Identification of Misclassified ClinVar Variants via Disease Population Prevalence

Naisha Shah,[1] Ying-Chen Claire Hou,[1] Hung-Chun Yu,[1] Rachana Sainger,[1] C. Thomas Caskey,[2] J. Craig Venter,[1,3,*] and Amalio Telenti[3,*]

There is a significant interest in the standardized classification of human genetic variants. We used whole-genome sequence data from 10,495 unrelated individuals to contrast population frequency of pathogenic variants to the expected population prevalence of the disease. Analyses included the ACMG-recommended 59 gene-condition sets for incidental findings and 463 genes associated with 265 OrphaNet conditions. A total of 25,505 variants were used to identify patterns of inflation (i.e., excess genetic risk and misclassification). Inflation increases as the level of evidence supporting the pathogenic nature of the variant decreases. We observed up to 11.5% of genetic disorders with inflation in pathogenic variant sets and up to 92.3% for the variant set with conflicting interpretations. This improved to 7.7% and 57.7%, respectively, after filtering for disease-specific allele frequency. The patterns of inflation were replicated using public data from more than 138,000 genomes. The burden of rare variants was a main contributing factor of the observed inflation, indicating collective misclassified rare variants. We also analyzed the dynamics of re-classification of variant pathogenicity in ClinVar over time, which indicates progressive improvement in variant classification. The study shows that databases include a significant proportion of wrongly ascertained variants; however, it underscores the critical role of ClinVar to contrast claims and foster validation across submitters.

Introduction

Currently, more than 68,000 clinical genetic tests are offered from more than 1,700 clinics and laboratories according to Genetests. While genetic testing is a powerful diagnostic tool, there are several challenges for interpretation and reporting of findings. Some of these challenges include the accuracy of variant calling, identification of pathogenic variants, and interpretation of low-penetrant variants.[1] Variant selection algorithms have been proposed to avoid false positive genomic screening results in the general population.[2] In addition, different laboratories have developed different protocols to handle the challenges, which leads to inconsistencies in the classification of variants and a bias toward overestimating pathogenicity.[3]

For many variants, the assignment of clinical significance reflects historical guidelines and evidence available at the time of the original interpretation. However, as additional information becomes available, the interpretation of pathogenicity of genetic variants may change. Data-sharing efforts have shown that 17% of the variants with clinical interpretations submitted by more than one laboratory had conflicting interpretations.[4] The American College of Medical Genetic and Genomics (ACMG) and the Association for Molecular Pathology (AMP) issued guidelines to support a standardized approach to variant classification.[5] Initiatives to curate existing knowledge and improve variant interpretation have been put in place. The NIH-based partnership between ClinVar and ClinGen is an example of such an initiative.[4,6] ClinVar implements a ranking system to denote the quality associated with each submission to the database. For example, a three- or four-star submission comes from "expert panel" and "practice guidelines" submitters, which are the most ClinGen-trusted sources for variant interpretation. Challenges lie ahead, as even the implementation of ACMG-AMP guidelines led to only a 34% concordance on the reporting of 99 variants across laboratories. After consensus discussions and detailed review of the ACMG-AMP criteria, concordance increased to 71%.[3]

Leveraging knowledge from shared data by categorizing variants based on clinical significance, the number of submitters, and their assertion criteria is an important step toward accurate interpretation of variant pathogenicity and diagnosis. However, there is a need for additional methods to detect misclassified pathogenic variants. Also, it is crucial to identify variants with low clinical penetrance (i.e., proportion of individuals with a variant that develop the disease or clinical symptoms) for proper clinical reporting. For example, the variant c.845G>A (rs1800562; p.Cys282Tyr) in *HFE* (MIM: 613609) associated with hereditary hemochromatosis was thought to be the main pathogenic variant;[7] however, as more individuals were genotyped, the variant's high population frequency appeared more compatible with low penetrance.[8]

Here, we revisit the topic of assignment of pathogenicity to a variant by assessing expected disease prevalence and observed genetic risk in a population. Genetic risk is defined here as the number of individuals that are at disease risk based on the presence of pathogenic variants

[1]Human Longevity Inc., San Diego, CA 92121, USA; [2]Baylor College of Medicine, Houston, TX 77030, USA; [3]J. Craig Venter Institute, La Jolla, CA 92037, USA
*Correspondence: jcventer@jcvi.org (J.C.V.), atelenti@jcvi.org (A.T.)
https://doi.org/10.1016/j.ajhg.2018.02.019.

and mode of inheritance. We used recent data from whole-genome sequencing of 10,495 unrelated subjects (Telenti et al.[9]) to identify individuals with clinically significant variants from ClinVar. We identify disease conditions with inflated values, i.e., the cumulative frequency of the disease risk variants in the population far exceeds the expected prevalence of disease. Here, we jointly considered several rare "clinically significant" variants to identify inflation for diseases rather than identifying individual misclassified variants based on their allele frequency. If the genetic basis and etiology of a disease is not well understood (incomplete knowledge on associated genes, variants, and other factors), the currently known variants will explain only some of the disease prevalence, i.e., they will appear deflated.

We performed the disease prevalence analysis as a function of pathogenic variants listed in ClinVar separately for the well-curated 59 genes and associated conditions recommended by the ACMG for reporting of secondary findings (SF v2.0) in clinical genome-scale sequencing[10–12] (herein referred to as ACMG-59) and for genes with available population prevalence information reported in OrphaNet/OrphaData, a data source on rare diseases.[13] We then replicate our findings of inflation in a large public dataset, gnomAD,[14] with more than 138,000 exomes and whole genomes.

Material and Methods

We used whole-genome sequences from 10,495 unrelated individuals sequenced at a 30× median coverage. Details are described by Telenti et al.:[9] participants were representative of major human populations and ancestries, and the study population was not ascertained for a specific health status. To avoid analyzing potentially inaccurate variant calls that lie within the areas of the genome prone to sequencing errors,[15] the analysis was focused on variants that fell within the high-confidence sequencing regions of the genome.[9,16,17] For the disease prevalence analysis, we calculated frequencies of individuals at genetic risk using variants deposited in ClinVar.[18] We performed disease prevalence analysis for two groups of conditions: (1) a well-curated list of the recommended ACMG 59 gene and associated conditions to report for incidental findings (referred to here as "ACMG-59")[10,12] and (2) a list of rare conditions collected by a consortium in a reference portal called OrphaNet/OrphaData.

To compare observed genetic risk and expected population prevalence, we used only the conditions with two or more at-risk individuals observed in the study. We calculated fold-change for genetic risk compared to population prevalence per condition using the formula: observed/expected. SnpEff[19] was used to annotate effect of the variants in ACMG-59 and OrphaNet genes using canonical transcripts.

