1 Title

Quantifying the impact of data sharing on variant classification

2 Authors and affiliations

James Casaletto [1], Melissa Cline [1], Brian Shirts [2]

1. Genomics Institute, University of California, Santa Cruz
2. Department of Laboratory Medicine and Pathology, University of Washington, Seattle

3 Corresponding authors

James Casaletto (jcasalet@ucsc.edu), Brian Shirts (shirtsb@uw.edu)

4 Abstract

Healthcare is increasingly leveraging genomic data to inform diagnosis, monitoring, and treatment of certain diseases with genetic predisposition. Associating patient data such as family history and *de novo* status with a genomic variant helps classify that variant as being pathogenic or benign. Indeed, many variants are already classified by experts, but the majority of variants are very rare, have no associated patient data, and are therefore of uncertain significance. This research models the hypothetical sharing of patient data across institutions in order to accelerate the time it takes to classify a variant. Using conservative assumptions described in the paper, we found that the probability of classifying a pathogenic variant which occurs at the rate of 1 in 100,000 people increases from less than 25% to nearly 80% after just one year when sequencing centers share their clinical data. After 5 years, the probability of classifying such a variant is nearly 100%.

5 Introduction

Targeted gene sequencing is becoming more common for patients to determine if they have known pathogenic variants for symptoms they present or diseases to which they are susceptible. These pathogenic variants may inform doctors and clinical geneticists how to manage their patients' health. For example, a patient with a known pathogenic variant in BRCA1 or BRCA2 should, at the very least, be screened more often for breast, ovarian, and pancreatic cancer. Similarly, asymptomatic patients with familial cardiomyopathy might consider certain lifestyle choices such as losing weight, reducing stress, quitting smoking, sleeping well, and perhaps taking ACE inhibitors and/or beta blockers. [2]

The American College of Medical Genetics (ACMG) and the Association for Molecular Pathology (AMP) define evidence-based guidelines for classifying genomic variants. Evidence for variant classification can come from many sources including family history, clinical data, functional assays, and *in silico* predictors. When sufficient evidence is present, a variant curation expert panel (VCEP) may classify the variant as likely benign (LB), benign (B), likely pathogenic (LP), or pathogenic (P). Variants with little or no evidence to support classification are called variants of uncertain significance (VUS). The classification of VUS is the objective of this research. [1]

Variants of uncertain significance pose problems in healthcare. One problem is the preponderance of VUS among genes implicated in disease, as VUS by definition don't provide any medically actionable information. Ultimately, every variant is either physiologically benign or pathogenic, so the significance of a VUS is only uncertain until there is sufficient evidence to classify it. Furthermore, clinical data is usually needed to classify VUS [1]. A second problem is the lack of centrally available clinical data sufficient to classify a variant. Molecular testing laboratories and sequencing centers are the largest source of variant data, but they hold their data privately. Yet a third problem is that the classifications of some genetic variants may vary or even conflict between laboratories, depending on the amount and nature of the evidence provided. [3]

One solution to all these problems is for laboratories to share their data. The more that laboratories and clinics pool their variant data, the more likely and more quickly a VUS may be classified.[17-19] Such pooling of data leads to more expedient VUS classification which is of utmost importance to the ultimate benefactors of this information: the patients with these VUS. Factors such as maintaining market advantage and preserving privacy reasonably preclude data sharing, but those obstacles may be overcome by, for example, leveraging federated computing.

The purpose of this research is to model how long it is expected to take and simultaneously how likely it is for a VUS to get classified when sequencing centers pool their clinical data with other sequencing centers. This model predicts the probability of variant classification over time given the clinical data currently available in laboratory databases and estimates of future variant-specific data accumulation rates. Underlying the model are factors such as prior probability of pathogenicity and allele frequency specific to the genes of interest. Our modeling approach may be applied to any gene implicated in a monogenic disease. The output of this model can guide VCEPs in prioritizing their efforts, inform functional assay developers on high-impact variants which cannot be classified through patient-derived data alone, and enable healthcare providers to develop better strategies for managing patients with VUS.

6 Materials and Methods

In this section, we define a statistical model that combines clinical information from multiple sequencing centers to create an aggregate, pooled center so that

VUS may be classified faster.

6.1 Combining multiple pieces of evidence to classify variants

The evidence that the ACMG/AMP uses to classify variants encompasses several sources of data, including the type of variant (e.g. nonsense or frameshift), *in vitro* functional studies, *in trans* co-occurrence with a pathogenic variant, co-segregation in family members, allele frequency, and *in silico* predictions. Tavtigian et al [4] showed that the ACMG/AMP variant classification guidelines could be modeled as a Bayesian classification framework. Specifically, the ACMG/AMP classification criteria were translated into a Bayesian classifier, assuming four levels of evidence and exponentially scaled odds of pathogenicity. These four levels include "supporting", "moderate", "strong", and "very strong". For example, one category of evidence is called PP1 which represents co-segregation of the disease with multiple family members. The PP1 evidence is considered "supporting" evidence for pathogenicity. Another example evidence category is BS4 which represents the lack of segregation in affected family members. The BS4 evidence is considered "strong" evidence against pathogenicity. These odds are combined and compared to thresholds to determine the variant's pathogenicity.

We leverage this Bayesian framework using a frequentist approach to model probabilities of pathogenicity conditioned on the presence of one or more pieces of evidence for a given variant. A single piece of evidence is represented as an odds of pathogenicity. Clinical evidence that is observed for the same variant from unrelated patients is independent, so odds from multiple observations may be combined multiplicatively. The odds of a variant V_i being pathogenic (belonging to the class P) given all the evidence X_j is the product of all the evidence, expressed as odds, as shown in Equation 1.

$$odds(V_i \in P|X_j) = \prod_j X_j \tag{1}$$

The product of all the evidence may yield a very large number, causing numerical overflow on a 64-bit machine. We convert the odds of pathogenicity to a log scale by taking the log of both sides of equation (1), as shown in equation (2).

$$log(odds(V_i \in P|X_j)) = \sum_j log(X_j) \tag{2}$$

For a single variant, we simply compare this sum to the thresholds for benign, likely benign, likely pathogenic, and pathogenic in log scale, as defined in Table 4. The same logic may be applied to calculate the overall odds that the variant is benign, as shown in equation (3).

$$log(odds(V_i \in B|X_j)) = \sum_j log(X_j) \tag{3}$$

Our model calculates two odds of pathogenicity: the odds of a VUS being benign and the odds of a VUS being pathogenic, both of which are conditioned on statistically sampled evidence.

