Response to the FDA Draft Guidance (NGS Testing & Public Variant Databases)

Global Alliance for Genomics and Health
ClinGen Principal Investigators

Submitted: 6 October 2016

Contact:

Adrian Thorogood
Coordinator, Regulatory and Ethics Working Group
adrian.thorogood@genomicsandhealth.org

Outline

Summary of the FDA Draft Guidance on "Use of Public Human Genetic Variant Databases to Support Clinical Validity for NGS-Based In Vitro Diagnostics" (8 July 2016) (here)

The FDA is moving to create a flexible and adaptive approach to regulate next generation sequencing (NGS). (*see* "Use of Standards in FDA Regulatory Oversight of Next Generation Sequencing (NGS)-Based In Vitro Diagnostics (IVDs) Used for Diagnosing Germline Diseases" (here)).

To streamline premarket review of an NGS test, the FDA would like to allow test sponsors to refer to public variant databases as evidence of clinical validity. To this end, the FDA has published **draft recommendations for public variant databases** concerning transparent aggregation, curation, and interpretation of variant data.

Generally categories of recommendations include:

A. Database Procedures and Operations: define, version processes; make processes transparent; preserve data and linkages; procedures for closure; security/privacy compliance; commonly accepted data formats.
B. Data Quality (and Currency): consistent nomenclature; detailed metadata; deduplication.
C. Curation, Variant Interpretation and Assertions: well-defined SOPs carried out by qualified personnel; assertions should be appropriate, not (clinically) misleading, and versioned.
D. Professional Training: training, qualifications, oversight, and conflicts of interest.

The draft also outlines a process for FDA "recognition" of public variant databases, potentially relying on assistance of "third parties" in the recognition process. Recognized databases could be relied on as evidence of clinical validity during review of NGS test submissions.

Organizations

The Global Alliance for Genomics and Health (GA4GH)

The GA4GH is an international coalition of over 400 research institutions, health centers, patient groups, and life science and information technology companies. The goals of GA4GH are to enable effective and responsible sharing of genomic and clinical data to improve human health, and to support projects that demonstrate the value of data sharing. For more information, please see our recent Perspective Article here: (GA4GH, A federated ecosystem for sharing genomic, clinical data, *Science* 352(6291),1278-1280 (2016)). The article advocates for improving the sharing of genomic data so that all individuals share in the benefits of scientific and medical advances, as stipulated by Article 27 of the Universal Declaration of Human Rights.

The GA4GH has a strong interest in facilitating the aggregation, curation, and interpretation of variant data and the responsible sharing of such data through public variant databases. The GA4GH demonstration project, the BRCA Challenge, is an international collaboration among data holders, variant curators, researchers, and clinicians to improve our understanding of the genetic causes of breast and ovarian cancer, and to make this information publicly available and easily accessible. The first work product of the BRCA Challenge is the BRCAExchange.org, a publicly available portal displaying *BRCA1/BRCA2* variant classifications and other variant information. A number of GA4GH members support or administer public variant databases, such as ClinVar (http://www.ncbi.nlm.nih.gov/clinvar/), the Leiden Open Variation Database (LOVD) (http://www.lovd.nl/3.0/home), BRCAShare (http://www.umd.be/BRCA1/), International Society for Gastrointestinal Hereditary Tumors (INSIGHT) (http://insight-group.org/variants/database/), the Harvard Personal Genome Project database (evidence.personalgenomes.org), the Clinical Interpretations of Variants in Cancer database (CIViC) (https://civic.genome.wustl.edu), the National Cancer Institute Genomic Data Commons (NCI GDC) (https://gdc.cancer.gov/), and the Bionimbus Protected Data Cloud (https://bionimbus-pdc.opensciencedatacloud.org/).

The Clinical Genome Resource (ClinGen)

ClinGen is dedicated to building an authoritative central resource that defines the clinical relevance of genes and variants for use in precision medicine and research. Feedback noted in this document is provided on behalf of the principal investigators of ClinGen. It does not necessarily represent the views of NIH staff involved in ClinGen or the institutions funded by ClinGen. For more information about ClinGen, please see our website and marker paper.

We, the executive committee of the GA4GH and on behalf of our member organizations, as well as the principal investigators of ClinGen, strongly support sharing of genomic data and clinical interpretations. We would like to begin by first applauding the efforts of the FDA to develop a flexible and adaptive approach to regulating NGS tests. Our membership expressed general support for use of public variant databases as sources of evidence of clinical validity, and general support for creating a recognition process for variant databases. The FDA has identified the crucial importance of public variant databases in interpreting the clinical significance of genomic variants.

