promote the sharing of data which can be trusted for clinical use, and accelerate the delivery of actionable clinical reports to improve patient care.

[5] Saunders, C.J., Miller, N.A., Soden, S.E., Dinwiddie, D.L., Noll, A., Alnadi, N.A., Andraws, N., Patterson, M.L., Krivohlavek, L.A., Fellis, J., Humphray, S., Saffrey, P., Kingsbury, Z., Weir, J.C., Betley, J., Grocock, R.J., Margulies, E.H., Farrow, E.G., Artman, M., Safina, N.P., Petrikin, J.E., Hall, K.P., Kingsmore, S.F., 2012. Rapid whole-genome sequencing for genetic disease diagnosis in neonatal intensive care units. Sci. Transl. Med. 4 (154), 135.

1 Purpose

These standards are intended to be a high level, generic set of standards which are applicable to sequence variant databases serving clinical purposes. In order to establish the context in which a sequence variant database may be utilised, the purpose must be clearly articulated. It is important for database users to know what is available within the database, and what to expect from the database.

S1.1 The intended purpose of the database must be clearly defined and documented, and be made available through appropriate media such as internally controlled documentation and, if publicly accessible, on the database website.

C1.1 The elements which characterise a DNA variant database must include:

- The context in which the database is intended to be used e.g. clinical diagnostic, clinical research, or clinical theranostic purposes, and whether access will be restricted to particular users (in-house, password protected) or in the public domain.
- The nature of the information included in the database. This description may include:
  - disease specific information,
  - gene specific information,
  - phenotype description),
  - whether the database contains deidentified data only (knowledge database), personal data (Healthcare database), or both related to submitted variants.
  - germline or somatic data, or both
- There must be a clear distinction between a database for use by in-house staff only i.e. all users are accountable to the custodian, versus a public database that may be used by people who are not accountable to the custodian (See S2.2). The requirements for databases in these two settings differ. If a database is transitioned from being in-house to being public, there must be a review of all aspects of the database operation to ensure that the different requirements are met.
- The basic limitations of the database including, but not limited to, criteria for inclusion and exclusion of data, the types of data included, the level of curation undertaken, and the mode and level of access that is facilitated.
- The technical and administrative functions that the database custodian has implemented to ensure the integrity of the data held by the database.

Examples of databases with well defined purpose include:

- COSMIC: http://cancer.sanger.ac.uk/cancergenome/projects/cosmic/about
- ClinVar: http://www.ncbi.nlm.nih.gov/clinvar/intro/
- DMuDB:http://www.ngrl.org.uk/Manchester/projects/dmudb

- Decipher: http://www.decipher.sanger.ac.uk
- BIC: http://research.nhgri.nih.gov/bic

S1.2 There must be a clearly defined procedure for regular review of the purpose and use of the database.

C1.2 Regularity of the review of the purpose of the database should be stated, and the date when the review of the purpose and use of the database was last addressed **must** be readily accessible.

S1.3 There must be a clearly defined procedure for reviewing the quality processes to ensure they are appropriate for any revision of the purpose and use of the database.

C1.3 Quality parameter audits must be readily available and provide a clear description of the frequency of the audit cycle, what is undertaken, the degree to which the audit parameters have been achieved (i.e. were the minimum specifications achieved and/or exceeded), and actions arising from the outcomes of each audit.

2 Governance

S2.1 The governance structure for the operation of and access to the database must be clearly defined and be readily accessible

C2.1 The governance structure for the database must address oversight of all aspects of the database including privacy, secure access and sharing, content, quality, accuracy, curation, and clinical utility.

S2.2 A database custodian [ownership] must be identified, this custodian being accountable and responsible for the operation of the database, and who is accountable to a clinical institution or incorporated entity

C2.2(i) The custodian **must** ensure adequate curation and management of the data such that it is secure, accurate and subject to regular review.

C2.2(ii) The custodian **must** have sufficient authority to carry out or delegate actions as required.

C2.2(iii) The custodian is the entity with responsibility and accountability for the operation of the database. This may be a position, committee, board of the host institution, or incorporated entity which can ensure transparency, and business continuity.

S2.3 The custodian must appoint an appropriately qualified person as curator whose specific role is to oversee management of the database and be accountable to the custodian. The curator must have the requisite authority to carry out and/or delegate the tasks required.

C2.3(i) The custodian **must** ensure the curator has the appropriate skillset, knowledgebase, and experience to oversee the content and operations of the database (See Appendix – Curation); and provide access to continuing professional development

C2.3(ii) the custodian and curator may delegate responsibilities to other parties, but such delegation must incur accountability to the curator or custodian.

C2.3(iii) In instances where the database is small, the custodian may also be the curator. It should be noted however that this does not necessarily constitute good governance.

C2.3(iv) If a multidisciplinary committee is appointed to assist the curator, the terms of reference of the committee, its constituent members (by position and expertise, not necessarily by name), and operational aspects such as meeting schedules and decision analytics models should be clearly defined and readily available for reference .

C2.3(v) Other approaches to curation may be considered, however accountability and responsibility to demonstrate efficiency and quality of the curation process reside with the custodian.

S2.4 The custodian must ensure that an appropriate policy regarding intellectual property held in the database is in place, that the policy is readily accessible to current and potential users of the database, and that implementation of the policy is audited.

C2.4(i) In the case of a public database, a copyright notice and method of citation may be included to protect material rights.