To replicate our findings, we used gnomAD exomes (n = 123,136) and genomes (n = 15,496) datasets (v.2.0.1.).[14] Since the datasets do not have sample-level information available, we calculated the number of individuals at risk for a condition to be the number of alleles observed for all pathogenic/likely pathogenic variants (i.e., assuming independent samples).

ACMG-59 Conditions

For each of the ACMG-59 conditions, we searched OrphaNet, GeneReview, and other published sources for the available population prevalence of disease. In case of multiple prevalence information available, we chose the maximum prevalence for the purpose of the study.

OrphaNet Conditions

We used the OrphaNet v.1.2.4 for this analysis. We selected OrphaNet conditions that had at least one associated ClinVar variant and had a defined mode of inheritance and population prevalence information available (i.e., prevalence type of "Point prevalence," "Lifetime prevalence," and "Prevalence at birth"). Only the conditions that had the following mode of inheritances were considered: autosomal dominant, autosomal recessive, X-linked recessive, or X-linked dominant. In case of multiple population prevalence information available, we chose the maximum prevalence.

ClinVar Variants

We used the newly available ClinVar VCF_2.0 file (version: clinvar_20170905.vcf.gz; GRCh38 reference) to obtain the disease-associated variants including single-nucleotide variant and indels. We filtered out variants that were considered "included" variants; i.e., variants that were interpreted as part of a set of variants such as a haplotype, and somatic variants. Following the ACMG guidelines for clinical interpretation, we removed variants with greater than 5% allele frequency in any ethnic populations except in Ashkenazi Jewish population due to founder effect. For variant allele frequency, we used both our database and the Genome Aggregation Database (gnomAD; genome and exome datasets).[14] For gnomAD datasets, we filtered out variants without PASS calls.

We divided variants deposited in ClinVar using ClinVar/ClinGen's ranking system[4] and its definition of variant classification into four sets. Set 1 included pathogenic and likely pathogenic (P/LP) variants with ClinVar star 2+ (i.e., multiple submitters with assertion criteria, expert panel or practice guideline). Set 2 included P/LP variants with ClinVar star 1 (i.e., one submitter with assertion criteria). Set 3 included P/LP variants with ClinVar star 0 (i.e., submitter without assertion criteria). Set 4 included variants with conflicting clinical significance with assertion criteria provided. We used clinical significance values from CLNSIG to group "Pathogenic," "Likely pathogenic," and "Pathogenic/Likely pathogenic" as P/LP variants and to select set 4 variants with "Conflicting interpretations of pathogenicity." For categorizing variants with ClinVar star 0 to 4, we used values in CLNREVSTAT. Values "criteria_provided,_multiple_submitters,_no_conflicts," "reviewed_by_expert_panel," or "practice_guideline" were grouped as star 2+, "criteria_provided,_single_submitter" as star 1, and "no_assertion_criteria_provided" as star 0. Each of the sets of variants was used separately to perform the disease prevalence analysis for ACMG-59 and OrphaNet conditions.

Disease-Specific Minor Allele Frequency Threshold

To identify and remove potentially benign variants from the sets above, we applied disease-specific minor allele frequency (dMAF) threshold per condition. We defined the dMAF threshold as follows. (1) For autosomal/X-linked-dominant conditions, assuming there is one highly penetrant variant causing 100% of

the disease cases, then the frequency of heterozygous should not be greater than the disease prevalence. Thus, the AF for the variant should not exceed 1/2*(disease prevalence). (2) For autosomal/X-linked-recessive conditions, assuming there is one highly penetrant variant causing 100% of the disease cases, then the frequency of homozygous recessive should not be greater than the disease prevalence. Thus, the AF for the variant should not exceed the square root of the disease prevalence.

To account for penetrance, these formulas can be generalized as follows. (1) For dominant conditions, dMAF threshold = 1/2*(disease prevalence)*(1/penetrance). (2) For recessive conditions, dMAF threshold = sqrt(disease prevalence*(1/penetrance)).

However, for the study, since one of our goals is to highlight conditions with inflated genetic risk compared to disease population prevalence, we use a more stringent threshold assuming 100% penetrance. To avoid by chance occurrences, we applied the threshold to variants observed in more than one individual in the study.

Change in ClinVar Variant Classification

To investigate how ClinVar variant classification changed over time, we compared September 2017 (the version used in our analysis) to the May 2016 version. For this, we used ClinVar XML files (ClinVarFullRelease_2017-09.xml.gz and ClinVarFullRelease_2016-05.xml.gz) instead of VCF_2.0 version, which was available only for the September 2017 version. We observed an artifact in the XML file where ClinVar's clinical significance for some records ("RCV") was mislabeled "conflicting interpretations of pathogenicity" (e.g., RCV000036715.4). To avoid this, using ClinVar's guidelines, we aggregate clinical significance from all records per variant by extracting each submitter's clinical significance.

We selected variants that were common between the two versions and that had clinical significance terms recommended by ACMG (i.e., pathogenic, likely pathogenic, benign, likely benign, and VUS). We grouped together variants that had a clinical significance of pathogenic and/or likely pathogenic as P/LP. Similarly, we grouped together variants with clinical significance of benign and/or likely benign as B/LB.

Results

For genetic screening, we used disease-associated variants deposited in ClinVar.[4] We divided the variants into four sets based on ClinGen's ranking system using clinical significance and review stars (see Material and Methods for more details). Set 1 included 9,638 pathogenic and likely pathogenic (P/LP) variants with ClinVar star 2+ (i.e., multiple submitters with assertion criteria, expert panel, or practice guideline). Set 2 included 26,873 P/LP variants with ClinVar star 1 (i.e., one submitter with assertion criteria). Set 3 included 18,978 P/LP variants with ClinVar star 0 (i.e., submitter without assertion criteria). Set 4 included 11,529 variants with conflicting clinical significance. In total, our study would consider 67,018 variants described in ClinVar (Table S1).

We performed genetic screening of 10,495 unrelated individuals, whose whole genomes were sequenced with a mean coverage of 30×. Below we present the analyses for ACMG-59 and for OrphaNet genes and conditions.

ACMG-59

Twenty-six sets of medical conditions and 59 genes are represented in ACMG-59. There are 16,781 ClinVar variants associated with the ACMG-59 genes ($n_{set\text{-}1}$ = 6,415; $n_{set\text{-}2}$ = 5,860; $n_{set\text{-}3}$ = 1,683; and $n_{set\text{-}4}$ = 2,823). All of the 26 sets of medical conditions had at least 1 variant in set 1, which is the most reliable and agreed-upon classified variant set according to the ClinGen/ClinVar ranking.

