Making Genetic Data Public: Privacy Considerations

Final Project: DATASCI 231- Summer 2023

Megan Martin (mmartin131@berkeley.edu)
Sanjiv M. Narayan (sanarayan@berkeley.edu)

Abstract

ClinVar, which is maintained by the US National Center for Biotechnology Information at the National Institutes of Health, has been a critical tool in determining the role of genetic variation in human disease for nearly a decade. ClinVar is heavily utilized in the clinical genetic testing and research industries, and relies on voluntary submissions of data on genetic profiles ('variants') into a curated database. Submissions include varying degrees of information about the variant including the genomic location and biochemical mutation, and the expected clinical significance - whether a variant is associated with disease or not. The database also includes information about each patient such as demographics and other co-existing diseases. Of note, access to information contained within the database is public, and made available through either the ClinVar web application (https://www.ncbi.nlm.nih.gov/clinvar/) or through direct download of the database (XML, VCF, or TSV format). The availability of such data has caused controversy both for the general public (1) and in scientific literature (2). This paper provides an overview of ClinVar and the types of data submitted, sources of privacy risks for patients, and a privacy analysis utilizing several widely-used privacy frameworks. Finally, recommendations are provided regarding additional consent, notices, and privacy security practices that may be considered by both ClinVar, submitters and users of the database.

Positionality and reflexivity statement

Megan Martin works at a publicly traded genetic testing company that utilizes ClinVar for clinical genetic testing services. Her background is in genetic counseling and has provided patient guidance for the integration of rare genetic variants into clinical care, guidance on genetic testing technology development, operational optimization, and compliance oversight. Additionally, she has been involved in various research initiatives for rare genetic disease characterization and identifying patient perspectives. As such, she has been both a submitter and user of ClinVar during clinical and research efforts. She recognizes that she receives compensation (in the form of salary and stock) in her current role which heavily utilizes and relies on databases such as ClinVar being publicly available.

Sanjiv Narayan is a professor at a private university who uses data science to integrate laboratory, physiological and genetic data to improve patient therapy. His background is focused on research to improve outcomes, and he is deeply committed to ensuring patient privacy and avoiding privacy breaches. He acknowledges that he receives compensation for his clinical and research efforts, and that he is less exposed to patient advocacy efforts for privacy and data breaches. In this team project, he strives to balance the limitations of his background.

Background

Untangling the human genome to identify disease-causing genetic variations is a daunting task that relies heavily on publicly available data in the form of peer-reviewed

publications and databases. As more people are tested and increasing numbers of genetic variants are curated, the importance of fast and easily searchable databases has become increasingly critical in reducing the time to interpret clinical genetic tests and cost reduction. One of the most commonly utilized databases globally is ClinVar, which is maintained by the US National Center for Biotechnology Information at the National Institutes of Health(3). Since 2013, ClinVar has provided a freely accessible public archive of curated human genetic variations and their associated evidence for disease causality (pathogenicity).

The ClinVar website, https://www.ncbi.nlm.nih.gov/clinvar/, contains multiple links about the stated mission, the data upload procedure, and intended use of ClinVar (Figure 1). A disclaimer at the bottom of the web page states that the information contained within the website is not intended for direct diagnostic use and health decisions should not be made based on information contained within the website alone. General information is provided to submitters regarding assumed consenting and de-identification procedures. Submission is voluntary, but must contain certain data fields in order to be published in the database. Users of the database can access information about genetic variations through a search tool or bulk data files are available for download and incorporation into bioinformatics or interpretation pipelines.

Figure 1. ClinVar Public Website, with key links for (a) How to; (b) Data inquiry submission, review and analyses; (c) Selected secondary resources; (d) Data Community; (e) Disclaimer

Throughout this paper, we will further detail the data submission and de-identification process, and discuss data privacy and security concerns for patients whose data has been submitted to ClinVar. We will provide a privacy analysis utilizing various frameworks, and provide an ethical and legal discussion. Finally, we will provide recommendations regarding additional consent, notices, and privacy security practices that may be considered by both ClinVar, submitters, and users of the database.

Introduction of potential data privacy and security risks

ClinVar does not provide consent procedures and instead, simply states that consent and de-identification should be performed by the submitter. ClinVar thus assumes the role of 'data broker' and attempts to take a position of impartiality. We feel that such processes are inadequate for the scope of ClinVar - and could be described as lack of accountability - because it involves the public disclosure of highly sensitive genetic data. We will elaborate on this in our Ethics discussion below.

ClinVar was designed with a view to maintaining data privacy and security, but we feel that critical risks remain. HIPAA, the Health Insurance Portability and Accountability Act of 1996, has title II rules that were extended in 2013 to electronic data entities to regulate security (use and disclosure) of personal health information (PHI)(4). Under HIPAA, PHI specifically includes biometric data such as genetic information such as that identified via clinical genetic testing.

A key unknown is whether HIPAA applies to ClinVar. HIPAA covers entities and not data. HIPAA permits data distribution on a "need to know" basis with health-care providers, health plans, health-care clearinghouses and electronic entities, and business associates. Our interpretation is that HIPAA applies to ClinVar as a data clearing house, althoughClinVar 'trades data online' and so HIPAA may not apply. Documents on ClinVar are not fully clear on this issue (5), and may remain open until determined by litigation.