C2.4(ii) In the case of a public database, a disclaimer notice may be displayed to limit or exclude liability. Acknowledgement of this disclaimer may be a requirement for access to and use of the database

S2.5 The curator must ensure that data sets submitted to the database comply with relevant professional standards and/or privacy legislation.

C2.5(i) There **must** be a defined mechanism to facilitate communication between the authorised user and the curator. This should include the ability to provide free text information which may be periodically reviewed. This should be included in audit processes.

C2.5(ii) There should be defined procedures relating to both clinical, technical, and regulatory issues; with regular quality audits to minimise occurrence or recurrence of operational issues identified by users or operators.

S2.6 When associating variant records in another or multiple databases, the external database/s must be audited prior to use.

C2.6(i) An audit should be performed on external databases prior to utilisation of any data/information from it to reduce the risk of introducing errors into the database

C2.6(ii) Results or certificates of audits of a database conducted by a trusted third party may be accepted as evidence to minimise repetition and work burden .

C2.6(iii) To increase efficiency and accuracy of information sourced from external databases, automation of the interrogation function to mine and update most recent data in the database is recommended. Such processes must be subject to control and regular audits.

2.7 The custodian must ensure there is appropriate ethical oversight of the database through an appointed ethics committee

C2.7(i) The ethics committee may be a dedicated ethics committee, an institutional committee, or hosted by a professional society / organisation

C2.7(ii) Useful reference materials include:

- National Pathology Accreditation Advisory Council Requirements for Medical Pathology Services (First Edition 2013)[6]
- The NHMRC National Statement of Ethical Conduct in Human Research[7].

S2.8 The custodian must ensure there is a procedure to follow in the event that the database no longer meets the stated purpose, or closes.

C2.8(i) The demise of a database may include abandonment, falling in to disuse due to reduced relevance, or being closed / discontinued due to obsolescence.

C2.8(ii) There should be a clear and detailed policy for transfer of data to ensure provision of continuity of access to data in the event of demise of a database. This may occur due to loss of funding to maintain the database, loss or change of custodianship, or force majeure.

C2.8(iii) There should be a clear and detailed policy for destruction of data in the event the database is closed because it no longer meets a need or is obsolete.

C2.7(iv) The RCPA Guideline *Privacy Guidelines – Managing Healthcare Information in Laboratories*[8] discusses privacy principles related to pathologists and their laboratories. It also addresses the application of these principles when a pathology practice faces a change in business circumstance or closure.

[6] National Pathology Accreditation Advisory Council Requirements for Medical Pathology Services (First Edition) Tier 2 document, Commonwealth of Australia (2013) Online ISBN: 978-1-74241-914-5

[7] National Statement on Ethical Conduct in Human Research 2007 (Updated March 2014).
The National Health and Medical Research Council, the Australian Research Council and the Australian Vice-Chancellors' Committee. Commonwealth of Australia, Canberra.

[8] http://www.rcpa.edu.au/getattachment/a631a573-0d07-4bd4-ba67-cfe545618dd1/Managing-Privacy-Information-in-Laboratories.aspx

3 Establishment of Databases

This section outlines the basic requirements to ensure a fully functional and efficient database which is capable of maintaining data integrity in a secure environment.

S3.1 The infrastructure and storage capabilities of the custodian institution must be fully functional.

C3.1(i) The authenticity of the data **must** be maintained. The authenticity of a digital record refers to its trustworthiness i.e. that it is what it purports to be and it is free from tampering or corruption. Authenticity has two components:

- integrity: the quality of being complete and unaltered in all essential aspects, and
- identity: the characteristics of a record that uniquely identify it and distinguish it from any other record.

Any unintended change to a record or its identifiers as a result of storage, retrieval, processing and operation, including malicious intent, unexpected hardware failure, is a failure of data authenticity

C3.1(ii) The underlying technical infrastructure used to implement the database must be capable of supporting the functionality required by this standard.

- A spreadsheet is not considered to be sufficient means of storing data.
- Examples which may meet the requirements include:
  - SQLite3, Microsoft SQL Server, PostgresSQL, MySQL, MongoDB, Apache Cassandra.

C3.1(iii) If the database is designed for public access, there should be a web interface to enable efficient access by users, and facilitate data gathering, sharing and report retrieval.

C3.1(vi) Data should be exportable and compatible with other data repositories used in healthcare to allow for efficient data sharing.

S3.2 Any modifications or updates **must** be new records, version controlled, and linked to the initially created record.

S3.3 **C**omplete provenance information for all records must be stored within the database to ensure that records are effectively permanent and the state of any record at any point in time can be viewed.

C3.3(i) Provenance information should indicate:

- the origin;
- intermediate source(s); and
- complete modification history of the data.

C3.3(ii) The provenance information should be visible to all users, with the exception that any free text which might identify the patient should only be visible to authorised users (see 3.7).

S3.4 There must be a policy regarding audit of the database. This policy must be readily available, together with the last date on which the audit was performed.

C3.4(i) There must be a complete audit trail of changes to any record to ensure that all records are effectively permanent.

C3.4(ii) The complete audit trail should be visible to viewers, with the exception that any free text which might identify a patient should only be visible to authorised users (see S3.7).

C3.4(iii) A spreadsheet will not support the journaling requirements for audit trails and is therefore not an appropriate form of media for use as a database.