6.2 Selecting categories of benign and pathogenic variant evidence for our model

Some sources of evidence are always available and relatively stable over time, including information about the nature of the variant in its protein and information from *in silico* predictions. Increased functional information will contribute to variant reclassification, but there are risks to using such data without any clinical data to classify variants. We will not use those categories of evidence in our model, but rather we will focus on the clinical data that is available in sequencing facilities. Several sources of case and family information will contribute to variant classification over time. As clinical databases grow and data is shared more effectively across institutions, more variants will be classified. Increased clinical information is the major source for variant reclassification as well. [7] We selected only those categories of pathogenic variant observations that relate to clinical information in the development of our model, as shown in Table 1. [1]

Category	Criteria
PM6	Assumed *de novo*, but without confirmation of paternity and maternity
PP1	Co-segregation with disease in multiple affected family members in a gene definitively known to cause the disease
PS2	*De novo* (both maternity and paternity confirmed) in a patient with the disease and no family history

Table 1: ACMG/AMP criteria for categories of pathogenic variant observations

We equivalently selected only those categories of benign variant observations that relate to clinical information in the development of our model, as shown in Table 2.

Category	Criteria
BP2	Observed in trans with a pathogenic variant for a fully penetrant dominant gene/disorder or observed in cis with a pathogenic variant in any inheritance pattern
BP5	Variant found in a case with an alternate molecular basis for disease
BS4	Lack of segregation in affected members of a family

Table 2: ACMG/AMP criteria for categories of benign variant observations

Over time, the more evidence that is gathered, the sooner and more likely a

VUS will be classified as either benign or pathogenic. [8] However, not all the evidence that is gathered over this time will be concordant. For example, some patients who have a VUS which is actually pathogenic may occasionally present evidence from one or more benign categories. This presentation of conflicting evidence for a given variant occurs at a low, non-zero frequency. Therefore, we use a combination of pathogenic and benign evidence in the classification of every VUS.

6.3 Making assumptions for the model

Here we discuss the various assumptions we made about the data in order to build our simulation.

6.3.1 Assumption 1: Frequency distribution for evidence

We made assumptions about the frequencies at which each of the ACMG/AMP categories described in Tables 1 and 2 are observed. We derived these assumptions from the scientific literature. [9-15] These assumptions allowed us to simulate the gathering of evidence over time at different distributions of participating sequencing centers in the effort of classifying VUS as either benign or pathogenic. If sequencing centers were to participate and share data and we suggest here, these frequencies would be replaced with the data provided by each of the participating centers to build a more accurate model. This is the data that we suggest to share across participating institutions for more accurately building these classification timeline models.
Tavtigian et al calculated the odds of pathogenicity for each category of ACMG/AMP evidence. Specifically, they determined that, for pathogenic evidence, the odds for "strong" is 18.7, for "moderate" it's 4.3, and for "supporting" it's 2.08. For benign evidence, the odds for "strong" is 1/18.7, for "moderate" it's 1/4.3, and for "supporting" it's 1/18.7. Table 3 depicts these odds and their associated assumed frequencies for each ACMG/AMP evidence category.

benign observations

Category	Frequency	Odds
PM6	0.0035	4.3
BP2	1.0*frequency	1/2.08
BP5	0.07	1/2.08
PP1	0.01	2.08
PS2	0.0015	18.7
BS4	0.1	1/18.7

pathogenic observations

Category	Frequency	Odds
PM6	0.007	4.3
BP2	0.005*frequency	1/2.08
BP5	0.0001	1/2.08
PP1	0.23	2.08
PS2	0.003	18.7
BS4	0.0001	1/18.7

Table 3: Frequencies and odds per ACMG/AMP category for benign and pathogenic observations

There may be pathogenic categories of evidence observed for benign variants and benign categories of evidence observed for pathogenic variants, though these

evidence categories which conflict with the variant pathogenicity generally occur at a low rate. We are assuming that the frequency of BP2 evidence (*in trans* co-occurrence with a known pathogenic variant) for pathogenic variants is very rare, except in tumors or in the case of rare diseases such as Fanconi anemia. Conversely, we assume that the frequency of BP2 evidence for benign variants is quite common and so occurs at the same rate as the variant itself.

6.3.2 Assumption 2: Thresholds for odds of pathogenicity

Tavtigian et al, in interpreting the ACMG/AMP variant classification guidelines as a naive Bayesian classifier, defined four threshold ranges for the odds of pathogenicity for each of the four variant classifications (benign, likely benign, likely pathogenic, pathogenic). These odds threshold ranges are shown in Table 4.

Classification	Odds threshold range
Benign	$[-\infty, 0.001)$
Likely benign	$[0.001, 1/18.07]$
Likely pathogenic	$[18.07, 100]$
Pathogenic	$(100, +\infty]$

Table 4: Odds threshold ranges for naive Bayesian classifier

These odds threshold ranges are based on the guidelines set forth by the ACMG/AMP. For example, the ACMG/AMP defined the term "likely pathogenic" to mean > 0.90 certainty of a variant being disease-causing but below the higher pathogenic threshold of 0.99. Translating the ACMG/AMP certainty cutoff into Bayesian terms, 0.90 is a posterior probability which corresponds to a pathogenicity odds of 100. Similarly, the likely pathogenic odds are between 0.90 and 0.99, the likely benign threshold is between 0.001 and 0.10, and benign threshold is less than 0.001.

6.3.3 Assumption 3: Data from participating sequencing centers

In advance of getting real data from any sequencing centers, we made certain assumptions about how much data each general size of sequencing center has as well as how many new observations are added per year. Table 5 depicts these assumptions.

Sequencing center size	Initial size	Tests per year
small	15000	3000
medium	150000	30000
large	1000000	450000

Table 5: Data size and rate assumptions per sequencing center size

These sizes and test rates were inferred from experience and online public business reports from genetic testing companies. Should a group of sequencing centers decide to pool their data, they will be able to replace these values with better estimates.

6.3.4 Assumption 4: Ascertainment bias

Another assumption we make relates to ascertainment bias at testing centers. Ascertainment bias, the medical term for statistical sampling bias, describes systematic deviations from an expected result due to the sampling processes used to find genomic variants and estimate their population-specific allele frequencies. People who go to testing centers are often referred there because their medical care providers suspect they may have a genetic disease. How much more likely is a person to present pathogenic evidence than benign evidence is captured in our model as a real-valued variable called pathogenic selection factor (PSF). We conservatively estimated this term to be 2 based on experience at the University of Washington Department of Laboratory Medicine.

6.3.5 Assumption 5: Prior odds of pathogenicity

Yet another assumption we made was regarding the prior odds of a variant's pathogenicity. This metric comprehends all other criteria that are not clinical and doesn't change much, if at all, over time. For the sake of this implementation, we sampled a random odds from a uniform distribution between 0.11 and 9 (which are the odds associated with probabilities between 0.1 and 0.9, respectively).