Our response is divided into three parts. In Part I, we provide direct feedback on the FDA's draft guidance for public variant databases. In Part II, we respond to the specific questions for consultation posed by the FDA in the Federal Register. In Part III, we propose how the FDA and GA4GH can work together to coordinate the development of international standards and recognition processes for public variant databases.

I. Feedback on Draft Guidance

Data Sharing Incentives: It would be helpful to clarify if and how the FDA plans to incentivize data sharing by genomic testing laboratories through public variant databases. Is there a scheme to ease regulatory approval burden, e.g., by imposing a lower risk category of tests submitted by sharing laboratories? How would conflicts of interest be handled in this regard?

Distributed Responsibilities for Data Quality and Regulatory Compliance: Many of the overall recommendations (summarized as lines 224-230) cannot always be fulfilled directly by public variant databases, but instead must be distributed to submitters (e.g., data aggregation, curation, and interpretation; transparency about variant interpretation processes; regulatory compliance for ethics approvals, consents, and de-identification). Could further guidance be provided on how databases may share such responsibilities with submitters, e.g., by imposing submission policies or agreements, or requiring submitter interpretations and transparency around the methods and evidence for those interpretations? Could a collaboration among databases and expert curators be "packaged" and recognized together? For example, the *BRCA Exchange* shares variant data curated by an expert group - the *ENIGMA Consortium*. The data reside in different databases, but the Exchange has links to the data sources and reports the expert curation results. Could a joint application be submitted? Likewise, ClinVar supports submissions from many expert groups, and draws on many databases. Could ClinVar be recognized for certain aspects of the guidance and the submitters apply for recognition of their content in ClinVar? The FDA could explicitly point out in lines 124-128 that the database role is often different from the submitter role, and may want to consider separate guidance for the "database-level" and the "submitter-level".

Recognition of a Subset of a Public Variant Database: There are doubts that many databases will be able to meet these standards and achieve recognition at the entire database level. Could a database receive recognition for a subset of variants? This would be especially

important in cases of “crowd sourced” databases where the general public can edit but expert curators can “sign off” on specific versions of a curated variant. For example, the ClinVar database contains a proportion of its variants (3 and 4 star level) that undergo expert review through an approach that is reviewed and approved by ClinGen.

Distinction Between Variant-Level Data and Case-Level Data: Most publically accessible databases cannot share patient data directly. The utility of this draft guidance is best focused on sharing of variant-level data such as interpretations and supporting evidence for variants and not the sharing of genotype and phenotype data directly. As such, we recommend that edits be made to lines 230 to clarify that the database is not required to store case-level sequence or phenotype data from research participants and patients.

Data Uniqueness: It will be difficult if not impossible for variant databases that do not have access to case-level data to ensure individual data points are not duplicated (lines 315-316). We suggest databases provide general guidance to submitters of interpretations to take into account what is known or not known about the uniqueness of case-level evidence (also note that a GA4GH Task Team is tackling this issue: see Part III). Reliable variant databases will need to be able to track back to case-level data, but do not necessarily need to retain such data themselves. The key point is to be able to verify variant interpretation by ensuring access to case-level data, but case-level data in a public database will face serious issues of data security, confidentiality and privacy.

Documentation of Analytic Validity: With regard to Metadata referenced in lines 308-310, can the FDA clarify if the requirement for test name applies only to interpretations supported by case reports, or also to variant interpretations relying on many different tests and methods contributing to variant-level data? This is an area where databases may need to rely on submitters to supply detailed metadata. Would a submitter need to submit metadata on analytic validity for each case or only some form of summary metadata? There may also be privacy issues where reporting the laboratory name may increase the identifiability of the individual. We suggest more general guidance that databases take available information about analytic validity be taken into account during variant interpretation.

Refining Definitions: The definitions provided for assertion, aggregation, curation, and interpretation are helpful and add clarity, as these terms are often ascribed widely different meanings.[1] The guidance may want to note that not all activities will clearly fit in one or the other of these definitions. Wherever possible, the guidance should speak to more specific activities.

[1] FDA definitions: **Assertion:** the informed assessment of a genotype-phenotype correlation (or lack thereof) given the current state of knowledge for a particular variant. An assertion is generally noted in the genetic variant database entry for a particular variant (e.g., benign, drug resistant, etc.). **Aggregation:** the process by which variant data are systematically input into a genetic variant database. This process may require that data conform to specified formats. **Curation:** the process by which data regarding a specific variant are collected from various sources, annotated, and maintained over time. **Interpretation**: the process by which genetic variant database personnel evaluate the evidence regarding a linkage between a genetic variant and a disease or condition and make an assertion about that linkage (or lack thereof).

One thing to keep in mind is that assertions for a given variant may differ depending on context e.g., germline predisposition to disease v.s. response to treatment. The definition of assertion should note that assertions are often context-specific and such context should be indicated along with an assertion.