S3.5 The database backup must contain all required information to reinstate the database with minimal reconstruction in the event of a catastrophic failure.

C3.5(i) The data backup policy **must** ensure that the hardware, software, geographic location of redundant dataset/s, network accessibility, and personnel responsible for data backup are consistent with a high level of protection of patient privacy and confidentiality.

S3.6 The database must be backed up at regular intervals to minimise loss of data and the required reconstruction in the event of a catastrophic failure.

C3.6(i) There must be a policy which specifies the frequency and type of backups, together with regular audits to ensure that this policy is implemented.

C3.6(ii) Backups should be in a form which can be reloaded in to a database with minimal effort.

C3.6(iii) A three tier back up system should be employed

(a) The original database (current in daily use)
(b) A separate local copy via network or manual
(c) A separate offsite copy (i.e.: a separate location such as an encrypted portable hard drive or cloud based solution)

S3.7 To minimise the risk of inadvertent disclosure of private information, free text must not be included in a record that may be accessed by users without appropriate access authorisation.

C3.7(i) Free text data that is submitted must be reviewed by the curator to exclude personal identifiers before a record is made public.

4 Privacy, Confidentiality, Ethics, and Data Security

There are significant ethical, legal, and social issues that must be considered and handled responsibly when developing, operating and de-commissioning a sequence variant database. These issues relate to concepts of privacy and confidentially for both patients and health system workers, the right of autonomy for individual patients and the related right to make decisions about the way information about them and their health care is used and disclosed, as well as broader societal concerns regarding the public interest and the benefits that can be derived from the use of genetic variation information. The responsible handling of these issues in Australia is mandated by a complex mix of state, territory and commonwealth legislation—often informed by international declarations and treaties from bodies such as the OECD, APEC and UNESCO—regulation, advice from the Office of the Australian Information Commissioner, relevant professional practice standards and professional codes of ethics.

Exactly which components of the above mix apply to the development, operation and de-commissioning of any one database is dependent on the jurisdiction under which the database custodian operates, whether the database custodian is a health service provider and if it operates in the public or private sector, the annual turnover of the database custodian, the primary use for which the data included in the database was collected, the intended use or uses of the database—i.e. for clinical or research purposes—and whether the information stored in the database is considered "personal information" under the *Privacy Act 1988* (Cth).[9] The RCPA Guidelines *The Ethical and Legal Issues in Relation to the Use of Human Tissue and Test Results in Australia*[10] and *Managing Privacy Information in Laboratories*[11] provide a more in-depth discussion of how these issues are controlled and regulated in Australia for accredited pathology laboratories.

The development, operation and de-commissioning of databases where the database custodian is located in a country other than Australia (international databases) will be regulated under the legislative and regulatory requirements of that country. Importantly, if an Australian entity transfers "personal information" about any Australian individual to international databases, then the Australian entity is responsible for ensuring that the international databases only use or disclose that information in accordance with the Australian Privacy Principles[12] or be "subject to a law, or binding scheme, that has the effect of protecting the information in a way that, overall, is at least substantially similar to the way in which the Australian Privacy Principles protect the information."

This document covers sequence variant databases used for clinical purposes and, as such, the Standards and Commentaries in this section should be considered to only apply to such databases. The Standards and Commentaries may not be applicable to databases that are used for research purposes.

[9] http://www.comlaw.gov.au/Series/C2004A03712

[10] http://www.rcpa.edu.au/getattachment/b52a239d-c5da-4f9c-8670-c65b14380e8f/Ethical-Legal-Issues-Use-Human-Tissue-Test-Results.aspx

[11] http://www.rcpa.edu.au/getattachment/a631a573-0d07-4bd4-ba67-cfe545618dd1/Managing-Privacy-Information-in-Laboratories.aspx

[12] http://www.oaic.gov.au/privacy/privacy-resources/privacy-fact-sheets/other/privacy-fact-sheet-17-australian-privacy-principles

S4.1 The database custodian must comply with relevant local legislation, regulations and professional practice standards in all aspects of the development, operation and de-commissioning of the database.

C4.1(i) The collection, storage, use and disclosure of all information **must** comply with all legislation and regulations that deal with the privacy and confidentiality of:

- the patients from whom the data is derived; and
- the laboratories, clinicians and laboratory staff who are submitting the data.

The RCPA Guideline *Privacy Guidelines – Managing Healthcare Information in Laboratories*[13] discusses privacy principles related to pathologists and their laboratories.

C4.1(ii) In deciding what legislative and regulatory requirements must be met, a decision **must** be taken as to whether any information collected, stored, used or disclosed by the database custodian would constitute "personal information" under the *Privacy Act 1988* (Cth). The Act defines personal information as information or an opinion about an identified individual, or an individual who is reasonably identifiable:

- whether the information or opinion is true or not; and
- whether the information or opinion is recorded in a material form or not. [14]

The Act only applies to personal information. Guidance exists for what constitutes personal information under the Act. Whether an individual can be identified or is reasonably identifiable depends on context and circumstances. While it may be technically possible for an agency or organisation to identify individuals from information it holds, it may not be practical to do so. For example, logistics or legislation may prevent such linkage. In these circumstances, individuals are not 'reasonably identifiable'. Whether an individual is reasonably identifiable from certain information requires a consideration of the cost, difficulty, practicality and likelihood that the information will be linked in such a way as to identify him or her.[15]

De-identified information is not 'personal information.'[16] The Office of the Australian Information Commissioner provides guidance on what constitutes de-identification of data.

