The ClinGen Individual Level Database (ILDB)

Vision statement, description and scope of project

Draft 2
August 3rd, 2015

Contributions by Snehit Prabhu, Heidi Rehm, Sharon Plon and Carlos Bustamante
On behalf of
The ClinGen ILDB workgroup

Aims and Objectives

The ClinGen ILDB project is an umbrella initiative to help patients, laboratories and healthcare entities to contribute clinical and genomic data for disease-related research. This will be done through the development of consent standards, data representation models, secure storage repositories, data submission and QC protocols, quantitative and qualitative privacy guarantee mechanisms, data usage guidelines, and lastly, regulatory standards and oversight measures for users of this data by the ILDB workgroup. The overarching aim of the ILDB workgroup will be to facilitate the transparent and seamless use of (and conversely, explicitly define the conditions for non-use of) clinical genomic, phenomic and health data – acquired in the context of delivering patient care – in medical research.

Genomic information generated in clinical settings today is often locked away in data silos at their source institutions. Despite its immense scientific value and potential, such data is unable to contribute meaningfully towards patient-health oriented research (often contrary to the intentions of the patients from whose samples the data were derived in the first place). A variety of causes lie at the root of this status quo, including (a) legal hurdles that discourage organizations from contributing sensitive data, (b) regulatory hurdles that dis-incentivize their participation, and (c) technical hurdles pertaining to the ease with which such data might be securely deposited at a single site.

The existing state of affairs has given rise to several organic, inefficient and potentially error-prone mechanisms through which the medical professionals, each with access to a single institutional data silo, share their limited insights pertaining to the molecular mechanisms of disease. For the task of clinical interpretation, an illustration of the predominant workflow in use today for pathogenicity assessment of genomic variation is provided here:
https://github.com/snehitp/ILDB/blob/e614badc96b8d8fe44b89de6eddb405199d651e9/Documents/Variant-Interpretation-Today.pdf

In contrast, we envision a new workflow by which patient data relevant to the task of interpretation can easily and efficiently be shared among the research community. An illustration of this workflow is provided here:
https://github.com/snehitp/ILDB/blob/e614badc96b8d8fe44b89de6eddb405199d651e9/Documents/Variant-Interpretation-ClinGen.pdf

The ILDB is thus a critical component of the ClinGen variant interpretation efforts. By making large sets of clinical data available to the community through a safe, secure and robust framework, it will enable each ClinGen domain workgroup to process the deluge of clinically relevant variation in a more scalable manner. The quality of interpretations will also directly benefit from the increased evidence made available by the ILDB, resulting in more biological insight and greater clinical actionability of genomic tests.

Innovations required for the use of clinical data in genomic research.

In order for hitherto unused clinical genomic data from patient records to be leveraged, the ILDB workgroup will need to provide solutions to the following list of key issues:

1. **Consent standards**
   [Sub-group: Snehit Prabhu. Kelly Ormond? Heidi Rehm? Sharon Plon?]

   [Snehit: This is not my area of expertise so please add content if you are qualified. Many aspects relating to meaningful consent, patient education and engagement, and legal aspects like HIPAA protections need to be addressed systematically before the ILDB can go online].

   a. For legacy samples, is re-contact with patient required before use in ClinGen ILDB? If not, as a guideline to participating labs, what type of language, or specific clauses, constitute adequate consent?

   b. Going forward, what type of consent will participating labs need to be request from their patients?

   c. What privacy-related risks should be communicated to the patient?

   d. What potential research-related benefits should be communicated to the patient?

   e. Should consent requested be dichotomous (yes/no for blanket use of patient's data), or more nuanced and multi-category? If so, how will data with partial or conditional consent be stored and used in the ILDB?

   f. Other …

2. Data models and IT infrastructure

[Sub-group: Snehit Prabhu. Larry Babb? Sandy Aronson?]

[Snehit: For ILDB version 1, a lot of this will just be a description of the relevant pieces of GeneInsight's software architecture.]

a. Which fields of information and data types from patient records are to be captured? Which fields will be mandatory vs. optional?

b. What controlled vocabularies will be used for each field of entry?

c. What will the database schema (relational or document model) be?

d. What web-based interfaces will be made available to (i) data uploaders, (ii) data administrators, and (iii) data users?

e. How will we authenticate each type of user?

f. What type of IT infrastructure will be used, and how will it be accessed? What IT security provisions will be made?

g. What redundancy and data-backup provisions will be made in case of system failure/crash?

h. What software production and maintenance protocols will be in used?

i. Other …

3. Data submission and QC protocols

[Sub-group: Snehit Prabhu. Sam Baxter? Samuel Aronson? Larry Babb?]