6.4 Implementing the simulation

The implementation is very straightforward. The simulation has only one random variable as input which is the allele frequency of the VUS of interest. The code iterates through the sequencing centers one year at a time to accumulate new evidence for a variant of the input frequency, where the initial size and yearly growth rate are depicted in Table 4. Because the variant is of uncertain significance (i.e. we don't know if it's benign or pathogenic), we gather evidence for both classifications simultaneously. So, for example, if we simulate gathering evidence for 1,000 VUS, we would have 2,000 sets of observations - 1,000 given the variant is benign and 1,000 given the variant is pathogenic.

Pseudo-code is shown below for simulating the gathering of evidence for a single variant using probabilities and odds for pathogenic and benign evidence derived from Table 3. The initial sizes and tests per year parameters are derived from Table 5.

```
years = 5
vusFrequency = 1e-05
s=10, m=7, l=3
centers = [ s*Center('small') ] +
          [ m*Center('medium') ] +
          [ l*Center('large')]
for center in centers:
    center.initialize(vusFrequency)
    allCenters += center.getEvidence()
    for year in range(years):
        center.run(vusFrequency)
        allCenters += center.getEvidence()
```

The `initialize()` method creates the initial set of evidence based on the initial sequencing center size from Table 5. Using the `getEvidence()` method, all the evidence that is initialized at each of the individual testing centers is then added to the `allCenters` object which represents the comprehensive shared data set across all the testing centers. The `run()` method simulates people getting tested at a sequencing center. Here we use a Poisson distribution sampling method when determining how many times variant is observed, given the presumed `vusFrequency`.

We execute the `run()` method once for each year in the range of years parameter and accumulate simulated data discretely (as opposed to continuously) over all the centers. The size of a center's population from which we sample and the number of tests per year are coded in the `Center` object according to small, medium, and large sizes described in Table 5. Each year, and for each center, we combine the odds of the pathogenic and benign observations to `allCenters` to simulate the sharing of data across all the sequencing centers over time.

Running the simulation as shown in the pseudo-code generates data points across sequencing centers over time for a single VUS of the given frequency. We ran this simulation several times to generate data points for multiple VUS of the same given frequency. These results are discussed in the next section.

7 Results

For each sequencing center, we simulated the gathering of evidence for 1,000 different variants with allele frequency 1e-05 over the course of 5 years. We generated data each year and took snapshots to create histograms and scatter plots that show the distribution and progression of the evidence over time. We plot the probability of classifying the variant as benign (B), likely benign (LB), likely pathogenic (LP), or pathogenic (P). Additionally, we calculated the probability of a variant being classified at any sequencing center, assuming that all these centers would share their variant interpretation even if they don't share their clinical data. Finally, we performed a sensitivity analysis to show

the impact each of the model parameters has on the probability of being either benign or pathogenic. We randomly selected one small, one medium, and one large sequencing center to show in the plots.

7.1 Histogram plots show that the distributions of evidence when sharing clinical data are sufficiently wide to cross classification thresholds

Figure 1 shows the distribution of evidence gathered at particular small, medium, and large sequencing centers as compared to the combined data across all sequencing centers.

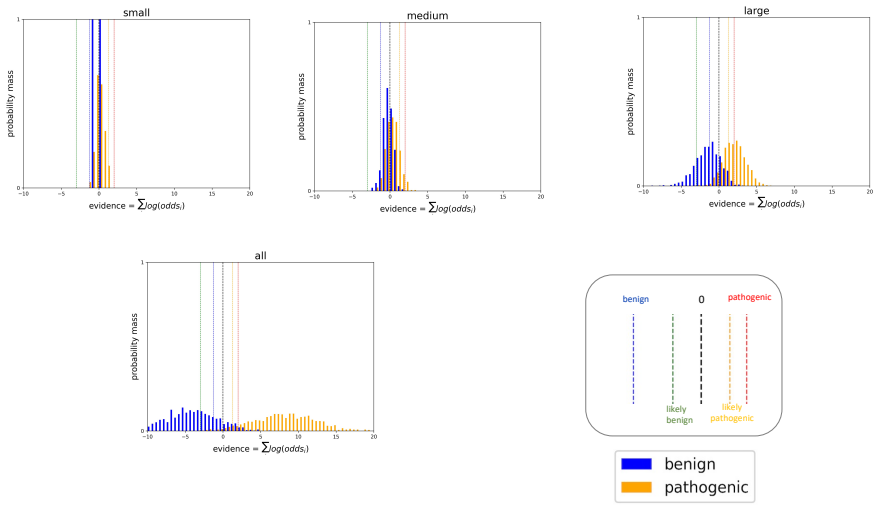

Figure 1: Distribution of accumulated evidence over 5 years at sequencing centers

The histogram plots in Figure 1 show that all the evidence in small-sized sequencing centers, except for some evidence which just crosses the threshold for pathogenic, is insufficient. Similarly, most of the evidence from medium-sized sequencing centers is insufficient for any classification. But when taken across all centers together, the evidence exceeds the thresholds required to classify approximately one-half of the benign variants and over 90% of the pathogenic variants. The distributions appear bell-shaped and are presumably normal distributions. We expect this to reflect the real distributions of evidence as the amount of evidence increases, as described by the central limit theorem in frequentist statistics.

7.2 Scatter plots shows that evidence for a given variant may contradict over time

Figure 2 shows the trajectory of evidence gathered at a small, medium, and large sequencing center as compared to the combined data across all sequencing centers. Where the thresholds were demarcated as vertical hash lines in the previous plots, they are demarcated as horizontal hash lines in these plots.

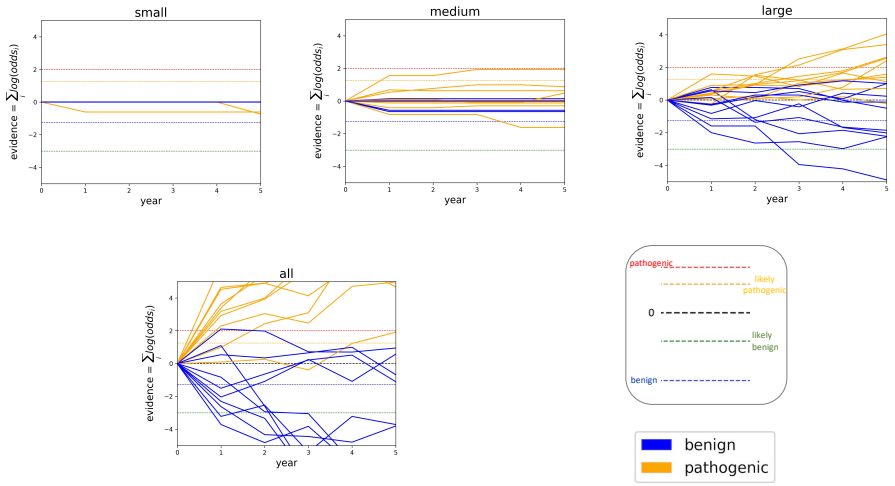

Figure 2: Example trajectories of accumulated evidence for 10% of variants modeled in each scenario

The scatter plots in Figure 2 show the trajectory of evidence over time. The trend of evidence for pathogenic variants is more clearly classifiable than the trend of evidence for benign variants. In both cases, the evidence goes in both directions. That is, sometimes there is benign evidence for a pathogenic variant and vice versa. Importantly, the scatter plots cross the classification thresholds more quickly when all the variants are combined as compared to individual sequencing centers.