In the study population, we observed 22,454 variants with allele frequency less than 5% in the coding regions of the 59 ACMG genes, including 7,162 missense and 245 loss-of-function (LoF) variants. Of these, 1,796 variants matched ClinVar records: $n_{set\text{-}1}$ = 139, $n_{set\text{-}2}$ = 107, $n_{set\text{-}3}$ = 17, and $n_{set\text{-}4}$ = 1,533. Thus, in the present study, we observed 10.7% of the 16,781 ClinVar variants associated with ACMG-59. Screening for the P/LP variants from set 1, set 2, and set 3, we observed that 2.6% of the individuals would be predicted at risk for disease (herein referred to as "genetic risk") for 16 of the 26 ACMG-59 conditions, and 4.9% of the individuals were carriers for 17 of the 26 ACMG-59 conditions. This is within the estimated range (1.5%–6.5%) of screened individuals that would have an incidental finding for the ACMG-56.[20] Three individuals (0.03%) in the study were at genetic risk for two ACMG-59 conditions.

We wanted to investigate whether the variant ranking (here, broken down by variant sets) is indicative of misclassified variants. Thus, using variants from each set separately, we compared the observed genetic risk to the reported population prevalence for the conditions (Figure 1A). We would expect, by using subsets of variants, that the observed genetic risk would be lower-bound if the variants were truly pathogenic and with high penetrance. We observed inflated (heuristically defined as more than 10-fold increase; see Material and Methods) genetic risk for several conditions using variants from set 1, set 2, and set 3. This may indicate that some of the variants have either low penetrance or inaccurate pathogenicity assignment. As we sequentially added ranked sets of variants to calculate genetic risk, the fold change of observed genetic risk compared to population prevalence gradually increases (Figure 1B). This suggests that with the addition of more variants with lower ranks, more misclassified variants and/or variants with low penetrance accumulate and contribute to the inflation.

We found three conditions (11.5% of the 26 ACMG-59 conditions) with more than 10-fold increase (i.e., inflated) when using P/LP variants from set 1, set 2, and set 3 (Table S2). These conditions included malignant hyperthermia susceptibility (MIM: 145600), multiple endocrine neoplasia type 1 (MIM: 131100), and hereditary paraganglioma-pheochromocytoma syndrome (MIM: 168000 [PGL1], 601650 [PGL2], 605373 [PGL3], 115310 [PGL4]). All three conditions were also inflated using only the

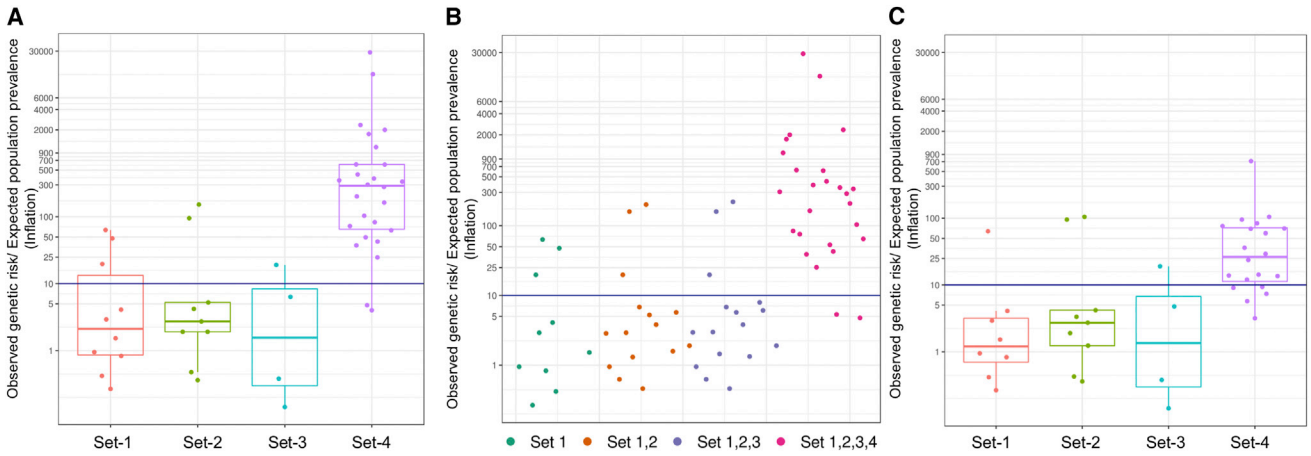


Figure 1. Genetic Risk in ACMG-59 Conditions

Fold-change of observed genetic risk over expected population prevalence using ClinVar variant sets for the ACMG-59 conditions. Each point represents a condition; each condition may be represented in more than one set. The navy blue line at a fold-change of 10 (i.e., inflation) indicates a theoretical penetrance of 10%. Observations above this line are highly suggestive of misclassified variants. The boxplot shows median (horizontal line in the box), first and third quartile (lower and upper hinges of the box, respectively). The upper whisker extends from the hinge to the largest value no further than 1.5 * inter-quartile range (IQR) from the hinge. The lower whisker extends from the hinge to the smallest value at most 1.5 * IQR of the hinge.

(A) Fold-change was calculated using variants per variant set: set 1 consists of variants with 2 or more ClinVar review stars (i.e., two or more submitters with assertion criteria, expert panel, and practice guideline); set 2 consists of variants with 1 star (i.e., one submitter with assertion criteria); set 3 consists of variants with 0 star (i.e., submitter with no assertion criteria submitted in ClinVar); set 4 consists of variants with conflicting interpretations of pathogenicity.

(B) Fold-change was calculated by using variants cumulatively from each set; i.e., set 2 includes set 1 variants, set 3 includes set 1 and 2 variants, set 4 includes all variants.

(C) Fold-change was re-calculated after variants were filtered for disease-specific minor allele frequency thresholds.

variants in set 1 (i.e., a concordant variant interpretation by two or more submitters with assertion criteria) (Figure 1A). Below we discuss several of the conditions.

Variants with conflicting interpretations from set 4 have an observed genetic risk that inflates massively (Figure 1A). 24 of the 26 (92.3%) ACMG-59 conditions were inflated. This is a strong indication of misclassified variants in the set of conflicting interpretation, in particular as inflation goes far beyond what could be assumed to reflect low penetrance. It has been suggested that ClinVar's mis-interpretation of some OMIM entries as "pathogenic" might be a source of conflict in set 4. However, we observed that only 4% of the set 4 variants have OMIM as a submitter, i.e., a majority of the conflict comes from multiple submitters with different interpretations of pathogenicity. There may be a few exceptional variants within the set that may be pathogenic; however, without supporting data they should be removed from consideration.