[13] http://www.rcpa.edu.au/getattachment/a631a573-0d07-4bd4-ba67-cfe545618dd1/Managing-Privacy-Information-in-Laboratories.aspx

[14] *Privacy Act 1988* (Cth) s 6 (definition of 'personal information'). http://www.comlaw.gov.au/Series/C2004A03712.

[15] Explanatory Memorandum, Privacy Amendment (Enhancing Privacy Protection) Bill 2012 (Cth) 61. http://www.comlaw.gov.au/Details/C2012B00077/Explanatory%20Memorandum/Text.

[16] Office of the Australian Information Commissioner, Australian Privacy Principles guidelines. B.53 page 11. http://www.oaic.gov.au/images/documents/privacy/applying-privacy-law/app-guidelines/APP-guidelines-combined-set-v1.pdf.

De-identification involves removing or altering information that identifies an individual or is reasonably likely to do so. Generally, de-identification includes two steps:

- removing personal identifiers, such as an individual's name, address, date of birth or other identifying information, and
- removing or altering other information that may allow an individual to be identified, for example, because of a rare characteristic of the individual, or a combination of unique or remarkable characteristics that enable identification.

De-identification may not altogether remove the risk that an individual can be re-identified. There may, for example, be a possibility that another dataset or other information could be matched with the de-identified information. The risk of re-identification must be actively assessed and managed to mitigate this risk. Relevant factors to consider when determining whether information has been effectively de-identified could include the cost, difficulty, practicality and likelihood of re-identification.[17]

NHMRC Guidelines approved under Section 95A of the Privacy Act 1988 (the Guidelines)[18] provide a framework to ensure privacy protection of health information that is collected, used or disclosed in the conduct of research and the compilation or analysis of statistics, relevant to public health or public safety, and in the conduct of health service management activities. The Guidelines form part of compliance requirements under the Australian Privacy Principles established in the Privacy Act 1988 (Cth).

C4.1(iii) If data are being submitted from outside the jurisdiction of the custodian, the requirements of other jurisdictions should be considered e.g. HIPAA[19], GINA[20]

C4.1(iv) If the database is stored in the Cloud, the regulatory requirements of the jurisdiction in which the custodian is located must be met.

C4.1(v) Where the purpose of the database includes medical research, the custodian **must** ensure that the management of the database also complies with the NHMRC national statement on ethical conduct in human research[21] or local equivalent.

S4.2 The database must have a readily accessible policy regarding the management of information that reflects the purpose of the database.

C4.2(i) Disclosure in the context of this document means authorised access to data by a third party where the data remains in the control of the

[17] Ibid B.54-55 page 12.

[18] Guidelines approved Under Section 95A of the Privacy Act 1988, National Health and Medical Research Council, Commonwealth of Australia, March 2014.

[19] Health Insurance Portability and Accountability Act.

[20] Genetic Information Non-Discrimination Act.

[21] NHMRC national statement on ethical conduct in human research.

database custodian or the sharing of data with a third party by the database custodian whereby the database custodian relinquished control of the data.

C4.2(ii) The information policy **must** include, but not be limited to, descriptions of how the data are collected, stored, used and disclosed by the database custodian.

C4.2(iii) If the database collects, stores, uses or discloses personal information, the information policy **must** address how the personal information is managed.

The Australian Privacy Principles [22] contain guidance on what information such a policy must contain.

C4.2(iv) The information policy should include information on how informed consent is collected for the collection, storage, use and disclosure of the data included in the database.

The RCPA Guideline *Privacy Guidelines – Managing Healthcare Information in Laboratories* (March 2014)[23] contains guidance on the consent requirements under the *Privacy Act 1998* (Cth).

C4.2(v) If a determination is made that informed consent is not required for the collection, storage, use, and/or disclosure of personal information, the information policy **must** contain information on who made the determination and the reasoning used to justify the determination.

The *Privacy Act 1988* (Cth) includes provisions for exceptions to the requirement to collect informed consent for the collection of personal information, including when such information is collected for:

- the compilation or analysis of statistics relevant to public health or public safety;
- the management, funding or monitoring of a health service.[24]

C4.2(vi) The information policy should address the manner of any data de-identification employed, as appropriate to the database. This should include a justification of the effectiveness of the de-identification techniques employed and an assessment of the risk of unauthorised re-identification of patient data. The policy should address data that are shared externally, e.g. by upload to a central data repository, to other databases, or other third-party users.

C4.2(vii) The International Code of Conduct for Genomic and Health Related Data Sharing, developed by the Regulatory and Ethics Working Group, Global Alliance for Genomics and Health (GA4GH), provides a

[22] http://www.oaic.gov.au/privacy/privacy-resources/privacy-fact-sheets/other/privacy-fact-sheet-17-australian-privacy-principles.

[23] http://www.rcpa.edu.au/getattachment/a631a573-0d07-4bd4-ba67-cfe545618dd1/Managing-Privacy-Information-in-Laboratories.aspx.

[24] *Privacy Act 1988* (Cth) s 16B(2)(a)

principled and practical framework for the disclosure of genomic and health-related data[25]

S4.3 There must be mechanisms in place to control disclosure of data held in the data repository.