7.3 Probability plots show that variants are classified sooner and with higher probability when data is shared

Figure 3 shows the probability of classifying a variant as either benign, likely benign, likely pathogenic, or pathogenic over time. Here we consider a small, medium, and large sequencing center which are not sharing anything as compared to two forms of sharing: centers sharing their all of their variant interpretations but none of their clinical data (labeled "any"); and centers sharing all their clinical data (labeled "all").

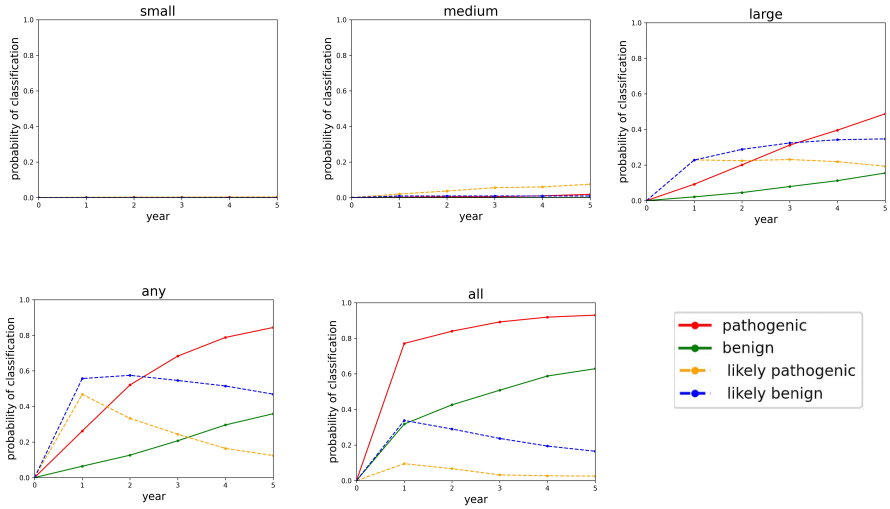

Figure 3: Probability of classification over 5 years at sequencing centers

The probability plots in Figure 3 show the trend that neither a single small-sized nor medium-sized sequencing center has sufficient evidence to classify VUS as benign within the first five years. Even the large-sized center only classifies about 20% and 40% of variants as benign or pathogenic, respectively. When centers share their variant interpretations, these probabilities increase to about 40% and 80%. When centers share their clinical data, there's almost a 60% and 90% chance of classifying a variant as benign or pathogenic after five years. The "any" plot showing interpretation-sharing includes discrepancies between different laboratories, for example, the same variant could be likely pathogenic or likely benign at different laboratories. This feature minimizes the apparent benefit of true data sharing.

7.4 Sensitivity Analysis

We used confidence intervals (low, expected, and high values) around the frequencies defined in Table 3 in order to determine how sensitive the probabilities of classification were to each of the ACMG/AMP evidence values. We held all other parameters constant (equal to their expected values) while changing one frequency at a time to a low and high value in the confidence interval. After running through each of the possible values, we arrived at the tornado plots in Figures 4 and 5 for benign classifications and pathogenic classifications, respectively.

7.4.1 Sensitivity analysis shows that the BS4 and BP5 ACMG criteria have highest impact on benign classification

The tornado plot in Figure 4 shows the ACMG/AMP evidence criteria that have the most profound impact on the probability of classifying a variant as likely benign or benign.

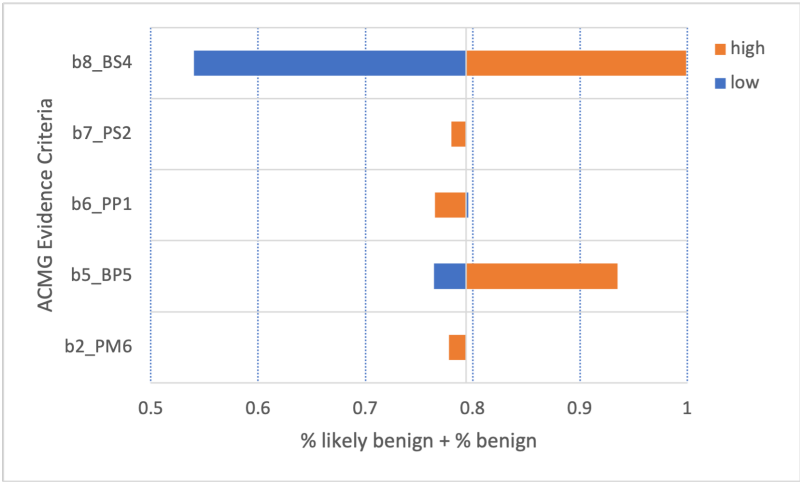

Figure 4: Tornado plot for benign classification

The parameters BS4 and BP5 have the highest impact on a benign classification. The BS4 evidence category represents the non-segregation of the disease with family members. The frequency of BS4 is 0.1 and it's odds of pathogenicity is 1/18.7. This is the both the highest frequency for benign observations as well as the lowest odds of pathogenicity among the ACMG criteria we included in our model. The probability of classifying a variant as benign is most sensitive to this value. Similarly, the frequency of BP5 (evidence that represents having found the disease due to another, non-genomic cause) is the second highest and it's odds of pathogenicity the second lowest among the ACMG criteria included in our model.

7.4.2 Sensitivity analysis shows that BS4 and PP1 ACMG criteria have the highest impact on pathogenic classification

The tornado plot in Figure 5 shows the ACMG/AMP evidence criteria that have the most profound impact on the probability of classifying a variant as likely pathogenic or pathogenic.