A recommended criterion to identify benign variants beyond ACMG-AMP's criteria of greater than 5% allele frequency (AF) threshold is to develop disease-specific minor allele frequency (herein referred to as dMAF) thresholds.[3] Several such dMAF methods have been proposed.[21–23] We compare our approach with a recent AF filtering framework by Whiffin et al.[21] The framework uses disease prevalence, penetrance, inheritance pattern, and maximum allelic contribution. The latter is a measure of allelic heterogeneity that is derived from the most common causative variant known in the literature for the specific disease. The variant frequency in disease cases is often derived from small sample size and thus is susceptible to biased estimation (e.g., frequency for most common causative variant in *PKP2* is estimated from only 361 cases). In addition, for most diseases, neither the allelic heterogeneity nor maximum genetic contribution is well characterized.

To compare the methods, we used diseases reported in Table 1 in Whiffin et al.,[21] which the authors used to test their framework (referred here as "W-framework"). These included hypertrophic cardiomyopathy/dilated cardiomyopathy, arrhythmogenic right ventricular cardiomyopathy (ARVC), and Romano-Ward long QT syndromes types 1, 2, and 3, Brugada Syndrome (LQTS/Brugada). Using the framework, for each of the diseases, we predicted maximum allelic contribution (95% CI); however, to maintain our conservative approach, we kept our penetrance at 100% unlike the assumed 50% penetrance used in the W-framework.

The W-framework flagged all the variants that our dMAF approach flags as potentially false positive variants. In addition, the W-framework flagged 217 variants including 15 variants from set 1 ("multiple submitters"), 35 from set 2, 4 from set 3, and 164 from set 4. One of the set 1 variants was rs373746463 in *MYBPC3* associated with hypertrophic cardiomyopathy/dilated cardiomyopathy. The predicted maximum allelic contribution for gnomAD Exome was 2; however, 4 were observed in gnomAD Exome. Due to the higher observed frequency than predicted, the W-framework filters out this variant, indicated

it to be a benign variant. However, there is a strong suggestion for the variant rs373746463 to be P/LP and is interpreted as such by 7 different submitters in ClinVar. We observed similar level of pathogenic confidence (multiple submitters) for the other 14 variants. These included rs267607554, rs794728583, rs397508118, rs12720458, rs397508097, rs199472815, rs397516005, rs757532106, rs137854604, rs199473284, rs199473283, rs199473220, rs199473097, and rs139794067. This suggests that the Whiffin et al.[21] framework may be prone to removing potentially pathogenic variants. Thus, for the rest of the analysis, we used our dMAF method.

After filtering out variants using dMAF, we observed an overall decrease in inflation, especially using set 4 variants (Figure 1C). The inflation dropped from 24 conditions to 15 conditions (57.7%), thereby confirming that a large proportion of the set 4 variants are benign. However, the genetic risks of two conditions (hereditary paraganglioma-pheochromocytoma syndrome and malignant hyperthermia susceptibility) showed inflation using P/LP variants (Table S2). Individually, variants for these disorders are rare; however, collectively, they add up to show an increase in the observed genetic risk compared to the estimated population prevalence. These conditions with inflated risks are discussed in detail below, and as a model case we also describe a well-studied condition: hereditary breast and ovarian cancer (MIM: 604370, 612555).

To further inquire the source of inflation, we divided the sets into sub-categories to distinguish likely pathogenic (LP; defined as greater than 90% certainty of variant being disease causing[5]) from pathogenic (P) variants. This deconvolutes two concepts: (1) confidence in assertion (via sets 1–4) and (2) likelihood of pathogenicity (P versus LP). We calculated inflation for each of the sets as above. We did not observe any indication of LP sets contributing more to inflation compared to P sets (Figures S1A and S1B).

Overall, of the 1,796 ClinVar variants from all sets (n = 263 in sets 1, 2, 3) observed in 10,495 individuals, using the dMAF filter, we removed 882 (49.1%) variants from all sets, most of which came from set 4 as expected (n = 870, 56.8%), while only 12 variants (4.6%) from sets 1, 2, and 3. Of the 26 ACMG conditions, the genetic risks for three conditions were inflated before the dMAF filter (sets 1, 2, 3), and for two conditions genetic risks were still inflated after the filter. For the critical sets (sets 1, 2, 3) only few variants appear responsible for the inflation.

Hereditary Breast and Ovarian Cancer (HBOC)

To test that the disease estimates for well-studied conditions are as expected, we studied the frequency of individuals at genetic risk for hereditary breast and ovarian cancer (HBOC). HBOC is estimated to have a disease prevalence of 0.2%–0.3% in the general population.[24] Variant classification for *BRCA1/2* (MIM: 113705, 600185) showed high concordance across seven established clinical testing laboratories.[25] Using P/LP variants, we observed 47 individuals with 39 variants ($n_{set\text{-}1} = 38$, $n_{set\text{-}3} = 1$) to be at genetic risk for HBOC (0.4% of the individuals); i.e., a frequency close to expected. The observed-over-expected fold change in genetic risk of less than 2% is mainly attributed to HBOC being one of the most well-studied conditions for germline mutations. An expert panel in ClinVar, known as ENIGMA (Evidence-based Network for the Interpretation of Germline Mutant Alleles), has submitted more than 6,000 variant interpretations for *BRCA1* and *BRCA2*. The expert panel comprises of international investigators focused on determining the clinical significance of sequence variants in breast cancer genes. This showcases the necessity to consolidate efforts around the world for correct annotation of variants.

Hereditary Pheochromocytoma-Paraganglioma

This is a rare condition characterized by growth of benign tumors in paraganglia and with an inheritance pattern of autosomal dominant. We observed five individuals in the study carrying five different variants ($n_{set\text{-}1} = 2$, $n_{set\text{-}2} = 3$) in the disease-associated *SDHB* (MIM: 185470) and *SDHD* (MIM: 602690) genes (observed genetic risk of 48 in 100,000 individuals). The population disease prevalence is not precisely known but the rate of incidence is approximately 0.3 in 100,000 individuals (Kirmani and Young in GeneReviews; see Web Resources). Disease penetrance for both the genes has the highest score (>40% chance) and clinical severity of 2 (i.e., reasonable possibility of death or major morbidity) according to ClinGen guidelines for actionability assessment for ACMG-59.[6] According to Benn et al.,[27] by the age of 40 years the penetrance for *SDHD* is 68%. Two variants (rs786202100, rs80338844) from set 1 had information available on its pathogenic effect on paraganglioma syndrome. However, the three variants from set 2 lacked sufficient information in ClinVar to determine pathogenicity. Reporting such variants that are not well recognized and do not have enough information available to clarify the association could lead to false reporting.