Specific Data Formats Standards: Regarding the *Data formats* section (lines 286-290), what specific data format standards will apply (e.g., HGVS nomenclature, recommended disease ontologies)? Please see our proposal, in Part III below, to coordinate an existing multi-agency collaboration to standardize the computational representation of sequence variation.

Definition of "Publicly Accessible": The guidelines could benefit from clarity in the definition of "publicly accessible database" and "publicly available documentation of evidence" (lines 124-128), especially relating to the FDA commitment to "public access to data" to support medical treatment decisions (line 184). How much data must be publicly available to qualify? Would databases that charge a fee for access qualify? Would controlled access databases qualify? If so, would the FDA want to specifically address conflicts of interest that can arise in controlled access processes (e.g., where there is a lack of independent access review or a lack of transparent access policies)?

Minor Comment: It is unclear why lines 379-381 ("curation procedures should ensure that all data has been collected in compliance with all applicable requirements" for privacy and research regulation) is included in the section on professional training and conflicts of interest, rather than in the Part A subsection on *Privacy and Security*.

II. Responses to Specific FDA Consultation Questions (here)

1. Should the quality recommendations outlined in the guidance apply equally to databases of somatic variants and to germline variants?

We suggest that this guidance apply to all types of human genetic variation, Many databases are not solely germline or somatic (i.e. ~1.5% of ClinVar is somatic/cancer variant submissions[2]). Some emerging projects will be largely focused on somatic tumor variation. Moreover, this is not always a straightforward distinction. Some germline variants affect drug response. Rare germline variants can be active in cancers; i.e., consider the recent example of intersecting COSMIC recurrent variants with germline variants (PMID: 26689913). The same variant may have different curation outcomes when somatic versus germline. What is important is that the data is the basis for clinical decisions. While the guidelines should in general apply to both germline and somatic variants, additional quality recommendations for somatic variants may be necessary such as accounting for evidence as to whether the variant derives from somatic/tumor only or from the tumor sequence subtracted from a matched normal. But in

[2] Another example of a database containing both germline and somatic variants is the IARC TP53 Database. http://p53.iarc.fr/

principle, all of the key approaches to database maintenance and expert review should apply regardless.

This is also why assertions need to be context-specific. In all instances classification of variants, germline or somatic, should be considered in the context of their proposed use/s for a given gene e.g., prediction of disease risk, prognosis, or response to therapy. Some variants in BRCA1, for example, appear to predict response to PARP inhibitors, but may not imply hereditary cancer risk. The “recognition” process too will need to address the different uses of variant interpretation.

2. While this document applies to NGS-based tests, FDA expects that it may also be relevant to genetic tests that use other technologies (e.g., polymerase chain reaction, micro-array technology, Sanger sequencing, etc.). Are any additional considerations necessary to support the use of these databases in the premarket review of tests using technologies other than NGS, should FDA decide to apply this approach more broadly in the future?

We believe this guidance should apply to all variation, regardless of the technology from which the primary evidence may have been generated. The details of quality control and analytical validity will of course depend on the underlying technology, but the need for verification from accurate measurement of the analyte to interpretation of clinical significance should apply to all technologies. Given that the focus of the guidance is on variant-level data, it would be hard to separate out the different technologies that are used in aggregate to build evidence on variants, though general guidance should be provided to ensure the underlying technology considered during interpretation and is recorded. It should be noted that variant interpretation typically assumes the range of all variation is known. Some technologies may never be able to capture certain types of variants. An assay-independent approach might also extend beyond solely in-silico discovery to include meta-analyses that use multiple platforms. Both have contributed to discovery of gene-disease associations (e.g., integrating multiple platforms, such as expression array with sequencing, across separate studies to identify a functional variant).

3. FDA recognizes that the evidence linking specific variants to diseases or conditions will change over time, and as such, assertions about those variants may also change. If an assertion regarding a variant changes over time, how should FDA assess what regulatory actions may be appropriate with respect to IVDs supported by such assertions? How often should FDA conduct ongoing review of an FDA-recognized database?

Test developers should be required to document how their test will stay up-to-date with evolving knowledge. Databases should at a minimum document the date the variant was last evaluated, and for what purpose e.g., disease risk, prognosis, response to therapy. Criteria for recognition are likely to be *process* criteria, about practices for reclassification, verification, and communication of changing interpretation. Additionally, databases should have processes in

place to track changes to variant assertions aggregated from elsewhere. Oversight should consider the frequency and extent (or clinical relevance) of changes to variant assertions over time. Variant assertions may change without substantially changing clinical actions or utility (e.g., variant moves from pathogenic to likely pathogenic, from benign to variant of unknown significance). If not elsewhere defined, the FDA needs to define what assertion changes would be substantial enough to require downstream action. The FDA should specifically consider how databases should highlight and communicate clinically important changes in assertion, such as a downgrade from pathogenic to uncertain.