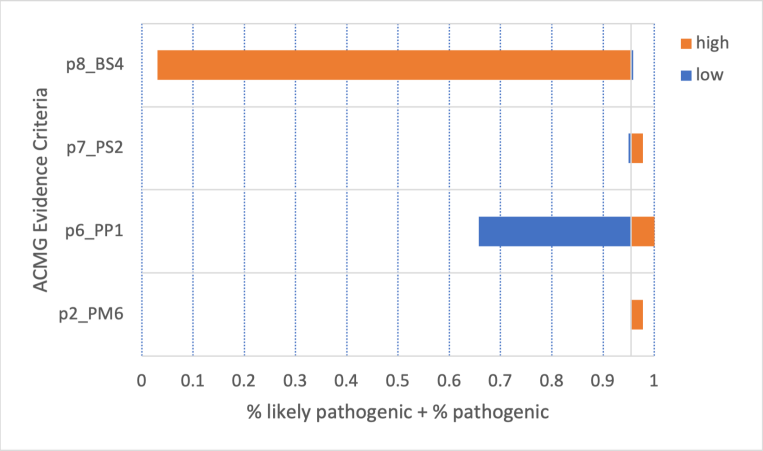

Figure 5: Tornado plot for pathogenic classification

The tornado plot illustrates that the PP1 evidence criterion has the most influence on the percentage of variants classified as pathogenic is PP1. The PP1 evidence category is derived from co-segregation of disease with multiple family members and has the highest frequency in our model (0.2) as well as the second highest odds of pathogenicity (2.08). Similarly, the ACMG/AMP evidence criterion we included in our model which has the most negative influence on the percentage of variants classified as pathogenic is BS4 (non-segregation with disease in family members).

8 Discussion

The simulations show that classifying pathogenic variants has a higher odds and quicker timeline than for classifying benign variants. This is because, according to the ACMG/AMP evidence criteria and classification guidelines, more evidence is required for benign classification. In terms of cost-benefit analysis in human variant curation, having a higher evidence threshold for benign variant classification makes sense as there is a higher cost associated with mis-classifying a variant as benign than as mis-classifying a variant as pathogenic.

The simulations also show that evidence for a given variant can be contradictory. As defined in the ACMG/AMP classification standards, evidence of pathogenicity may be presented for benign variants (and vice versa), though less frequently than for pathogenic variants. This is because there are other factors besides genetic variation involved in human health, and disease is not deterministic.

Most importantly, the simulations illustrate that sharing clinical data increases the probability of and accelerates the time to classify VUS. Knowing

how long and how likely classification is for a VUS with a particular frequency can guide important decisions for patients and their healthcare teams. For example, a patient with a known pathogenic variant on BRCA1 or BRCA2 may elect to have a prophylactic mastectomy. According to the National Cancer Institute, bilateral prophylactic mastectomy reduces the risk of breast cancer in women who carry a BRCA1 or BRCA2 by 95% [16]. A patient with a VUS on BRCA1 or BRCA2 may choose to wait for their VUS to get classified if their variant is likely to be classified in the near-term (e.g. within 5 years) but may not choose to wait if that variant will not likely get classified for another 20 years or more. The majority of variants in the BRCA1 and BRCA2 genes are of uncertain significance, yet these are two of the most widely studied and documented genes in the human genome. Other Mendelian diseases with highly penetrant alleles have an even higher percentage of unclassified variants. Similarly, having an approximation for the probability and timeline to classify a VUS can guide functional assay developers as to which variants they should include in order to maximize the impact of their efforts. Sharing clinical data across institutions is critical to human health outcomes.

9 Description of Supplemental Data

The supplemental data contain other experiments we ran using the model. Specifically, we include in the supplement results from a 20-year simulation, results from a different combination of centers, and results for a one-in-a-million variant.

10 Declaration of interests

The authors declare no competing interests.

11 Acknowledgements

The authors are very grateful to David Goldgar and Sean Tavtigian for providing their expert insight in the approximation of ACMG evidence criteria observation frequencies.

12 Data and code availability

The Python software we wrote to run these simulations is available on github at `https://github.com/jcasalet/data-sharing`.

13 References

1. Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, D., Hegde, M., Lyons, E., Spector, et al. (2015).Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Gentics in Medicine. Volume 70, Issue 5, 405-423.

2. Staff. (2020). Prevention and treatment of cardiomyopathy, https://www.heart.org/en/health-topics/cardiomyopathy/prevention-and-treatment-of-cardiomyopathy ; last accessed April 15, 2021.

3. Rehm, H., Berg, J., Brooks, L., Bustamante, C., Evans, J., Landrum, M., Ledbetter, D., Maglott, D., Martin, C., Nussbaum, R., et al. ClinGen - The clinical Genome Resource. New England Journal of Medicine Volume 372, Issue 23, 2235-2242

4. Tavtigian, S., Greenblatt, M., Harrison, S., Nussbaum, R., Prabhu, S., Boucher, K., Biesecker, L. (2018). Modeling the ACMG/AMP variant classification guidelines as a Bayesian classification framework. Genetics in Medicine. Volume 20, Issue 9, 1054-1060.

5. Bayes, T. 1764. "An Essay Toward Solving a Problem in the Doctrine of Chances", Philosophical Transactions of the Royal Society of London 53, 370-418. [https://web.archive.org/web/20190126111142/http://www.stat.ucla.edu/history/essay.pdf]

6. Russell, Stuart; Norvig, Peter (2003) [1995]. Artificial Intelligence: A Modern Approach (2nd ed.). Prentice Hall. ISBN 978-0137903955.

7. Berg, J. (2017). Exploring the importance of case-level clinical information for variant interpretation. Genetics in Medicine. Volume 19, Issue 1, 3-5.

8. Shirts, B., Jacobson, A., Jarvik, G. Browning, B. (2014). Large numbers of individuals are required to classify and define risk for rare variants in known cancer risk genes. Genetics in Medicine. Volume 16, Issue 7, 529-534.

9. Besenbacher, S., Liu, S., Izarzugaza, J., Grove, J., Belling, K., Bork-Jensen, J., Huang, S., Als, T., Li, S., Yadav, R., et al. (2015). Novel variation and de novo mutation rates in population-wide de novo assembled Danish trios. Nature Communications, Volume 6, Issue 1, 5969

10. Tsai, G., Ranola, J, Smit, C., Garrett, L., Bergquist, T., Casadei, S., Bowen, D., Shirts, B. (2019). Outcomes of 92 patient-driven family studies for recalssification of variants of uncerstain significance. Genetics in Medicine. Volume 21, Issue 6, 1435-1442

11. Hampel, H., Pearlman, R., Beightol, M., Zhao, W., Jones, D., Frankel, W., goodfellow, P., Yilmaz, A., Miller, K., BAcher, J. (2018). Assessment of tumor sequencing as a replcaement for Lynch Syndrome screening and current molecular tests for patients with colorectal cancer. JAMA oncology, Volume 4, Issue 6, 806-813

12. Shirts, B., Pritchard, C., Walsh, T. (2016). Family-specific variants and the limits of human genetics. Trends in Molecular Medicine, Volume 22, Issue 11 925-934.