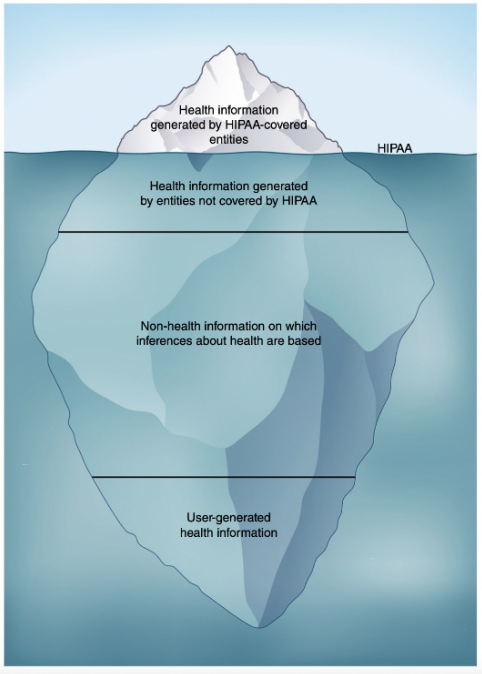

Figure 2. Data covered and not covered by HIPAA

Data from entities covered by HIPAA are regulated as shown in figure 2, unlike the bulk of health data including social media posts, data from wearables and 'traded online'. Data from non-covered entities could be used for aggregation, secondary uses and re-identification with little protection to the individual. Note that the ethical risks of data breach are similar whether or not HIPAA applies.

If HIPAA applies to ClinVar, data distribution for therapy is likely appropriate. However, data disclosure for non-treatment uses is less clear. HIPAA lists provisions to disclose PHI without authorization, and ClinVar may fall under the 4th provision (of 5): public health activities. This is one mission of ClinVar, the public good of disseminating data to other sufferers, and to facilitate treatment. The other 4 provisions for PHI disclosure without consent likely do not apply to ClinVar: for treatment or payment (only substance-abuse and written psychotherapy notes are prohibited), if the patient can agree to or object to any disclosure, during a natural disaster, and if requested by court orders(4)

Other laws may impact data security requirements by ClinVar. The Health Information Technology for Economic and Clinical Health (HITECH) Act requires that covered entities and business associates notify affected individuals, the Secretary of Health and Human Services, and in some cases the media of breaches of unsecured PHI. Unsecured PHI is defined as ''PHI that is not secure through the use of technology or methodology specified'' by guidance of the law. (6)

The role of ClinVar as a data broker introduces its own limitations. In particular, ClinVar has not had to register with relevant registries (such as the California Data Broker Registry) because of the complexity of consumer privacy laws vis-a-vis HIPAA. A more concerning issue is that ClinVar has no geographical boundaries, and data to ClinVar has been submitted

and utilized globally even though the site is U.S. based. Of course, nations have widely differing laws about data security, privacy, respect for PHI and notice and consent. A U.S. consumer could thus have their data used in Asia, Africa, or Europe, which may violate their contextual norms and expectations. It is difficult to see how ClinVar could satisfy GDPR, US consumer laws, and laws in other jurisdictions with its current model.

Our conclusion is that ClinVar's current data security leads to unsecured PHI. While presumably de-identified, the data are made publicly available, with no measures to prevent aggregation with secondary datasets. Secondary datasets include other genetic databases listed on the ClinVar website, which may enable some extent of patient re-identification as discussed below, and facilitate a plethora of unintended uses. If ClinVar data were aggregated with data from non-HIPAA covered 'electronic repositories', the resulting aggregated data could potentially have an opportunity to aggregate, share, re-identify or commercialize data without consent or even knowledge from the primary data sources. Some of these risks have been discussed by others (7).

Data management and submitter responsibilities

ClinVar relies on volunteer submissions from a variety of backgrounds including clinical genetic testing laboratories both in the United States and internationally and research centers. More patient-centric submitters are also accepted including patient registries and medical geneticist clinics. A list of submitters, with the top 5 submitters being clinical genetic testing companies, is here: https://www.ncbi.nlm.nih.gov/clinvar/docs/submitter_list/ and has a combined submission of over 25 million variants.

The notices provided to submitters states that ClinVar assumes that the submitter has obtained appropriate consent from the patients whose genetic variation data is being submitted. The ClinVar policy states: "It is your responsibility to ensure that the submitted information does not compromise participant privacy and is in accord with the original consent in addition to all applicable federal, Tribal, state, and local laws, regulations, statutes, guidance, and institutional policies"(8). Beyond this statement, the submission process does include a quality assessment that if personally identifiable information (PII) is discovered during submission, the submitter will be contacted and the submission halted. However, it does not state the protocol for how PII is detected during the submission process. If PII is discovered after the variant is published in the database, ClinVar will contact the submitter to request an update and if the submitter is unresponsive (no timeline for waiting for a response is provided), then ClinVar will delete the record. While the record will be deleted in the current and subsequent published versions of the database, there is no statement provided by ClinVar as to how previously downloaded (static) versions of the database would be corrected or if users of previous downloads would be notified for correction. It does not appear that there's a clear option for data removal if a patient identifies their information, although a contact email is provided for all general inquiries about the ClinVar database. As such, it appears that the responsibility for ensuring proper de-identification and removal of variant submissions is primarily the responsibility of the submitting party. Finally, the genetic variant data submitted to ClinVar appears to be stored indefinitely as there is no statement provided as to length of storage.

Given that obtaining proper consent is the responsibility of the submitter, we reviewed common consent language provided by the top three submitters. For equivalent comparison purposes, the consent document for clinical exome genetic testing was reviewed for each of the top three submitters and retrieved from the respective websites on July 18th, 2023.