Malignant Hyperthermia Susceptibility

This is a disorder of skeletal muscle calcium regulation and is inherited in an autosomal-dominant manner. This disorder is triggered by reactions to anesthetics and thus the exact population prevalence is unknown. However, prevalence in individuals undergoing surgery in a New York hospital was estimated to be 1 in 100,000 for adults and 3 in 100,000 in children (Rosenberg et al. in GeneReviews; see Web Resources). Incidence range is estimated much higher to be 1 in 30,000 to 50,000 uses of anesthetics (Rosenberg et al. in GeneReviews; see Web Resources). Using P/LP variants in the two associated genes, *CACNA1S* (MIM: 114208) and *RYR1* (MIM: 180901), from set 1, set 2, and set 3 (after applying dMAF threshold), the observed genetic risk was inflated to 133 in 100,000 individuals. A total of 12 P/LP variants in only *RYR1* were observed in 14 individuals.

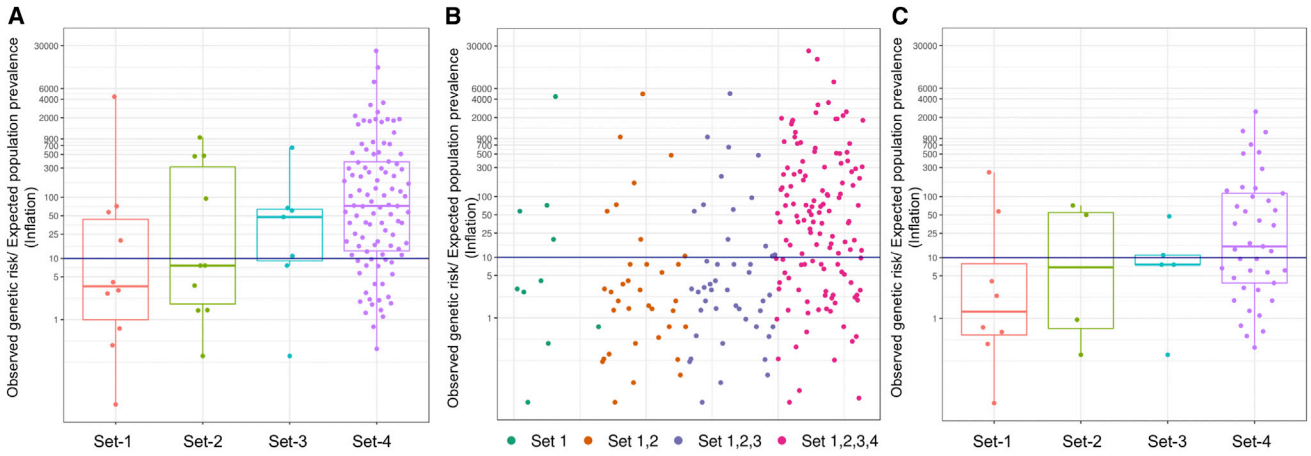


Figure 2. Genetic Risk in OrphaNet Conditions

Fold-change of observed genetic risk over expected population prevalence using variant sets from ClinVar for the OrphaNet conditions. Each point represents a condition; each condition may be represented in more than one set. The navy blue line at a fold-change of 10 (i.e., inflation) indicates a theoretical penetrance of 10%. Observations above this line are highly suggestive of misclassified variants. The boxplot shows median (horizontal line in the box), first and third quartile (lower and upper hinges of the box, respectively). The upper whisker extends from the hinge to the largest value no further than 1.5 * inter-quartile range (IQR) from the hinge. The lower whisker extends from the hinge to the smallest value at most 1.5 * IQR of the hinge.

(A) Fold-change was calculated using variants per variant set: set 1 consists of variants with 2 or more ClinVar review stars (i.e., two or more submitters with assertion criteria, expert panel, and practice guideline); set 2 consists of variants with 1 star (i.e., one submitter with assertion criteria); set 3 consists of variants with 0 star (i.e., submitter with no assertion criteria submitted in ClinVar); set 4 consists of variants with conflicting interpretations of pathogenicity.

(B) Fold-change was calculated by using variants cumulatively from each set; i.e., set 2 includes set 1 variants, set 3 includes set 1 and 2 variants, set 4 includes all variants.

(C) Fold-change was re-calculated after variants were filtered for disease-specific minor allele frequency thresholds.

RYR1 is also related to several distinct myopathies, including central core disease (MIM: 117000), Minicore myopathy with external ophthalmoplegia (MIM: 255320), King-Denborough syndrome (MIM: 145600), and neuromuscular disease (MIM: 117000) (Rosenberg et al. in GeneReviews; see Web Resources). Some of the myopathies are inherited in an autosomal-dominant manner and some are inherited in autosomal-recessive manner.[29] The confluence of several disorders, two modes of inheritance, and a disorder triggered by an exogenous exposure (anesthetics) leads to great imprecision in estimates of population prevalence and genetic risk. Studies such as Gonsalves et al.[30] and Merrit et al.[31] have identified misclassified and functionally validated pathogenic variants in *RYRI* and *CACNA1S*. However, there is a need for larger studies to estimate the disease prevalence and further assess pathogenicity of known and novel variants for malignant hyperthermia susceptibility.

OrphaNet Conditions

Next, we performed a similar prevalence analysis on OrphaNet conditions with at least one variant in any of the four sets, a stated mode of inheritance, and disease prevalence information. There are 12,997 ClinVar variants with allele frequency < 5% ($n_{set\text{-}1}$ = 2,778, $n_{set\text{-}2}$ = 4,454, $n_{set\text{-}3}$ = 3,249, and $n_{set\text{-}4}$ = 2,516) in 463 genes associated with 265 OrphaNet conditions.

In the study population, we observed 119,236 variants with allele frequency < 5% in the coding regions of the 463 genes, including 40,569 missense and 2,460 LoF variants. Overall, 2,830 were ClinVar variants associated with the OrphaNet conditions in the study population: $n_{set\text{-}1}$ = 448, $n_{set\text{-}2}$ = 391, $n_{set\text{-}3}$ = 199, and $n_{set\text{-}4}$ = 1,792. Thus, in the present study, we observed 21.8% of the 12,997 ClinVar variants associated with OrphaNet conditions. Screening for P/LP variants from set 1, set 2, and set 3, we observed 4.8% of the individuals with genetic risk for at least one of the 45 OrphaNet conditions, and 48.3% of the individuals were carriers for at least one of 168 OrphaNet conditions. Lazarin et al. identified 24% individuals of a large, ethnically diverse population as carriers for 108 Mendelian disorders.[32] While the present study is based on whole-genome sequencing and not limited to predefined sets of variants and disorders, the massive increase in carriers may alert of a significant number of misclassified variants.