It seems reasonable that database recognition be updated annually. However, this should not require that every variant be re-evaluated annually as many clinically significant variants for rare diseases will only ever be observed in one patient. FDA recognition should focus on the processes of updating variant interpretation, as opposed to requiring the database to review every variant annually. Databases and submitters are unlikely to have resources to review variants that are never observed subsequently. The review frequency for FDA-recognized databases may need to be modulated according to the variant content of the database.

4. FDA notes that databases may have "discordant calls" with other databases, where the assertions for a variant in each database vary. While FDA believes that these discordant calls often arise because one database has information the other does not and our proposed policy will mitigate these issues over time; what, if any, action should FDA take when it learns about discordant calls between two databases with respect to database recognition or IVDs supported by such calls in FDA-recognized databases?

The proposed policy does not address how the FDA, test developers, or databases would be notified of external discordance or how it would be resolved. To create feasible solutions, it is important to first understand the levels of internal (within databases that use a multi-submitter format) and external (between-databases) discordance. Internal discordance is addressed through tracking and flagging, and development of internal resolution plans. Amendola et al (PMID: 27181684) discuss discordance in variant interpretations between and within laboratories when tasked with using internal interpretation schema v.s. the ACMG guidelines. Through a moderated, discussion-based format, a substantial number of discordant variant interpretations were resolved. However, it is important to note that discordant variant interpretations remained even after using the same interpretive guidelines and participating in a moderated discussion of shared evidence to promote consensus. Practical analysis such as Amendola et al about internal discordance can inform the scope of potential external discordance and the reasoning that underlies a resolution or a sustained discordant interpretation. Based on this, the FDA should allow for databases to consider measures to handle a sustained external interpretive discordance (i.e., not resolving discordance despite best efforts to ensure databases have access to similar information.), consistent with the practice of medicine in all disciplines. Key to handling discordance are transparent (though not necessarily identical) interpretation SOPs, and links to underlying evidence.

5. FDA has requested information regarding conflicts of interest for curators and personnel of databases seeking FDA recognition. FDA acknowledges that many personnel involved with variant curation and interpretation may have some connection to NGS test developers. What type of information should FDA collect and what policies should it implement to mitigate such potential conflicts of interest in FDA-recognized databases?

We suggest that the FDA request documentation from the owners, curators and submitters of a database and/or dataset that requests FDA recognition as to whether any of those parties have a financial conflict of interest with respect to genetic service provision.

III. Future Coordination between the GA4GH and the FDA

International Standards and Recognition Process: We support the FDA's outline of a recognition process for public variant databases. As public variant databases often collect, store and share data internationally, we recommend that the FDA work to establish an international recognition process for public variant databases. As other national regulators impose approval processes, public variant databases may have to meet multiple, potentially divergent standards and recognition processes. It would be preferable if standards and recognition processes for public variant databases were developed through international collaboration from the beginning. As an example, there are already variant classification schemes (InSiGHT, ENIGMA) other than the ACMG's that have been developed with international consultation, and are internationally recognised. These examples show that international approaches can work and allow data to be combined across multiple countries. In its role as an international convenor, the GA4GH could coordinate with the FDA to ensure international alignment.

Data format and representation: Historically, the genetics community has not distinguished the format used for *presentation* of sequence variation to humans (typically based on the "HGVS guidelines", now known as the "Varnomen recommendations"[3]) and the *representation* of sequence variation within software and among computer systems. A common practice is to store HGVS-formatted variants as text strings within databases, which is inefficient computationally and limits the capabilities of the information model. The GA4GH, in partnership with representatives of ClinGen, ClinVar, FHIR Genomics, Sequence Ontology, and the Human Phenotype Ontology, is actively engaged in a project to harmonize the computational representation of sequence variation data among these efforts. We propose that the Variation Modelling Collaboration work with the FDA to set standards for the computational representation of variation that is used to exchange data among recognized variant databases.

Data Uniqueness: The FDA recommends methods to avoid duplication of individual records in datasets. We would like to bring to your attention current efforts of a joint GA4GH Task Team on Participant Unique Identifiers, in collaboration with the International Rare Diseases Research

[3] Varnomen: Variation Nomenclature recommendations

Consortium, which aims to develop a guiding policy for the generation of participant-specific identifiers (pseudonyms) that enable data from the same individual to be connected across multiple projects without directly revealing the participant's identity.
http://www.irdirc.org/activities/current-activities/participant-unique-identifiers/