- The top contributor, Invitae (https://www.invitae.com/) consent document states that de-identified data and samples could be shared with third parties for research purposes. There is the ability to opt-out by the patient through changing their preferences in the patient portal. Interestingly, the consent document states: “Recipients of the de-identified data and samples are prohibited from attempting to re-identify me. Recipients may link de-identified data from Invitae with other data sources to create a combined data set as long as the data remains de-identified.”(9) However, there are no statements as to how re-identification is prevented by Invitae.
- The second highest contributor, Ambry Genetics (https://www.ambrygen.com/) appears to be lacking any relevant consent language in either their exome test requisition or their exome patient consent form. A search of the Ambry website did not return any recent information about consenting practices for database submission or sharing. The website’s (https://www.ambrygen.com/legal/notice-of-privacy-practices) privacy practices only state that disclosure of PHI for research purposes is done if allowable by HIPAA. Based on this search effort via publicly available documents by Ambry, it is not clear that patients are informed about their genetic variant database submission practices. Additionally, it is unclear how patients would opt-out of such activities.
- The third highest contributor, GeneDx (https://www.genedx.com/) exome consent document has a clear Database Participation section which details how sharing of de-identified information is handled. The consent document also describes how a unique code for de-identification is used and that PII is removed prior to submission or sharing. It describes that there is a risk to be identified and that risk is higher if the individual has already shared their genetic or health information with public resources such as genealogy websites. This appears to be a more thorough explanation of database sharing practices and the inclusion of risks and examples of how the risk may be increased is commendable.

While this analysis of the top three submitters may not represent the full scope and type of consent documents made available to patients, it does demonstrate a spectrum of the type of information and options provided. As demonstrated by the GeneDx consent, some submitters may provide a more comprehensive explanation of the database sharing practices, but even with this more thorough consent document, it is still unclear if the patient has the ability to opt out of such practices or how such a request to opt out could be made. Generally, patients are used to signing HIPAA-related consents when undergoing healthcare procedures, but those consents often cover PHI sharing practices as it relates to HIPAA covered entities, which as stated in the above introduction is unclear if ClinVar meets the HIPAA entity definition. Furthermore, the HIPAA Privacy Rule establishes conditions under which PHI may be used for research purposes, of which either consent or authorization is required, or waiver of consent must be established and documented by an institutional review board (IRB)(10). Based on the available information collected for these example submitters, it appears that there may be

significant gaps in proper informed consent practices across institutions. From a societal expectation, historically, awareness of research activities at the time of a cancer diagnosis has been low with 85% of patients being unaware or unsure that participation in clinical trial was an option(11). While the findings from this survey may be outdated, it still emphasizes that the general public may be unaware that such research efforts are being conducted at clinical genetic testing institutions, thus highlighting the importance of informed consent procedures.

As genetic variant information identified through clinical genetic testing is likely classified as PHI under HIPAA, a review of additional data fields included with the variant submission is warranted to determine the potential extent of PHI exposure. The data preparation and submission process is complex with required and optional fields available to the submitter. The data dictionary (https://www.ncbi.nlm.nih.gov/projects/clinvar/ClinVarDataDictionary.pdf) states which fields are required by all submitters in order for a variant submission to be accepted vs. optional, but recommended fields. In total, there are 84 data fields for the variant worksheet and 38 additional fields in the clinical information worksheet that would be joined with the variant data. Table 1 provides select fields relevant to privacy concerns for the purposes of this paper.

Field	Optional or Required	Description
Local ID	Optional	Highly recommended, stable ID provided by Institution and should not contain PHI
Gene Symbol	Required	Gene involved unless variant is a deletion or duplication involving multiple genes
HGVS	Required	The c. or g. Nucleotide expression which identifies the specific variant. Can include multiple variants if compound heterozygous
Preferred Condition Name	Required	If gene/variant is associated with a known genetic condition, the name of the condition should be provided
Clinical Significance	Required	Drop down menu including terms like pathogenic, uncertain significance, benign, etc.
Date Last Evaluated	Required	Date the variant was interpreted. Use format yyyy-mm-dd
Allele Origin	Required	Indicate if inherited from a parent (paternal vs. maternal) or de novo, unknown, etc.
Clinical Features	Optional	A list of clinical features evaluated in the patient

Date Phenotype Evaluated	Optional	Date that the clinical features were evaluated in the patient. Use format yyyy-mm-dd
Sex	Optional	Allowed fields are female or male
Age Range	Optional	Range of age at time of testing for the patient. For Adults, report age in years; for under 1 year, report specific months; for over 85, use >85 and not specific years of age.
Population Group/Ethnicity	Optional	e.g. African, Hispanic, etc.
Geographic Origin	Optional	Can be used to indicate country or region. e.g Brazil, Asia, Western Europe.
Indication	Optional	Clinical indication that prompted the genetic testing in the individual.
Test Name	Optional	Name or type of test used to identify the genetic variant. E.g exome sequencing, genome sequencing, microarray.
Family ID	Optional	ID used to indicate that multiple cases in a submission are observed in a single family.

Table 1: Selected Fields, requirements, and description that are submitted to ClinVar.

As displayed in Table 1, even if only the required fields were provided for a genetic variant, a high level of information could be identifiable. For example, there are numerous examples of family-specific genetic variants that are unique to an individual or family(12). If this unique genetic variant was known, by utilizing the search functionality in ClinVar, a user could discover when the individual was tested (by using the “date last evaluated” as a proxy) and the genetic testing company would also be known as this is publicized as the submitting institution for the variant. The risk for discovering sensitive information expands further if the optional fields are submitted as one could discover what clinical symptoms the individual had, the date of the medical appointment where those symptoms were described, ethnicity and possibly even further identifiable regional identity, age at testing, and whether that same genetic variant was found in parents or other family members.