C4.3(i) The benefits of open sharing of data must be weighed against the risks to the privacy and confidentiality of the patients from whom the data in the database is derived. Where these risks outweigh the benefits, due consideration should be given to adequate deidentification of data and access to data should be through a register of approved users, rather than full open access. This should be driven by the clinical need to utilise the data for clinical or translational research and or clinical diagnostic purposes. Refer to S4.4 for descriptions of database user tiers.

C4.3(ii) Methods to ensure confidentiality should take into account the nature of the entity with whom data is being shared, the nature of the data being shared, and the use that the data will be put to. A risk-based approach, following current best practice guidelines should be taken.

S4.4 The database must be protected to ensure data security and to protect privacy and confidentiality of individuals.

C4.4(i) Users should only be permitted to gain access to the information that they are entitled to view. The management of access involves issues of computing security that lie beyond the scope of this document, and will involve liaison between the curator and the host institution's IT management. The National eHealth Transition Authority (NEHTA) has identified a number of standards that are pertinent to this issue (see management for Australian Clinical Quality Registries[26]).

C4.4(ii) The database **must** comply with relevant health information systems and security standards including:

- Health informatics – Functional and structural roles ISO/DIS21298 (draft international standard 2014)

C4.4(iii) Mechanisms should include password protected access, data encryption, licensing/certification for access, application for access, and access audit trails.

C4.4(iv) Physical access to the underlying technical infrastructure on which the database is hosted **must** be controlled to ensure security of information.

[25] **International code of conduct for genomic and health-related data sharing.** *The HUGO Journal* June 14 2014, **8**:1 doi:10.1186/1877-6566-8-1

[26] Australian Commission on Safety and Quality in Health Care (2012), *Infrastructure and Technical Standards for Australian Clinical Quality Registries,* ACSQHC, Sydney.

C4.4(v) The database must have a defined mechanism for managing appropriate access to different types of information held in the database.

For example, a four tier access level strategy could be considered for any database which operates beyond the scope of an individual laboratory, thereby facilitating the protection of the database contents**.**

(a) **Level 1: Unregistered Viewer:** This user has access to only collated information with no details regarding the patient sample, or laboratory submitting the information

(b) **Level 2: Registered Viewer:** this user has access to individual reports of the variant, including patient identifier and laboratory submitting the information. However, this information is read-only.

(c) **Level 3: Registered Submitter:** this user has access to read-only information (as for a registered viewer) and is also able to submit information from a specific laboratory.

(d) **Level 4: Curator:** access to all information, and the ability to annotate certain fields.

(e) DBAdmin: Administrative and operational access (IT staff).

S4.5 The privacy of laboratories, clinicians and laboratory staff who are submitting information to the database must be maintained.

C4.5(i) The RCPA Guideline document Privacy Guidelines – Managing Healthcare Information in Laboratories[27] discusses privacy principles related to pathologists and their laboratories.

C4.5(ii) Where it is common practice to list submitters to a database, authorisation to publicly list the submitter must be obtained from each individual submitter.

S4.6 The database custodian must ensure that the development, operation and management of the database complies with the information policy.

C4.6(i) As part of the user registration process, a user should explicitly acknowledge they have read and understood the policy and a method of recording their acceptance of the obligations it details should be available.

C4.6(ii) The information policy should include how a breach in the terms and conditions of use of the database will be dealt with.

[27] http://www.rcpa.edu.au/getattachment/a631a573-0d07-4bd4-ba67-cfe545618dd1/Managing-Privacy-Information-in-Laboratories.aspx

C4.6(iii) A registered user is defined as described in C4.4(v) above.

C4.6(iv) Audits of a database conducted by DBAdmin users should include:

- Review of access logs
- User management: information about the registrations to the database
- Database (usage) statistics: information about the number of contributors and contributions to the database
- Log in events and registration types

C4.6(v) Risk Analysis must be conducted periodically as a component of a systems test after hardware or software modifications or upgrades to identify and remove vulnerability to any threats or weaknesses identified. More information regarding quality systems can be found in the NPAAC document, Requirements for Medical Pathology Services (First Edition 2013)[28]

S4.7 There must be a documented procedure/policy readily available to be followed in the event of a security breach or unauthorized disclosure of information.

C4.7(i) This policy should address the technical, operational, legal, and ethical consequences of such a breach. Such a breach may carry legal or professional obligations to report the breach to clinicians, patients, the host institution, contributing laboratories, and regulatory authorities.

C4.7(ii) The Office of the Australian Information Commissioner has produced a useful guide to data breach notifications.[29]

[28] http://www.health.gov.au/internet/main/publishing.nsf/Content/health-npaac-docs-medpathserv.

[29] http://www.oaic.gov.au/privacy/privacy-resources/privacy-guides/data-breach-notification-a-guide-to-handling-personal-information-security-breaches.

5 Content

The scope of the data to be captured and maintained should be clearly articulated to encourage consistency across databases. It is essential there be adequate information provided to enable an external user to review and use the information with confidence in making clinical decisions.

S5.1 The means of submission of data to databases must include the use of commonly used data exchange formats

C5.1(i) The supported formats should be clearly stated and publicised through appropriate mechanisms such as internally documented procedures or on a database website

C5.1(ii) Submission using standard formats reduces the risk of corruption of data during upload into the database, and facilitates sharing of data, and federating of databases.