13. Susswein, L., Marshall, M., Nusbaum, R., Vogel , K., Weissman, S., Yackowski, L., Vaccari, E., Bissonnette, J., Booker, J., Cremona, M., et al. (2016). Pathogenic and likely pathogenic variant prevalence among the first 10,000 patients referred for next-generation cancer panel testing. Genetics in Medicine, Volume 18, Issue 8, 823-832

14. Mannan, A., Singh, J., Lakshmike, R., Thota, N., Singh, S., Sowmya, T.S., Mishra, A., Sinha, A., Deshwal, S. Soni, M., et al. (2016). Detection of high frequency of mutations in a breast and/or ovarian cancer cohort implications of embracing a multi-gene panel in molecular diagnosis in India. Journal of Human Genetics, Volume 61, Issue 6, 515-522

15. Santos, C., Peixoto, A., Rocha, P., Bizarro, S., Pinheiro, M., Pinto, C., Henrique, R., Teixeira, M. (2014). Pathogenicity evaluation of BRCA1 and BRCA2 unclassified variants identified in Portuguese breast/ovarian cancer families. The Journal of Molecular Diagnostics, Volume 16, Issue 3, 324-334.

16. Rebbeck TR, Friebel T, Lynch HT, et al. Bilateral prophylactic mastectomy reduces breast cancer risk in BRCA1 and BRCA2 mutation carriers: The PROSE Study Group. Journal of Clinical Oncology 2004; 22(6):1055-1062.

17. Harrison, Steven M., Jill S. Dolinksy, Wenjie Chen, Christin D. Collins, Soma Das, Joshua L. Deignan, Kathryn B. Garber, et al. "Scaling Resolution of Variant Classification Differences in ClinVar between 41 Clinical Laboratories through an Outlier Approach." Human Mutation 39, no. 11 (November 2018): 1641–49.

18. Cline, Melissa S., Rachel G. Liao, Michael T. Parsons, Benedict Paten, Faisal Alquaddoomi, Antonis Antoniou, Samantha Baxter, et al. "BRCA Challenge: BRCA Exchange as a Global Resource for Variants in BRCA1 and BRCA2." Edited by Charis Eng. PLOS Genetics 14, no. 12 (December 26, 2018): e1007752.

19. Harrison, Steven M., Jill S. Dolinsky, Amy E. Knight Johnson, Tina Pesaran, Danielle R. Azzariti, Sherri Bale, Elizabeth C. Chao, Soma Das, Lisa Vincent, and Heidi L. Rehm. "Clinical Laboratories Collaborate to Resolve Differences in Variant Interpretations Submitted to ClinVar." Genetics in Medicine 19, no. 10 (October 2017): 1096–1104.

14 Supplementary Information

In this section, we show the results of simulations for various combinations of input parameters that are outside the main scope of the paper. First we'll show the results from a 20-year simulation using the same mix of sequencing centers. Next we'll examine results from a simulation using a very small combination of sequencing centers (1 large, 3 medium, and 5 small). We'll look at the results of a simulation using a 1e-6 frequency VUS. And we conclude by examining the results of a simulation using a 1e-7 frequency VUS.

Note that throughout this supplement, we provide the plots only for one of the large sequencing centers in addition to the plots for the aggregated center. By this time, we've adequately demonstrated what the contributions of small and medium-sized sequencing centers look like by way of their histogram, scatter, and probability plots.

14.1 Results from 20-year simulations

Below, we show the histogram, scatter, and probability plots using the same mix of sequencing centers (10 small, 7 medium, and 3 large), but over the course of 20 years rather than 5 years.

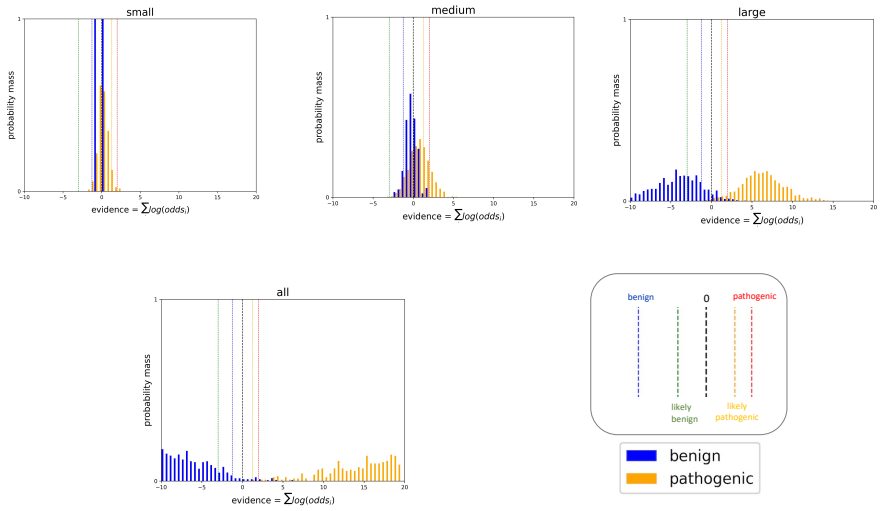

Figure 6: Distributions of accumulated evidence over 20 years

The histogram plots of the 20-year simulation in Figure 6 show the same trend: the distributions of evidence are "skinny and tall" for smaller collections of data and "wide and short" for larger collections of data. In fact, the vast majority of evidence in the aggregated center are outside the extreme thresholds required for classifying as benign or pathogenic and even exceed the horizontal boundaries of the plot itself.

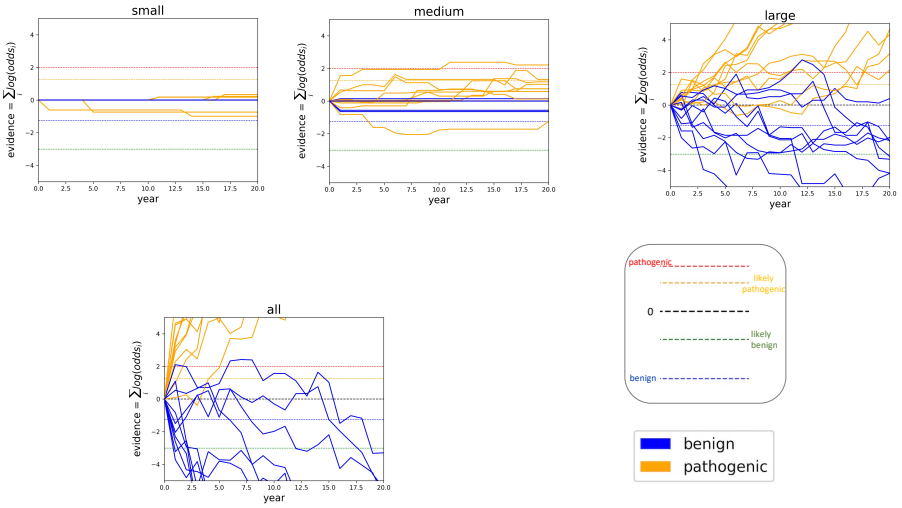

Figure 7: Trajectory of accumulated evidence over 20 years

The vast majority of pathogenic variants are classified within 7 years, and all of them appear classified after about 12 years. While there is conflicting evidence for benign variants after 20 years of collecting data at the large center, all the evidence for benign variants in the aggregated center is less than 0.