While the combination of required and optional fields may raise concerns about exposing sensitive patient information, this is balanced by the usefulness of the information in expanding medical and scientific knowledge. For example, it is helpful to understand the specific age that a person was observed with symptoms as this helps clinicians provide anticipatory guidance for

onset and treatment. Gender can also be informative as different genetic conditions may present with different symptoms in males or females, although it is important to note that the sex submission field being binary has its own inherent issues and limitations. One of the most important fields is the clinical indication or clinical features as clinical correlation is a critical tool that researchers utilize to understand whether a gene or variant is associated with a common set of clinical symptoms that would meet the threshold of being identified as a new genetic condition or disease. As described in the next section for the intended use, unfortunately, advancing clinical and scientific knowledge relies heavily on published details. The more information known about a rare genetic variant, the more helpful for developing treatments and clinical management.

Intended data use

ClinVar has several intended uses as a repository for genetic variants and how they relate to potential disease. Figure 2 summarizes some links provided by Clinvar for data analysis and review. First, it has become a dynamic catalog of variants which can be used to improve and accelerate diagnosis. Second, it can be used to put clinical variants into context to better interpret genetic testing for patients and caregivers. Third, ClinVar facilitates 'big data' analysis of curated genetic data for research and hypothesis testing. An ultimate goal of these threads is to ultimately enable personalized therapy for patients with genetic diseases. A far-future goal would be to identify non-disease causing variants which could still negatively impact lifespan or 'overall healthiness'.

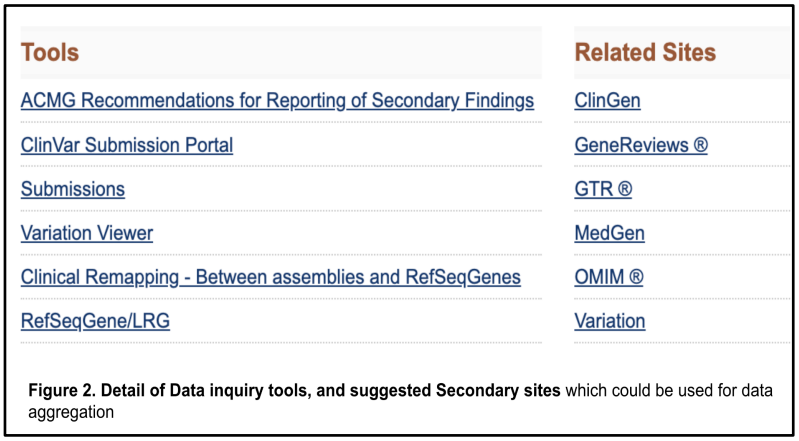

Figure 2. Detail of Data inquiry tools, and suggested Secondary sites which could be used for data aggregation

For data privacy and security purposes, a more important question would be "who uses ClinVar data?" The key actors would be patients and healthcare workers. Important additional users would be researchers at corporations focused on developing new curative therapy. A less appreciated actor could be corporations who sell curated genetic data to other corporations for profit, either as a primary business model or to subsidize diagnostic testing as is done by 23andMe (13).

We would like to emphasize that ClinVar does take its data security seriously, even if we conclude that its protections are suboptimal. In particular, violation of its terms of service falls under Title 18 of the U.S. code (Crimes and Criminal procedure). It is not clear, however, that any suits have been filed for violations of HIPAA by passing PHI to non-HIPAA bound entities, who may then sell the data for profit. We believe that this is a likely battle ground for future cases. To illustrate some risks, we set out to study if re-identification is possible using ClinVar data.

Re-identification example

We set out to study the possibility of re-identification utilizing information submitted to ClinVar. Importantly, in the era of social media, blog posts, and online news stories, individuals with rare genetic conditions may be re-identified by online search tools at scale. A particularly concerning application of this would be identification of a minor whose information is submitted to ClinVar. Unlike consumer tools, medical and genetic data on ClinVar can be posted on minors below age 13(14). For demonstration purposes, we selected a rare genetic condition called “GET4-related condition” and searched for this disease in ClinVar. This resulted in two submitted genetic variants. One variant entry contained extensive data about the patient’s symptoms, age range, date of interpretation (which could be used as proxy for testing date) and the institution which submitted the genetic variant (the NIH Undiagnosed Disease Network). A simple search was entered into google “GET4 rare genetic disease” and two news articles were returned. Given the rarity of this genetic condition, it is not surprising - yet also alarming - that both news articles identified an 11 year old boy diagnosed with this condition in early 2020 after consultation with the NIH (see Figure 3 for ClinVar entry and headline for the news story). Given the rarity of this condition, it is highly probable that this patient is the same individual in ClinVar. Furthermore, from the news article, the patient’s name is identified as well as the parent’s names. An interesting aspect of this case is that the genetic variant is described in ClinVar as being inherited from the father. Given that the father is identified from this news article, one could easily conceive of scenarios where this information could be used to alter (or deny) insurance coverage, lending, or employment decisions. This could result in extensive harm for a family who is likely already in a delicate situation given the care required for a child with special healthcare needs.