As with the ACMG-59 conditions, we compared the observed genetic risk to the reported population prevalence for the OrphaNet conditions separately for each set (Figure 2A). We observed four conditions (Birt-Hogg-Dubé syndrome [MIM: 135150], central core disease [MIM: 117000], multiple endocrine neoplasia type 2 [MIM: 171400, 162300], and hereditary chronic pancreatitis [MIM: 167800]) with more than 10-fold higher observed genetic risk compared to the expected population prevalence using variants from set 1. For hereditary chronic pancreatitis (HCP), the study population had 17 set 1 variants from *CFTR* (MIM: 602421) which is known

to be associated with autosomal-dominant HCP and autosomal-recessive cystic fibrosis (MIM: 219700). The mode of inheritance for *CFTR* and its associated condition HCP is incorrectly assigned as autosomal dominant in OrphaNet. Similarly, central core disease, usually inherited in an autosomal-dominant manner, can also be inherited in an autosomal-recessive manner. Such discrepancies can cause the observed genetic risk to be inflated. Using set 2 and set 3 variants, we observed 7 conditions with inflated genetic risk. Using P/LP variants from the union of set 1, set 2, and set 3, in total, there were 9 conditions with inflated observed genetic risk (3.4% of the 256 conditions; Table S3). As observed with the ACMG-59 conditions, genetic risk using set 4 variants inflated massively for 82 OrphaNet conditions; strongly suggesting the inclusion of misclassified variants (30.9% of the 256 conditions; Figure 2B).

We used dMAF threshold to filter out additional potentially benign and low penetrant variants from the variant sets (see Material and Methods). We observed an overall decrease in inflation (Figure 2C). Applying the filter, we not only removed potentially benign variants but also filtered out variants due to inaccurate disease information, including imprecise prevalence estimations. For example, a common pathogenic variant, rs6025 (also known as R506Q; MAF = 0.02), in *F5* (MIM: 612309) is well known for its association with a common clotting disorder called factor V Leiden thrombophilia (MIM: 188055), where the blood clots more easily than normal. Other mutations in this gene are known to cause a different, rare condition called factor V deficiency (MIM: 227400) (prevalence of 1 in 1,000,000),[33] where the blood does not clot easily. In ClinVar, the rs6025 variant is incorrectly associated with factor V deficiency due to the gene's association with both conditions. The dMAF filters out such cases of true pathogenic variants but where they are wrongly associated with another condition. Another example is the successful removal of variants in *CFTR* that were associated with HCP.

Even though the dMAF filter removed most of the noise, there were four conditions with autosomal-dominant inheritance that had inflated genetic risk using P/LP variants collectively from set 1, set 2, and set 3 (1.5% of the 256 conditions; Table S3). Of the four conditions, three were the same as the inflated conditions observed using the set 1 variants, namely Birt-Hogg-Dubé syndrome, central core disease, and hereditary chronic pancreatitis. The fourth condition, pulmonary arterial hypertension (MIM: 178600), had a fold-change of 11.

As with ACMG-59 analysis, we sub-divided the sets to include only LP variants to inquire whether LP variants contributed to inflation more than P variants. We calculated inflation for each of the sets as above. We did not observe any indication of LP sets contributing more to inflation compared to P sets (Figures S2A and S2B).

Overall, of the 2,830 ClinVar variants from all sets (1,038 in sets 1, 2, 3) observed in 10,495 individuals, we removed 1,064 (38%) variants from all sets (41, 4% in sets 1, 2, 3) using the dMAF filter. Of the 265 OrphaNet conditions, the genetic risks for 24 conditions were inflated before the dMAF filter (sets 1, 2, 3) and for four conditions were still inflated after the filter. As was the case for ACMG-59 conditions, only a few variants appear responsible for the inflation in sets 1, 2, 3.

Replication in gnomAD

To validate our findings, we looked for the inflation patterns in ACMG-59 conditions in large publically available datasets, gnomAD exomes (n = 123,136), and genomes (n = 15,496).[14] We observed 3,476 ClinVar variants from all sets ($n_{set\text{-}1}$ = 743, $n_{set\text{-}2}$ = 480, $n_{set\text{-}3}$ = 67, and $n_{set\text{-}4}$ = 2,186) in the ACMG-59 genes that had at least 1 sample count in gnomAD. As above, we filtered variants (sets 1, 2, 3) using dMAF filter. We observed three conditions inflated (11.5% of the 26 ACMG-59 conditions; Figures S3A and S3B), two of which were also inflated using this study population. The conditions included hereditary paraganglioma-pheochromocytoma syndrome, malignant hyperthermia susceptibility, and *PTEN* hamartoma tumor syndrome (MIM: 153480).

In OrphaNet, we observed a similar inflation pattern using gnomAD exomes and genomes data. We observed 5,026 ClinVar variants from all sets ($n_{set\text{-}1}$ = 1,056, $n_{set\text{-}2}$ = 1,167, $n_{set\text{-}3}$ = 584, and $n_{set\text{-}4}$ = 2,219) in the OrphaNet genes that had at least 1 sample count in gnomAD. As above, we filtered variants (sets 1, 2, 3) using dMAF filter and looked for inflation. We observed 17 conditions inflated (6.6% of the 256 OrphaNet conditions; Figures S4A and S4B), which included all four conditions that were observed to be inflated using this study's population.

Changes of Classification of ClinVar Variants over Time

The classification of variants in ClinVar evolves over time as reflect of better understanding of pathogenicity, but also as a result of more population representative frequency data emerging through large sequencing efforts. Compared to the May 2016 version of ClinVar, the September 2017 version added 179,432 new variants (see Material and Methods). Of these, 23,712 were classified as pathogenic/likely pathogenic (P/LP), 60,888 were classified as benign/likely benign (B/LB), and the rest were classified as a variant of uncertain significance (VUS), of conflicting pathogenicity, or other non-ACMG terms of significance.

We observed discrepancy in "review status" for a subset of variants with conflicting pathogenicity status in ClinVar's XML file that reports at a record-level ("RCV"). For example, record RCV000036715.4 (rs376225470) has two submitters with clinical significance of "benign" and "likely benign" but was categorized as "conflicting interpretations of pathogenicity," which according to ClinVar guidelines should be "Benign/Likely benign." To overcome the issue, we used submitter level clinical significance for all records to aggregate clinical significance for each

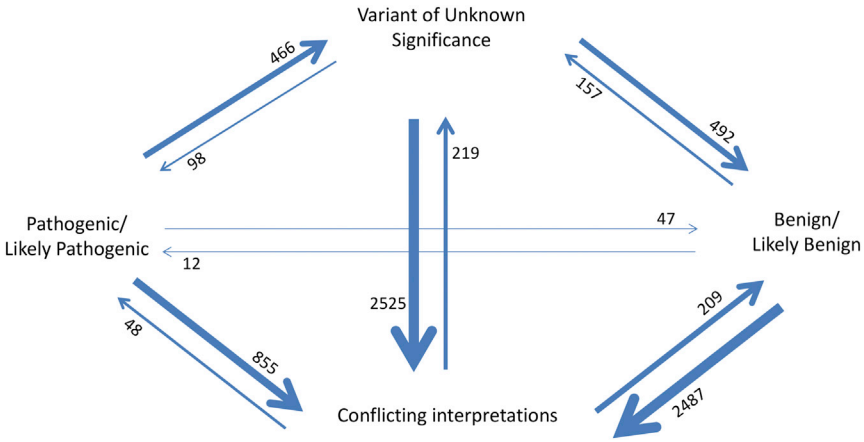


Figure 3. Change in ClinVar Variant Classification from May 2016 to September 2017