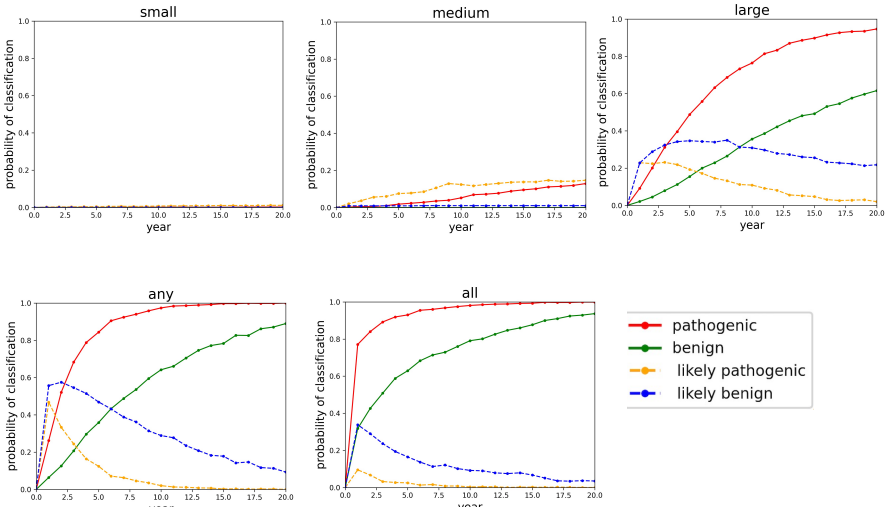

Figure 8: Probability of classification over 20 years

The probability plots of Figure 8 show that variants have only a 40% chance of being classified as benign after 20 years for the large sequencing center. The aggregated center, however, almost doubles that probability for benign variants.

Moreover, the probability of classifying a variant as pathogenic is nearly 1.0 after just 12 years.

14.2 Results from a combination of 1 large center, 3 medium centers, and 5 small centers

Below we show the results from simulations using a different combination of center sizes to demonstrate the value of clinical and variant data sharing even at a smaller scale. The other parameters of the model are the same as what was used in the simulations described in the main manuscript: 1e-5 variant frequency, 5-year time period, and 1,000 simulated variants.

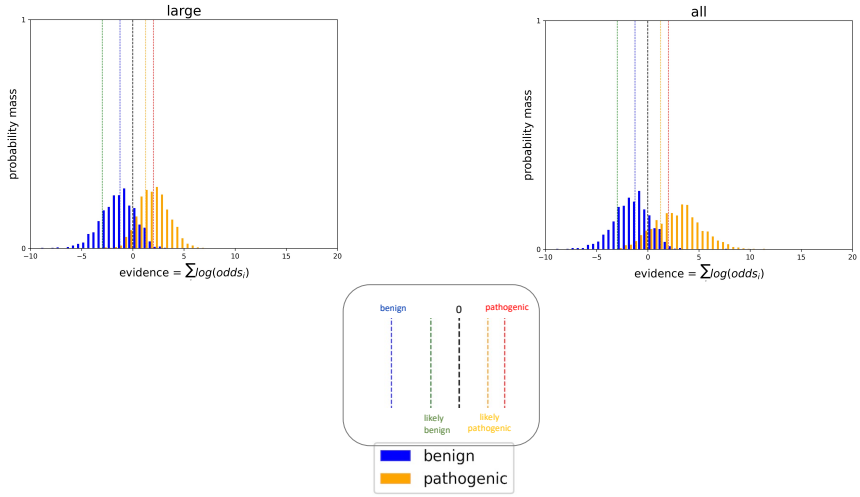

Figure 9: Distribution of accumulated evidence over 5 years at sequencing centers

The histogram plots of Figure 9 of this smaller combination of sequencing centers points out that with just one large sequencing center, there's not much difference in the distributions of that large center and the aggregated center.

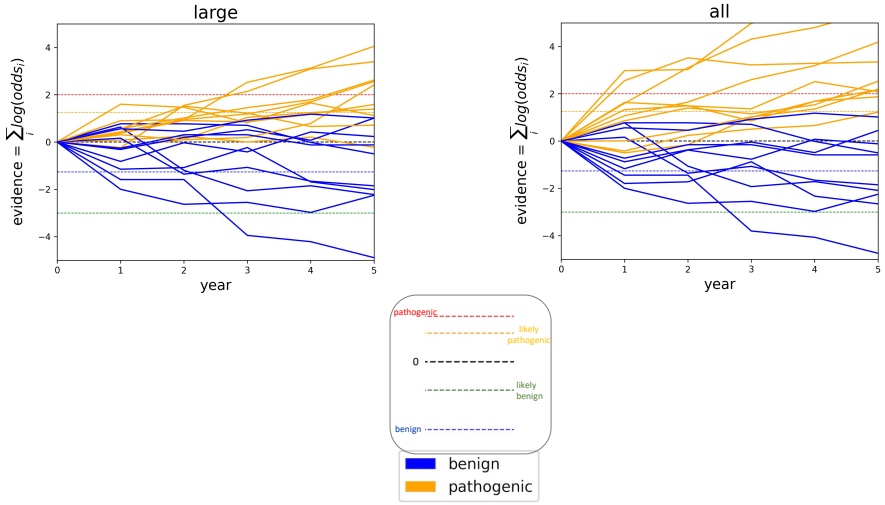

Figure 10: Trajectory of accumulated evidence over 5 years at sequencing centers

The trajectory plots of Figure 10 show that while there's not much difference in the evidence gathered for benign variants, there is a recognizable difference in the evidence gathered for pathogenic variants.