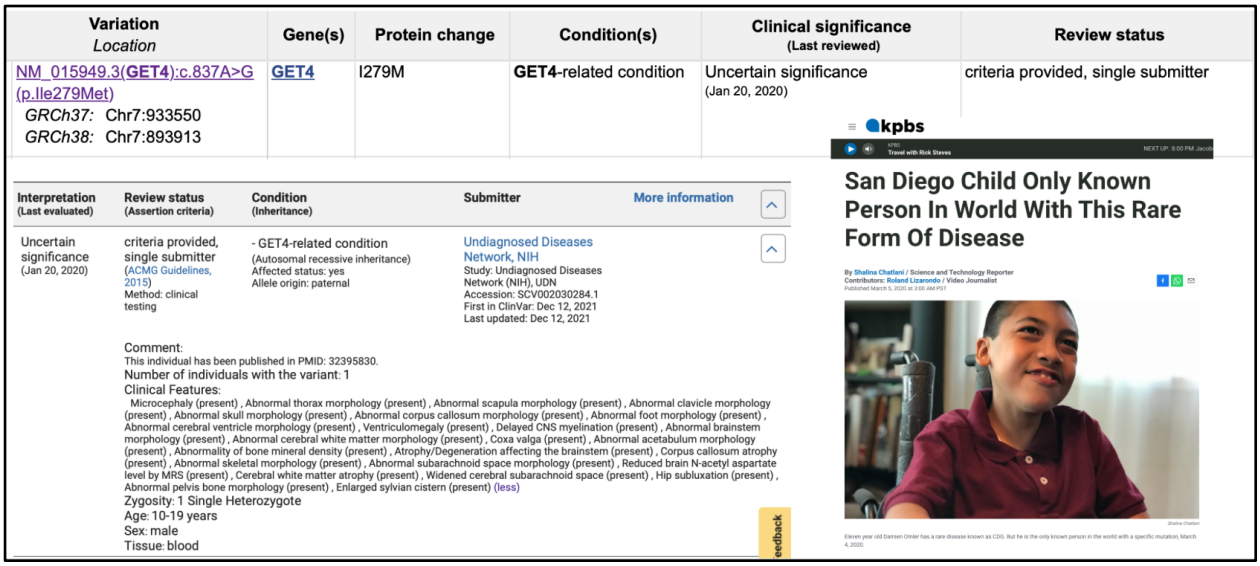

Figure 3. Re-identification Example. Results from ClinVar search for GET4-related condition including submitted clinical information. Screenshot of news article about patient with GET4-related condition likely matching the patient’s identity in the ClinVar entry.

Unfortunately, such potential risks are likely commonplace. As observed in this re-identification example, linking of variant information to PHI about the patient or extended family

is based on voluntary disclosure from the family/patient to the public. Given the rarity of some genetic conditions, it is not uncommon for the affected individual or their guardian to share their story publicly in hopes of connecting with experts or other patients who could share their lived experiences(15). Information that is collected and shared online may be linked to individuals as demonstrated in reports involving Facebook and companies in other domains (e.g. Cambridge Analytica). Facebook has been reported to approach healthcare organizations to obtain de-identified patient data to link to individual Facebook users. Self-disclosure online may include extensive developmental or medical history of the patient which may be more comprehensive than what is included in ClinVar. Nonetheless, it is important to note that in the case of self-disclosure through online posts or news articles, it is the patient or guardian making the decision to disclose the information as opposed to being potentially unaware of the extent of information submitted to ClinVar without their express consent or knowledge.

Re-identification at scale

Large scale re-identification could theoretically be achieved by joining ClinVar variant data with other public datasets to identify individuals who would not otherwise voluntarily disclose their genetic or health data. While the scope of secondary datasets was not extensively reviewed for this paper, a few examples are provided as demonstration of re-identification risk. One such example concerning minors would be identifying genetic conditions with early or sudden death. Date of death, location, and other identifiable information included in death certificate records and news articles could be utilized to identify subsets of patients with genetic variants submitted in Clinvar that are known to pose a risk for early death. A second example could connect genetic variants in ClinVar known to be associated with schizophrenia or other medical conditions associated with violent behavior with publicly available criminal records to identify risk for recidivism based on these genetic factors. While the attempt to connect recidivism to genetic profiles has largely been rejected by the medical and scientific community, it was not long ago that this was an active area of investigation by law enforcement(16).

As demonstrated from these examples of re-identification, genetic data publicly available in ClinVar could have far reaching implications for the patient. While common genetic variants which may be shared by thousands of individuals may be much less likely to be personally identifiable, rare or family-specific genetic variants pose a significant risk for re-identification. This risk poses a challenge to the clinicians and researchers submitting to ClinVar as these rare variants are the cases that are most likely to benefit from advancing clinical knowledge and research efforts.

Ethical considerations and privacy analysis

ClinVar presents some important ethical dilemmas. The organization receives and redistributes genetic data as part of an assumed deal: sources provide users with data that may violate some privacy concerns, in return for which all stakeholders benefit from advances in health care. ClinVar acts as a broker for this transaction. Although initially appealing, this ‘deal’ may become increasingly unequitable over time.

Data collection is likely not equitable, in that it is unlikely to be collected evenly between all entities who may ultimately benefit from the results from Clinvar. Therefore, results may not

be generalizable widely within society. Another source of inequitable utility of the results are well documented disparities in healthcare. Importantly, data sharing to entities not bound by HIPAA may lead to unintended consequences. Data sharing and technology transfer to commercial entities may financially benefit shareholders rather than others, even if a wider sector of society may benefit from improved therapies.