S5.2 Each record in the database must include the following data:

- **The variant described using a recognised (and specified) nomenclature that uniquely identifies the variant. This must be referenced to the sequence stipulated by the database or precisely state which reference has been used.**
- **The zygosity state must be provided if known.**
- **For variants described using genomic coordinates, the reference Genome Build must be stated**
- **The methodology of variant detection**
- **The reason for testing must be provided in the context of relevance to the database.**
- **If the submission includes a statement regarding clinical interpretation or significance, the basis of this statement must be provided. This may include reference to other published sources or to unpublished studies by the submitting laboratory.**

C5.2(i) The Primary descriptor of the variant should be provided in HGVS nomenclature with reference to a genome sequence (genomic coordinates). Secondary descriptors (non-standard aliases or reference to a transcript sequence) may be included. Mapping tools and conversions must be identified

C5.2(ii) Established legacy nomenclatures which are difficult to change may be used where conversion to HGVS is not possible, or difficult e.g. HLA haplotypes. The nomenclature being used must be clearly stated.

C5.2(iii) Where possible, information regarding the frequency of the variant in the tested population or a control group should be submitted.

C5.2(iv) Where available additional information should be provided e.g. protein function prediction, splicing abnormality prediction, literature evidence, familial studies on variant co-segregation with disease, or other relevant evidence.

C5.2(v) A reason for testing may be:

- Diagnostic test in affected person
- Predictive testing in an unaffected/affected person at high risk of a specific mutation
- Predisposition testing for a particular disease in an unaffected/affected person who is not at high risk of a specific mutation)
- Segregation analysis to assist with pathogenicity assessment
- Screening test for many disorders in a person who is not at high risk of specific disease.
- Theranostic testing in an affected person to guide therapeutic decisions.

C5.2(vi) Where available, the clinical phenotype and supporting multidisciplinary evidence should be provided. This may include:

- Patient history and diagnosis
- Inheritance information – This should include the number of affected and unaffected individuals tested for the variant – suspected mode of inheritance, consanguinity,
- ethnicity
- gender
- age at diagnosis.
- carrier status
- pathogenicity
- Relevant non-genetic pathology results
- Relevant non-genetic medical results

C5.2(vii) If phenotype data are submitted together with genotype data, the phenotypic information would preferably be reviewed by a relevant multidisciplinary team (MDT) or clinical service specialising in the disease.

S5.3 Standardised terminology and a recognised international ontology must be used within the database. The selected standard should be clearly stated and made available through appropriate media such as internally controlled operating procedures or on the database website.

C5.3(i) If phenotype information is provided in the database, the ontology system being used **must** be stated (e.g. SNOMED CT, Human Phenotype Ontology (HPO), etc.)

S5.4 There must be clearly defined guidelines for the classification of variants. Any pathogenicity classification must provide detailed information describing how and why the classification has been made.

C5.4(i) The criteria for classification of pathogenicity should be evidence based, clearly stated, and available through appropriate media such as internally controlled operating procedures, or on a database website. *NPAAC Requirements for Medical Testing of Human Nucleic Acids*[30] and *Requirements for Medical Pathology Services* [31] *,or local equivalents,* should be referred to regarding reporting decisions and classifications**.**

An example of this is the InSiGHT classification criteria for mismatch repair genes www.**insight-group.org/criteria**

C5.4(ii) If pathogenicity is determined by the submitter, it **must** be stated how the pathogenicity was determined. There may be multiple fields in the database where this information is recorded.

C5.4(iii) The database **must** flag where there are inconsistencies in the database, with a mechanism to resolve the discrepancies (e.g.: same variant submitted twice or more with different pathogenicity classifications). See Section 7 for more detail.

C5.4(iv) A description of "research" undertaken to reach the conclusion including citation of any peer reviewed papers should be included to enable the database user to make an informed professional judgment about the pathogenicity classification with a certain level of confidence.

C5.4(v) The level of confidence in the classification may be included.

C5.4(vi) The classification of pathogenicity should be described within the purpose of the database. This may include the reason for testing, and what pathogenicity means in what context (e.g.: describing the database as an LSDB versus a generic genome database).

S5.5 Each patient, and each family, must have a unique identifier applied.

C5.5(i) The unique identifiers should be system generated.

C5.5(ii) These unique identifiers are required to flag the frequency and co-occurrence of variants. For example, this gives the database the ability to flag to the user that there are multiple variants in different genes in a single individual, or that a single variant is in a number of related (familial) or unrelated individuals.

C5.5(iii) The system generated identifiers also provide a mechanism which allows the submitting laboratory to identify the individual patient data within the database submitted by that laboratory. This is to enable future updating and correction of information which may impact

[30] National Pathology Accreditation Advisory Council Requirements for testing of human nucleic acids (second edition 2013) Commonwealth of Australia

[31] National Pathology Accreditation Advisory Council Requirements for Medical Pathology Practices (first Edition 2013) Commonwealth of Australia

patient management. It is neither necessary nor appropriate that any user other than the submitter be able to make this association. This restriction on identifying a patient applies equally to the curator and custodian as to other registered and unregistered viewers of the database.

C5.5(iv) There **must** be a mechanism by which unique patient and family identifiers can be updated by the curator. This is necessary in the event that a submitter realises that multiple instances of a variant identified in supposedly unrelated people are actually from members of the same family, or that an individual's results have been submitted multiple times to the database as independent events.

C5.5(v) A submitter should have a mechanism to relate the database-generated unique identifier to their own in-house medical records.