In the study period, 7,615 ClinVar variants changed classification. Predominantly, variants were reclassified to "conflicting interpretation" (n = 5,867; 77%). Only 158 variants (2%) were reclassified as pathogenic or likely pathogenic. Thickness of the arrows corresponds to the number of variants reclassified.

variant (see Material and Methods). The newly released ClinVar VCF file (available from September 2017) that reports at variant level by collapsing all records for the variant did not have the discrepancy. Thus, the discrepancy did not affect the above inflation analysis, which was performed using the VCF file.

To better understand the evolution of the assignment of pathogenicity classification to variants, we identified the directionality of the changes, i.e., whether a P/LP variant was reclassified as VUS, B/LB, or of conflicting pathogenicity. Between the two ClinVar versions, we observed that a majority of the variants (n = 93,101 of 100,716; 92.4%) did not change its clinical significance classification over the period of 16 months. For the 7,615 (7.6%) variants that were re-classified, we noted the directions of change (Figure 3).

Overall, most of the re-classification in ClinVar feeds into "conflicting interpretation," B/LB and VUS, and away from P/LP. The same period observed the reclassification of a significant number of P/PL variants to VUS or "conflicting interpretation." The trend of re-classification is expected as more knowledge is acquired and shared. For these variants, there was an increase in the number of submitters in the newer version of ClinVar (Figure S5).

Discussion

It is known that the current knowledge on pathological variants is far from complete.[34] There have been several efforts to identify misclassification of variants via data sharing (e.g., ClinVar/ClinGen)[35] and genetic analysis of large populations. Some examples of the studies are identification of false positive interpretation for cardiomyopathy using 7,855 cases;[36] re-evaluating previously identified casual variants for hypertrophic cardiomyopathy;[37] and observational study of P/LP variants implicated in Mendelian pediatric disorders[38] and epilepsy[39] using ExAC.[14] Such large genomic screening studies have assessed frequencies of variants in the general population[2,40] and used variant selection algorithms to identify misclassified variants that are relatively common. Here, we identify not only conditions with relatively common variants but also conditions that have inflated genetic risk (i.e., observed genetic risk is higher than population prevalence) when several potentially misclassified rare variants are considered jointly. One of the limitations of accurate estimations of inflation is lack of validated prevalence for all diseases. For example, we showed inflation (fold-change = 133) in malignant hyperthermia susceptibility, a disorder triggered by reactions to anesthetics. However, the available estimated population prevalence may be underestimated as it is measured only in individuals undergoing surgery. By design, we used curated public resources (OrphaNet, GeneReviews, and ClinGen) for disease prevalence information. This way, improvement in estimations can be globally tracked and can be used by other researchers. To address imprecision and reported biases of disease prevalence, we applied extremely conservative criteria to identify inflation at disease level as discussed below.

The inflation analysis uses three main concepts: pathogenicity of the variants, variant penetrance, and disease prevalence. Given limited understanding of the inherently interconnectedness between the three concepts, we used an extremely conservative approach to identify inflation of genetic risk for multiple diseases in the following ways. (1) At variant level, we used a stringent disease-specific minor allele frequency threshold by assuming 100% penetrance. In addition to false positive variants, this stringent approach may remove common potentially pathogenic variants with lower penetrance and thus underestimate the observed at-risk rate used in calculating inflation. (2) At disease level, we defined inflation only when the observed rate is higher than 10-fold of the reported prevalence. This enabled us to highlight only the diseases with a much higher rate of observed at-risk rate than disease prevalence; thereby allowing for inaccuracies in disease prevalence estimations. (3) At disease level, when there were two or more reported prevalence information available, to err on the side of caution, we chose the maximum prevalence. Using the approach, the overall inflation was observed in up to 11.5% of disorders using pathogenic/likely pathogenic variant sets (only using sets 1 to 3; not set 4 as the analysis strongly indicates

misclassified variants in this "conflicting pathogenicity" set) and up to 92.3% of disorders using the variant set with conflicting interpretations (set 4). This improved to 7.7% and 57.7%, respectively, after filtering for disease-specific minor allele frequency (dMAF). It is possible that, despite the conservative approach, inflation may just alert about diseases with incomplete penetrance, and pathogenic variants that are not clinically manifest due to a number of modifiers.

This work revisits a well-understood relationship between minor allele frequency and disease prevalence. The pattern is one of an excess of individuals with genetic risk relative to the disease prevalence in the population. The inflation increases as the level of support for the pathogenicity of the genetic variants decreases in ClinVar. This means that a number of rare pathogenic variants have low penetrance or that those variants have an incorrect interpretation of pathogenicity. The present analyses strongly suggest that ClinVar includes significant amounts of misclassified variants and supports the important role of ClinVar to increase transparency, contrast claims, and foster validation across submitters. Overall, our work supports the observations of Yang et al.[35] that show that discordance is higher in non-clinical and older submissions and low-penetrance variants. Yang et al. also noted that the concordance differs among clinical areas, with highest consensus rate in hereditary cancer genes and lowest in genes related to cardiology and metabolic disorders.[35]

The ACMG-AMP standards and guidelines for variant interpretations recommends that variants with more than 5% allele frequency should be classified as benign variants (rule BA1).[5] To lower this relatively lenient threshold, a more recent recommendation supports the use of dMAF thresholds based on disease prevalence.[3] Application of the dMAF threshold does correct inflation; few variants contribute to inflation for many of the conditions. However, for some disorders inflation may be hard to pinpoint to low-frequency alleles, and rare variants needed to be considered jointly. This problem is compounded by the very large number of rare variants that are being identified though genome and exome sequencing.[14] Here, rare variation risks are being misconstrued as indicating functional relevance.

The present study also adds to the discussion on the importance of specifying review criteria with the variant submission to ClinVar, and improving the applicability of standardized criteria, such as those of the ACMG-AMP guidelines.[3,41] ACMG-59 genes are now being carefully annotated, and we see limited differences in inflation across star 0–4 levels. In contrast, OrphaNet conditions exhibit inflation as the star classification moves from highest to lowest confidence. We include examples that illustrate some of the sources of misclassification, error, and other issues in the reporting of pathogenic variants. Specifically, we observed cases of rare variants with low penetrance, incorrect mode of inheritance, conditions with unknown or older estimates of population prevalence, and variants that are incorrectly associated with disease.