The Belmont principles(17)

The Belmont principles apply directly to ClinVar, with ethical considerations that parallel the index Belmont case where cells extracted from Henrietta Lacks in 1953 without notice and consent inadvertently led to a class of “immortal” HeLa cells, and therapy which did not benefit the Lacks family and many other disadvantaged populations.

Respect for Persons. ClinVar collects key genetic data on individuals and makes them publicly available for clinical, research and commercial purposes, pointing to other repositories which could be used for data aggregation. ClinVar provides general information to submitters regarding recommended and assumed consent and de-identification procedures, but does not require evidence from the submitter to be provided, nor does ClinVar verify the suitability of the consent obtained. Attempting to serve as a “clearing house” by placing the burden of notice and consent on individual submitters will introduce variation across the data set. Moreover, it is unlikely that notice and consent is ever adequate here - as discussed in more detail below for Nissenbaum’s contextual integrity, in terms of intended and secondary uses.

Beneficence We believe that ClinVar adheres to beneficence. The primary intended use of data is to better diagnose and treat patients. This has substantial societal benefits. Intended secondary cases are for research to better understand the link between genetic variants and disease. However, some secondary uses have less clear public good. We are particularly concerned about the public good of selling personal genetic data to undisclosed entities, who are likely not bound by HIPAA. We would recommend some legal framework be built around ultimate secondary uses as discussed below

Justice. In principle, ClinVar is open to all, with no demographic exclusions. It is also international. Increasingly, mandates from the National Institutes of Health, analogous agencies internationally and societal pressure are ensuring that efforts such as ClinVar benefit society broadly. Unfortunately, many groups with protected characteristics are acknowledged to be under-represented in medical research and in receiving the benefits of healthcare. While ClinVar cannot address all of these enormous issues, certain principles can be stated. First, ClinVar could attempt to ensure wide and representative data inclusion. Traditionally, disadvantaged groups have been excluded from data inclusion. This is true for non-health data such as credit card use or Internet history — which cause biases in credit scores or consumer profiles— as well as big health data, such as genomic databases or EHRs. This lack of inclusion may favor an individual, but mostly it disfavors those whose data are missing. For instance, if one considers allocating a scarce therapy to patients, if a particular minority group responds less well to therapy, failure to collect information on that group might lead to giving that minority group more priority than had the data been included. If the minority group responds better than other groups, the opposite effect might result. Thus, these issues are complex and likely require case-by-case analysis. More broadly, disadvantaged populations have been excluded from the benefits of healthcare, as we and others have reported (18).

Nissenbaum's contextual integrity(19,20)

Our analysis suggests that the social context of ClinVar is that of a laboratory test at a medical facility. A major departure from this is that a subjects PHI are distributed beyond the "need to know" that is central to HIPAA. From a legal perspective, HIPAA lists circumstances under which PHI can be disclosed without patient authorization, and ClinVar may fall under the 4th provision (of 5): for public health activities. The mission of ClinVar, if handled well, is for the public good of disseminating data to other sufferers, and to facilitate treatment. The other 4 provisions for PHI disclosure without consent do not apply: for treatment or payment (only substance-abuse and written psychotherapy notes are prohibited), if the patient can agree to or object to any disclosure, during a natural disaster, and if requested by court orders(4)

We conclude that notice and consent are inadequate for the complex mission of ClinVar: to acquire, store and distribute data on individual genetic variants. In the current era of rapid genome sequencing and ' fingerprinting', genetic data is the richest source of PHI or PII that exists. No human being likely fully understands the implications of current or future uses of such data. Few subjects will understand that sensitive genetic data could be passed to entities not bound by HIPAA, for instance by their employer to deny a position or promotion, or by their insurance company to deny benefits. Individuals may not fully appreciate that data passed to commercial entities may also not be used to develop therapy, but to sell for profit such as reported for 23andMe(13). Even therapies developed by companies from these data may ultimately not benefit the subject or others in disadvantaged or vulnerable populations due to the very long time window from data acquisition to therapy availability, or to disparities in healthcare as discussed above in the Belmont Report on Justice. On the other hand, it is difficult to ignore the transformative potential of using genetic data to better detect sufferers of disease, improve diagnosis and develop new treatments.

Mulligan's privacy framework(21-23)

This framework can be discussed in detail here. In the **dimension of theory,** privacy could be construed as required to prevent potentially sensitive genetic information from going to unintended actors, from being used for unintended purposes, while data are available. The contrast concept is that if genetic information were freely shared, individuals could be denied jobs, life partners, or insurance based on inferred yet unproven risk of disease. The **dimension of protection** includes the subject, initially individuals who share their genetic data, and the target, those who use these data. In a co-existing scenario, those who share in the use of the genetic data could also be considered subjects, the target now being corporations, news media or insurance companies. In fact a complex web of subject-target interrelationships can be outlined due to the public nature of the data, and the fact that many who have access are not bound by HIPAA. **Dimensions of harm** include several actions, including publicizing information on genetic variants, and publicizing the identity of someone with a potential disease (from: individual, to: news agency). Other actions include denial of an individual with a pre-existing condition for health insurance (from: patient, to: insurance company), a job (from: patient, to: potential employer), or social interactions through stigmatization (from: patient, to: society). Others include similar interactions to the data user, if they are a patient (from: patient, to: patient). **Dimensions of provision** should include legal statements of permitted and non-

permitted use cases, mechanisms to enable data editing, correction and removal, full transparency on who is using data, and for what purposes. **Dimensions of scope** are broad, as are the potential uses of personal genetic data. We do not believe that there is clearly definable social scope, since publicly available data by non-HIPAA bound actors has no limit. Temporal scope cannot clearly be defined; HeLa cells are still being used after 70 years, for instance.