S5.6 The accreditation status of the laboratory which performed the analysis must be stated.

C5.6(i) The definition of an accredited laboratory in the context of the database **must** be clearly stated and made available on the database website interface or (in the case of a private or restricted database) in a documented policy.

S5.7 The analytical validity of the report must be clearly indicated, and documented in sufficient detail to enable assessment by a viewer.

C5.7(i) For each report of a variant, **the following information must be included:**

- A measure of the quality of the variant call. In a tightly controlled environment (such as a validated test in a laboratory accredited for clinical diagnostic service delivery), it may be sufficient to have a statement regarding quality which applies to the entire dataset. In a less controlled environment (such as an RUO or translational test in a research or unaccredited translational laboratory), it may be necessary to have a statement regarding quality for each reported instance of a variant.
- The consistency [accuracy of the nomenclature] of the variant events held in the database **must** be demonstrated, and the provenance of the data **must** be defined. This is a joint responsibility of the submitter and the curator.
- Indicate whether orthogonal method verification or previously validated test was performed. This is intended to lend more integrity to the data – if confirmed, or already validated, the user is likely to feel more confident utilising this information clinically than if it has not been confirmed by an alternative method or run as a validated test.

C5.7(ii) For further guidance regarding analytical validity parameters in medical testing, refer to ***The NPAAC reference material,***

***Requirements for the development and use of in-house in vitro diagnostic devices (IVDs)*[32] .** In relation to genomic sequencing, refer to *Assuring the Quality of Next Generation Sequencing in Clinical Laboratory Practice Working Principles and Guidelines*[33] developed by the Next Generation Sequencing Standardisation of Clinical Testing (Nex-StoCT) Workgroup.

S5.8 The clinical validity of the report must be clearly indicated and documented in sufficient detail to enable assessment by a viewer.

C5.8(i) For each report of a variant, **the following information must be included:**

- Any records considered valuable in defining provenance. This may include diagnostic records, peer reviewed papers, research reports, confirmation of variant by other methods
- Clearly indicate whether the consequences were **experimentally determined** or only **theoretically deduced**.
- When changes in patients with a recessive disease are described, it should be clear in which combination (phase) the changes were found

C5.8(ii) Further guidance regarding clinical validity parameters in NGS based genetic tests – can be found at: reference/s for guidance on clinical validity parameters

5.9 It must be stated clearly when a review of data interpretation has taken place.

C5.9(i) There should be a clearly defined policy regarding data re-analysis, and this should be made available through appropriate media such as internally controlled operating procedures or on a database website. Expectations for re-analysis and re-interpretation of data should be managed against laboratory/database resources and priorities.

[32] Requirements for the development and use of in-house in vitro diagnostic devices (IVDs) National Pathology Accreditation Advisory Council (NPAAC) Commonwealth of Australia 2007 edition.

[33] Gargis,A. et.al. Assuring the quality of next-generation sequencing in clinical laboratory practice.; Supplement 1, Nature Biotechnology, Vol30, No.11, November 2012

6 Functionality

A database is only as useful as the information contained within, and the ease of access to the information in a relevant, efficient and informative fashion. Ease of access includes the visual appearance and navigational qualities of a database and website interface, as well as quality and usefulness of interrogation capabilities.

S6.1 The database must have flexible search capabilities, including the ability to search the content of each record over time

C6.1(i) Search capabilities should be customisable to allow for multiple types of queries including orthogonal queries to increase filtering capabilities.

C6.1(ii) Examples of searchable fields include specific variant, gene name, alias (gene, disease), disease type and classification, phenotype, protein, codon, ethnic group, geographic location, author or citation.

C6.1(iii) Probabilistic search capabilities ("fuzzy" searches) may also be desirable

S6.2 The database must be capable of associating variant records from another or multiple databases complying with the requirement to audit such external databases prior to data import per S2.6.

C6.2(i) The database should support the ability to import good data to enhance the usefulness of a database.

S6.3 The database must have the functionality to allow for tracking for regular review and updates, and aggregate information in a version controlled manner.

C6.3(i) It is desirable that the database is able to track updates for entries and retrieve or receive data from participating sources automatically. In the event that this is not possible, the database should be capable of batch uploading from tables or spreadsheets to maintain most current information. This should include key metadata which defines the provenance and flags the status of the variant/s reported; noting any correction of nomenclature errors.

S6.4 The database must be able to account for the number of occurrences of the same variant in the same individual or in the same family.

C6.4(i) The database **should** have a means of flagging variants that have been reported in one individual and/or one family versus many unrelated individuals. This is needed to distinguish between rare / isolated events versus common events, and minimise "double" counting of the same variant event.

S6.5 The database must be able to generate summary reports for viewers

C6.5(i) Summary data of information held in the database should be provided to the users of the database. This may include (but not limited to) the number of:

- Genes described
- Samples entered (single entities)
- Coding mutations
- Papers cited, authors
- Unique Variants
- Fusion genes
- Genomic rearrangements
- Whole Genomes
- Whole Exomes
- Copy Number anomalies
- Mutation maps
- Graphics, tools, location of a variant in a gene

C6.5(ii) Summary reports may be customised according to context of the reports and needs of the end users, and be downloadable[34].

For example: COSMIC Keyword search using the term "Lung" provides a summary report, and selection of the Primary site provides a customisable report including further internal links to more detailed information

S6.6 The database must be able to support the transition of existing data to newer version of the Human Genome reference build as they become available.