[Snehit: From my discussions with Sandy Aronson and Sam Baxter, it appears that we will require an admin team for day-to-day logistics. Besides that, I anticipate that we will also need an executive committee to make higher-level decisions, around topics like those listed below]

a. For large bulk uploads from participating laboratories, their data + consent types will need to be reviewed by an ILDB committee before incorporation of such data into the ILDB.
    i. Who will be in the committee (e.g. patient advocacy, legal, technology experts, etc.)?
    ii. What will be its guiding principles?
    iii. What will the guidelines for providing/denying such approval be?

b. Alternately, are we either going to
    i. Externalize responsibility to the data-uploader through a legal disclaimer, or
    ii. Programmatically enforce barriers against uploading sensitive data fields by mistake?

c. What is the list of QC measures to be put in place for maintaining data integrity and consistency across uploads from multiple labs?

d. Who will resolve any QC related conflicts that arise during upload and what procedure will be followed to resolve them (versioning protocol, etc.)?

e. Other …

4. Patient privacy and security frameworks

[Sub-group: Snehit Prabhu. Carlos Bustamante? Sam Baxter?]

[Snehit: This section partially deals with pragmatic/logistical issues relating to exposing clinical data fields, as well as an open-ended research component to develop privacy enhancing mechanisms for use of sensitive data in research settings]

a. What are the different risks associated with sharing patient health data? List, categorize, describe and define these.

b. Which data fields from their patient records will labs upload to the ILDB?

c. What privacy framework will be used to judge the risk associated with sharing a particular field?

d. Where will the committee's judgment about each field be recorded? How will it be justified and communicated?

e. If fields are judged to have differing levels of privacy related risks, how will this determine their usage in various types of research?

f. What kinds of read/write/download privileges will be assigned to different categories of users by the ILDB?

g. What (if any) are the quantitative privacy/utility trade-offs associated with each data field in the ILDB?

h. Other …

5. Data usage guidelines and oversight/regulation of users
[Sub-group: Snehit Prabhu. Sharon Plon? Heidi Rehm?]

[Snehit: This section deals with the criteria for and logistics of approving a data-access request. A decision has been made to provide controlled access to known entities only. There will be NO anonymous/public access for now.]

a. What kinds of entities may request access to and be granted access to the ILDB?

b. What does a *bona fide* data-access request look like? What materials must the access-requesting entity provide (web-form/write-up) to establish validity of research agenda?

c. Who will adjudicate over each request (ILDB committee, patient advocacy group, data contributor, etc.)

d. What criteria of the proposed research agenda will be used to establish whether a request for access warrants approval or denial (e.g. requester has to provide power studies to demonstrate the utility of additional case records, etc.)?

e. Once access is granted to a subset of data (e.g. Noonan's syndrome cases), how will we limit the set of queries available to the user?

f. What mechanisms will be used to enforce the terms of data usage?

g. What mechanisms for attribution and reporting (to the ILDB and the data sources) will be established?

h. Other …

Concluding Remarks

The proposed framework has the potential to directly and positively impact patient care, in exchange for minimal to no additional losses to patient privacy than those afforded by current practice. Through research and development of the mechanisms outlined above, the ILDB workgroup will provide a comprehensive solution through which appropriate individual-level clinical data might be incorporated into research studies.

The immediate focus of the pilot ILDB will be to support various ClinGen disease-domain working groups in their respective variant interpretation and curation efforts. In this regard, the ILDB will provide a superior alternative to *ad-hoc* solutions that are in use today, whereby collaborative research around genomic interpretation is mostly done using simple worksheets. In current practice, investigators often manually "scrub" case-level clinical data using rough intuition-based guidelines to protect their patients' identities. Such data is then shared between peers, often via email, in the interest of discovering relevant patient commonalities or to bolster evidence for/against a VUS's pathogenicity. The ILDB will provide a more unified, robust and safe solution to replace these practices. Besides variant/gene interpretation, extensions to other areas of medical and genomic research may also be charted out, but are of secondary importance.