Solove's taxonomy (24)

There are many aspects of Solove's taxonomy which are pertinent to ClinVar, since Fair Information Practice Principles (FIPPs) should mitigate against harms. Information collection is essentially by interrogation, although some uploaded data could be aggregated and collected by surveillance. Collection excludes several individuals from society, as discussed in the Belmont principles, and there should be provision for increasing awareness of this data repository for the public good. Information Processing is for a series of use cases as described above. We believe that certain secondary uses violate contextual integrity (see above in Nissenbaum's framework) and should be excluded or at least subject to opt-in by specific notice and consent. Aggregation of data with other sources is a major source of ethical angst, yet is also the most powerful scientific tool to use these data to find cures - aggregating with clinical data (e.g. "did the patient with this variant respond to therapy?"). This requires societal debate. We believe that data insecurity has not been well addressed. Substantial harm could come from incorrect data without adequate provisions for security, from leakage of data, and from re-identification as discussed above. Dissemination of data is essentially by user 'pulls' from the public repository. We believe that this should be restricted, so that uses can be better directed by data providers and society. The risks of sharing correct or incorrect data could be devastating. Invasions cause intrusions if genetic data are used by unintended non-HIPAA bound actors for non-healthcare purposes. This may lead to decisional interference - in that an individual may no longer have access to specific employment (e.g. the military or other agencies) or insurance without having given consent for this.

Legal considerations

Several legal frameworks are applicable to genetic information in addition to HIPAA and HITECH described above. This legal analysis will touch on the most relevant considerations for privacy risk and is not an exhaustive analysis. In regards to genetic data, the most well-known legislation is the Genetic Information Nondiscrimination Act (GINA) which was established in 2008 and is intended largely to prevent discrimination in employment and insurance coverage based on genetic test result or genetic predisposition alone(25). Importantly, GINA does not provide protection for those enlisted in the military, or from denials for life insurance or long term care insurance coverage. Some states have enacted additional protections on top of GINA, such as CalGINA which extends protections to include non-discrimination in housing, mortgage lending, education, and other state funded programs.

Additional laws may also apply to genetic variant data. For example, there are many citizens of the European Union represented in ClinVar for which the General Data Protection Regulation (GDPR) would apply. Specifically, GDPR considers irreversibly de-identified data as

non-personal data, which further introduces the question of whether the variant information stored in ClinVAR is irreversibly de-identified (26). Arguably, in the situation where a genetic variant is unique to an individual and additional data fields are available such as location of testing lab, date of testing, sex, age, etc, this is not irreversibly de-identified. It is unclear if there are any additional protections available under state-specific privacy laws such as CalOPPA as these laws are often related to the information that is tracked and collected by the website entity itself. In the case of ClinVar, the privacy concerns don't necessarily relate to tracking, but the submitted data that is made publicly available.

Consumer protections for individuals exist, but it is important to note that protections would apply to the services the patient received by the genetic testing laboratory and would likely not implicate ClinVar itself. To this point, the Federal Trade Commission (FTC) provided a policy statement on biometric information in May 2023(27). This policy statement specifies that genetic information is included under the definition of biometric information and that companies have a responsibility for providing clear terms of use and privacy practices notifications as well as proper security practices to ensure protection of genetic data. The FTC has taken action against companies who have had insufficient notices and protection practices(27). If a consumer of a genetic test was to encounter harm due to the data sharing practices with ClinVar, they may have avenues for litigation under consumer protection laws such as the FTC or other state-specific consumer laws.

Recommendations for future iterations of ClinVar

ClinVar provides a data resource of critical value which can literally help to save lives. While acknowledging this immense benefit as healthcare professionals, we also conclude that ClinVar publicly provides PHI that introduces several unintended risks. When aggregated with secondary data, it may be relatively easy to re-identify individuals which violates privacy and has substantial ramifications. Given these risks, the following recommendations may be considered in order to reduce risk and potential harm:

- The issues of benefit versus harm for this healthcare application are likely to be common debates in coming years, yet quite distinct to non-health domains. We would ordinarily recommend tougher controls for ClinVar including strict differential privacy – data access for specific healthcare workers but not other members of the public. This would satisfy several privacy concerns. However, such data safeguards come with a real human cost: they could greatly hinder the ability for other patients, relatives or companies to use data to assist in diagnosis and therapy, and this could stifle life-saving innovation.
- It is still our conclusion that stricter safeguards are required, but this is clearly a societal debate with broad ramifications that will affect all of digital health care. At a first level, we believe that ClinVar should provide a more comprehensive statement of intended uses, to attempt to constrain certain types of secondary use. This may data aggregation for purposes not designed to advance treatment, such as data sharing with employers or insurance companies. Such intended use disclosures could be based on input from patient advocacy input and professional societal guidelines.

- ClinVar should better enable data security at least to make patients and submitters better aware of the ability to edit data, and potentially develop intuitive tools for this goal. Such tools may themselves advance the field and thus could be an initiative of the NIH. In this way, privacy and security may become design features, which develop as uses and therapies arise.
- ClinVar should discuss differential privacy, making certain data available only to specific users such as acknowledged treatment providers. Additional features of differential privacy include strategies such as injecting noise into the dataset, yet this itself poses real harm for medical or genetic data. The accuracy of data is integral to the mission of understanding and treating disease. Again, these topics are likely to be ongoing societal debates.
- Finally, it would be useful to provide a process for notifying previous users who downloaded static versions of the database containing PII/PHI that an update is required and previously stored versions should be destroyed.