C6.6(i) As newer versions of the Human Genome Reference build are released, the existing database content should be re-mapped to the new reference within a reasonable time frame.

C6.6(ii) The old variant description should be listed with the new variant description to allow searching using any version.

C6.6(ii) The conversion should be automated where possible. Automated transition modules are available (e.g.: the batch liftOver tool by UCSC, or Remap by NCBI).The module which is used **must** be accredited or optimised and validated by the laboratory. This is to ensure the robustness of the program and harm minimisation (such as corruption of existing data***).***

C6.6(iii) Any conversion should be clearly indicated to alert users.

S6.7 The database must be able to identify incorrect and inconsistent data entries.

[34] http://chromium.liacs.nl/LOVD2/colon_cancer/variants_statistics.php
http://www.ncbi.nlm.nih.gov/variation/view/
http://www.ncbi.nlm.nih.gov/clinvar/submitters/
http://asia.ensembl.org/Homo_sapiens/Gene/Variation_Gene/Table?db=core;g=ENSG00000134982;r=5:112707498-112846239

C6.7(i) This should be an automated process with notification of incorrect or inconsistent records to the curator for review and correction where required.

C6.7(ii) The database should have a means of flagging duplicate entries of the same variant event.

S6.8 Mechanisms for access and sharing between data repositories must be supported.

C6.8(i) There should be mechanisms for importing data and exporting data for inclusion in an external database in a compatible format.

C6.8(ii) The type of data that can be shared should be clearly stated. Sharing of aggregated data may be a more readily accepted method for sharing data.

C6.8(iii) Mechanisms for supporting useful links such as clinical content / information websites may be included.

C6.9(iv) A citation list provided for individual or aggregated data should be included.

Appendix

1 The Curator

Curation is the activity of managing and promoting the use of data from its point of creation, to ensure it is fit for contemporary purpose, and available for discovery and re-use[35]. It involves the selection, preservation, maintenance, collection, and archiving of data in order to establish, maintain, and add value to repositories for current and future use.

Challenges of DNA data repository curation include:

- The increasing rate of creation of data sets as MPS reduces in cost and increases in output
- Standardising terminology and ontology within a database/set of data/federated databases
- Filtering and triaging variant calls and evaluating level of confidence in accuracy
- Maintaining relevance, and accuracy of data within the database
- Maintaining currency of genome build and compatibility of variants recorded to current / updated genome builds (i.e. correct alignment of sequence to reference)
- Facilitating access and sharing through secure links
- Compatibility of database schema with external and/or federated databases

Good curation means checking provenance of the variants and any associated information: was it from an accredited source, is the evidence robust, and is there any metadata to support the variant identification. What, if any, is the evidence for pathogenicity?

Exemplars of well curated databases include DMuDB, ClinVar, and COSMIC (Catalogue of Somatic Mutations in Cancer)

When a database has been established, the Curator's role includes the undertaking or delegation of the following tasks:

Curation:

- Provide general information about the database contents and functionality
- Evaluate and register new submitters
- Curate new submissions
    - ensure standards that are set for data collection / submission are met
        - formatting of submitted data to be uploaded to the database
        - ensure information within the database is accurate, up to date, and accessible
        - understand the ways in which genetic variant information are presented / stored (e.g.: nomenclature used), and utilise automated tools (such as Mutalyser) to perform variant curation, data formatting, and related tasks

[35] Lord, Philip, and Alison Macdonald. e-Science Curation Report: Data curation for e-Science in the UK: an audit to establish requirements for future curation and provision. Digital Archiving Consultancy Limited, 2003.

            - Ensure the consistency [accuracy of the nomenclature] of the variant events held in the database are demonstrated. If this is not possible, provenance of the data must be defined and ensured
        - Assist any committee which is in place to classify or approve classification of genetic variants prior to inclusion in database (for example: providing supporting data, and arranging virtual / face to face meetings for discussion
    - Check URLs for currency
    - Prevent duplication of entries; and linking of entries within the database or between databases
    - Extract relevant information from published literature (data mining) and other resources. Upload this data as required and or organise in a manner which allows them to be used for the classification of variants
    - ensure mechanisms for separating public and non public data are maintained and comply with local privacy laws
    - review and approve any Locus Reference Genomic (LRG) sequences
- Undertake audits according to predefined schedule
    - Review and action any database content audit outcomes
- Update data entries (enter new publications, conclusions on functional consequences)
- Update variant descriptions as they evolve
- Promote the database
    - Apply the concept of microattribution to acknowledge database submitters for their contributions
    - Collaboration with other groups, national and global, to encourage and facilitate the sharing of data
    - Contact database submitters to inform them of other parties which are interested in their data / similar patient cohorts
    - Analysis of database statistics for presentations, grants, and papers.
- Publish regular updates including
    - Summary reports of database contents
    - Frequency of database updates, and last update,
    - audit outcomes,
    - notifications of corrections and or removal of data
- Ensure contingency plans for emergencies (back-up curators/ administrators, database transfer/archive) are in place

Maintenance (The following may be performed by the curator or delegated to a database administrator; however curator is responsible for overseeing):

- Server administration
- Maintain server and data security (operating system, firewall, back-ups, etc.)
- Assist with technical development of the database
    - Update software platform regularly (bug fixes, latest functionality)
- Review and action any database infrastructure audit outcomes