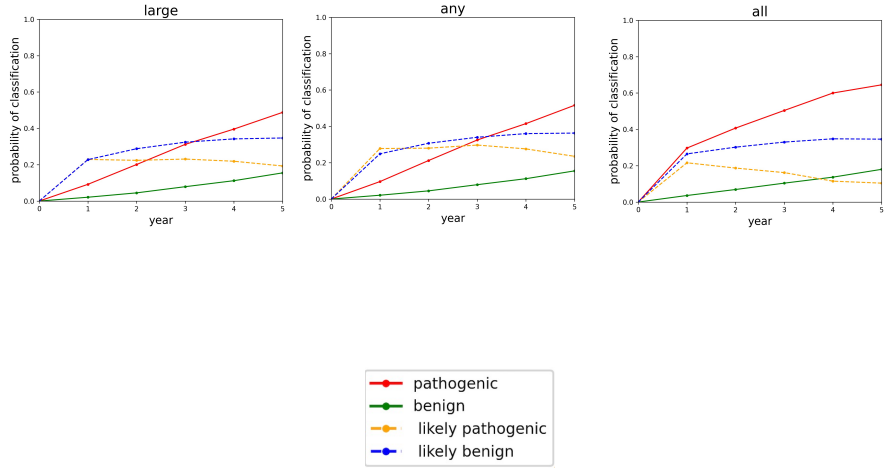

Figure 11: Probability of classification over 5 years at sequencing centers

The probability plots of Figure 11 show that there's not much difference between one large center and any center in the probability of classifying a variant. When all centers share their clinical data, there's a remarked improvement in the probability of classifying pathogenic variants.

14.3 Results for a one-in-a-million variant

Below we show the results from simulations for a variant with frequency 1e-6.

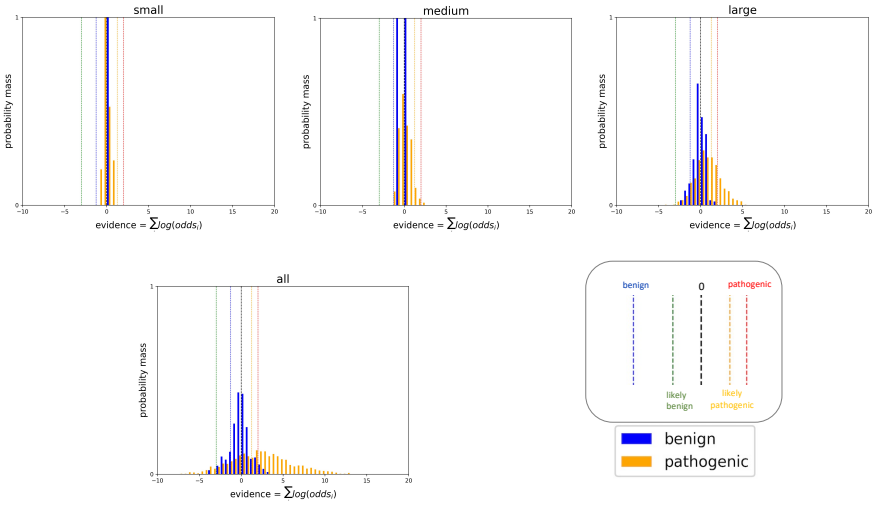

Figure 12: Distribution of accumulated evidence over 5 years

The histogram plots of Figure 12 show that all the evidence follows a "skinny and tall" distribution, even in the aggregrated center.

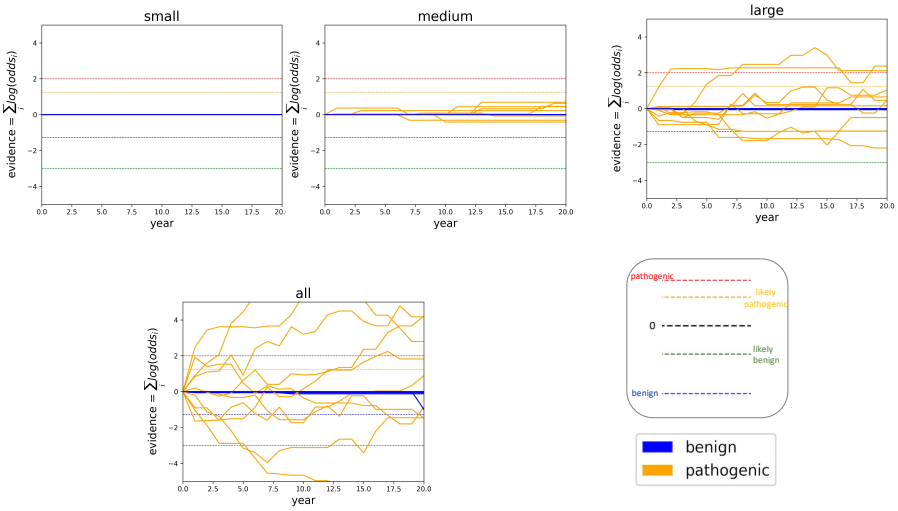

Figure 13: Trajectory of accumulated evidence over 5 years

The trajectory plots of Figure 13 show that most of evidence for pathogenic

variants is insufficient and often contradictory to classification.

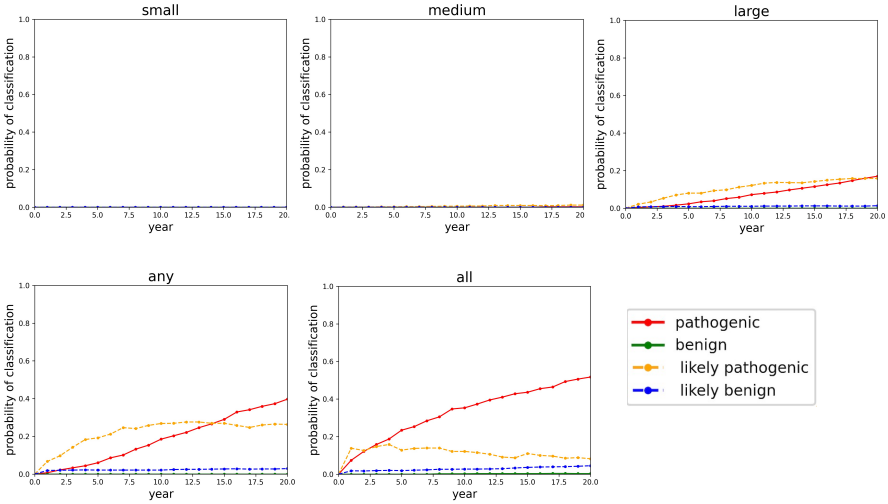

Figure 14: Probability of classification over 5 years

The probability plots of Figure 14 show that, in the best case of the aggregated center, pathogenic variants have almost a 60% chance of being classified. The probability of classifying benign or likely benign is negligible, even when sharing data.