Conclusions

ClinVar is a unique and highly valuable resource for the advancement of health. Data from Clinvar has directly and indirectly saved many people's lives. Nevertheless, the richness and depth of genetic data on specific individuals raises many concerns and opens ClinVar to privacy and security vulnerabilities. Next iterations must integrate privacy and security as design features, which may develop with novel use-cases and evolving medical therapies over time. Moreover, some secondary are not clearly in the public good and could be curtailed or subjected to strict differential privacy protocols. We are particularly concerned about the current practice of the for-profit sale of personal genetic data to undisclosed entities who not bound by HIPAA. We would recommend legal frameworks be built around ultimate secondary uses. Such revisions would provide better tools, hand-in-hand with improved security. Additionally, while not within the scope of our commissioned report, we recommend that ClinVar rises to the need for a formal privacy impact assessment.

References

1. Callaway E. Supercharged crime-scene DNA analysis sparks privacy concerns. Nature 2018;562:315-316.
2. Kuo TT, Jiang X, Tang H et al. The evolving privacy and security concerns for genomic data analysis and sharing as observed from the iDASH competition. J Am Med Inform Assoc 2022;29:2182-2190.
3. Srivastava AK, Wang Y, Huang R et al. Human genome meeting 2016 : Houston, TX, USA. 28 February - 2 March 2016. Hum Genomics 2016;10 Suppl 1:12.
4. Moore W, Frye S. Review of HIPAA, Part 1: History, Protected Health Information, and Privacy and Security Rules. J Nucl Med Technol 2019;47:269-272.
5. Azzariti DR, Riggs ER, Niehaus A et al. Points to consider for sharing variant-level information from clinical genetic testing with ClinVar. Cold Spring Harb Mol Case Stud 2018;4.
6. Moore W, Frye S. Review of HIPAA, Part 2: Limitations, Rights, Violations, and Role for the Imaging Technologist. J Nucl Med Technol 2020;48:17-23.
7. Price WN, 2nd, Cohen IG. Privacy in the age of medical big data. Nat Med 2019;25:37-43.
8. National_Institutes_of_Health. NCBI Website and Data Usage Policies and Disclaimers https://www.ncbi.nlm.nih.gov/home/about/policies/. 2023.
9. Invitae. Invitae Consent: genetic counseling and telehealth, https://www.invitae.com/us/individuals/consent/gc-and-telehealth. 2023.
10. DHSS. Health Information Privacy. https://www.hhs.gov/hipaa/for-professionals/special-topics/research/index.html. 2023.
11. Commonwealth_fund. 2001 International Health Policy Survey https://www.commonwealthfund.org/publications/surveys/2002/may/2001-international-health-policy-survey. 2001.
12. Shirts BH, Pritchard CC, Walsh T. Family-Specific Variants and the Limits of Human Genetics. Trends in molecular medicine 2016;22:925-934.
13. Segert J. Understanding Ownership and Privacy of Genetic Data https://sitn.hms.harvard.edu/flash/2018/understanding-ownership-privacy-genetic-data/. 2021.
14. Dewar R, Claus AP, Tucker K, Ware R, Johnston LM. Reproducibility of the Balance Evaluation Systems Test (BESTest) and the Mini-BESTest in school-aged children. Gait Posture 2017;55:68-74.
15. Iyer AA, Barzilay JR, Tabor HK. Patient and family social media use surrounding a novel treatment for a rare genetic disease: a qualitative interview study. Genet Med 2020;22:1830-1837.
16. Tehrani JA, Sarnoff A. Mednick. Genetic Factors and Criminal Behavior. FEDERAL PROBATION 2000;64.
17. National_register. Protection of human subjects; Belmont Report: notice of report for public comment. National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. Fed Regist 1979;44:23191-7.
18. Gomez SE, Fazal M, Nunes JC et al. Racial, ethnic, and sex disparities in atrial fibrillation management: rate and rhythm control. J Interv Card Electrophysiol 2022.
19. Nissenbaum H. A Contextual Approach to Privacy Online. Daedalus 2011;4:32-48.
20. Nissenbaum H. Contextual Integrity Up and Down the Data Food Chain. Theoretical Inquiries L 2019;221:20
21. Mulligan DK, Griffin D. Rescripting Search to Respect the Right to Truth. Georgetown Law Review 2018:557.

22. Mulligan DK, Kluttz D, Kohli N. Shaping Our Tools: Contestability as a Means to Promote Responsible Algorithmic Decision Making in the Professions. 2019.
23. Mulligan DK, Koopman C, Doty N. Privacy is an essentially contested concept: a multi-dimensional analytic for mapping privacy. Philosophical Transactions of the Royal Society A 2016.
24. Solove DJ. A Taxonomy of Privacy. University of Pennsylvania Law Review 2006;154:477.
25. National_Human_Genome_Research_Initiative. Genetic Discrimination https://www.genome.gov/about-genomics/policy-issues/Genetic-Discrimination. 2023.
26. Shabani M, Vears D, Borry P. Raw Genomic Data: Storage, Access, and Sharing. Trends Genet 2018;34:8-10.
27. FTC. Privacy and security of genetic information: Putting DNA companies to the test https://www.ftc.gov/business-guidance/blog/2023/06/privacy-security-genetic-information-putting-dna-companies-test
. 2023.