We explored another strategy to understand the mechanisms of correction of the data currently deposited in ClinVar. We compared two versions of ClinVar released data (May 2016 and September 2017). Changes in classification in successive versions of ClinVar favors a general direction away from pathogenic/likely pathogenic toward VUS and benign/likely benign. However, the bulk of reclassified variants are reassigned to the "conflicting interpretation" category. This observation and our analysis of massive inflation in genetic risk for variants classified as having conflicting interpretation reinforce the notion that pathogenicity of these variants is questionable. More generally, the study supports the aim of reaching a definitive classification for this set of variants to avoid repeated re-assessment in the clinics.

In addition to classifying a variant as pathogenic or benign using ACMG-AMP guidelines, there is a need for a quantitative measure of risk (e.g., penetrance of the variant).[34] Larger genetic studies integrating phenotype data to estimate variant risks using age/sex-specific incidence are needed. Although the concept of this work is anchored in the current practice of genetics, it serves to document the use of current large databases on the general population to better support the classification of variants. The observation of excessive numbers of individuals carrying genetic risk alleles both in well-curated genes sets (ACMG-59) and in other resources (OrphaNet) provides a useful benchmark for the improvement of variant annotation.

Supplemental Data

Supplemental Data include five figures and three tables and can be found with this article online at https://doi.org/10.1016/j.ajhg.2018.02.019.

Acknowledgments

We thank the anonymous reviewers for excellent advice. All authors are current or former employees of Human Longevity.

Received: November 8, 2017
Accepted: February 22, 2018
Published: April 5, 2018

Web Resources

ClinVar, https://www.ncbi.nlm.nih.gov/clinvar/
GeneReviews, Rosenberg et al. (1993). Malignant Hyperthermia Susceptibility. https://www.ncbi.nlm.nih.gov/books/NBK1146/
GeneReviews, Kirmani, S., and Young, W.F. (1993). Hereditary Paraganglioma-Pheochromocytoma Syndromes. https://www.ncbi.nlm.nih.gov/pubmed/20301715
Genetests, http://www.genetests.org/
gnomAD Browser, http://gnomad.broadinstitute.org/
HLI Open Search, http://www.HLI-opensearch.com

OMIM, http://www.omim.org/
OrphaData, http://www.orphadata.org/cgi-bin/index.php/
OrphaNet, http://www.orpha.net/consor/cgi-bin/index.php

References

1. Eilbeck, K., Quinlan, A., and Yandell, M. (2017). Settling the score: variant prioritization and Mendelian disease. Nat. Rev. Genet. *18*, 599–612.
2. Adams, M.C., Evans, J.P., Henderson, G.E., and Berg, J.S. (2016). The promise and peril of genomic screening in the general population. Genet. Med. *18*, 593–599.
3. Amendola, L.M., Jarvik, G.P., Leo, M.C., McLaughlin, H.M., Akkari, Y., Amaral, M.D., Berg, J.S., Biswas, S., Bowling, K.M., Conlin, L.K., et al. (2016). Performance of ACMG-AMP variant-interpretation guidelines among nine laboratories in the clinical sequencing exploratory research Consortium. Am. J. Hum. Genet. *98*, 1067–1076.
4. Rehm, H.L., Berg, J.S., Brooks, L.D., Bustamante, C.D., Evans, J.P., Landrum, M.J., Ledbetter, D.H., Maglott, D.R., Martin, C.L., Nussbaum, R.L., et al.; ClinGen (2015). ClinGen–the clinical genome resource. N. Engl. J. Med. *372*, 2235–2242.
5. Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde, M., Lyon, E., Spector, E., et al.; ACMG Laboratory Quality Assurance Committee (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. Genet. Med. *17*, 405–424.
6. Hunter, J.E., Irving, S.A., Biesecker, L.G., Buchanan, A., Jensen, B., Lee, K., Martin, C.L., Milko, L., Muessig, K., Niehaus, A.D., et al. (2016). A standardized, evidence-based protocol to assess clinical actionability of genetic disorders associated with genomic variation. Genet. Med. *18*, 1258–1268.
7. Rossi, E., Henderson, S., Chin, C.Y., Olynyk, J., Beilby, J.P., Reed, W.D., and Jeffrey, G.P. (1999). Genotyping as a diagnostic aid in genetic haemochromatosis. J. Gastroenterol. Hepatol. *14*, 427–430.
8. Rossi, E., Olynyk, J.K., and Jeffrey, G.P. (2008). Clinical penetrance of C282Y homozygous HFE hemochromatosis. Expert Rev. Hematol. *1*, 205–216.
9. Telenti, A., Pierce, L.C., Biggs, W.H., di Iulio, J., Wong, E.H., Fabani, M.M., Kirkness, E.F., Moustafa, A., Shah, N., Xie, C., et al. (2016). Deep sequencing of 10,000 human genomes. Proc. Natl. Acad. Sci. USA *113*, 11901–11906.
10. Green, R.C., Berg, J.S., Grody, W.W., Kalia, S.S., Korf, B.R., Martin, C.L., McGuire, A.L., Nussbaum, R.L., O'Daniel, J.M., Ormond, K.E., et al.; American College of Medical Genetics and Genomics (2013). ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. Genet. Med. *15*, 565–574.
11. American College of Medical Genetics and Genomics (2013). Incidental findings in clinical genomics: a clarification. Genet. Med. *15*, 664–666.
12. Kalia, S.S., Adelman, K., Bale, S.J., Chung, W.K., Eng, C., Evans, J.P., Herman, G.E., Hufnagel, S.B., Klein, T.E., Korf, B.R., et al. (2017). Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): a policy statement of the American College of Medical Genetics and Genomics. Genet. Med. *19*, 249–255.
13. Rath, A., Olry, A., Dhombres, F., Brandt, M.M., Urbero, B., and Ayme, S. (2012). Representation of rare diseases in health information systems: the Orphanet approach to serve a wide range of end users. Hum. Mutat. *33*, 803–808.
14. Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al.; Exome Aggregation Consortium (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.
15. Goldfeder, R.L., Priest, J.R., Zook, J.M., Grove, M.E., Waggott, D., Wheeler, M.T., Salit, M., and Ashley, E.A. (2016). Medical implications of technical accuracy in genome sequencing. Genome Med. *8*, 24.
16. Zook, J.M., Chapman, B., Wang, J., Mittelman, D., Hofmann, O., Hide, W., and Salit, M. (2014). Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. Nat. Biotechnol. *32*, 246–251.
17. Zook, J.M., Catoe, D., McDaniel, J., Vang, L., Spies, N., Sidow, A., Weng, Z., Liu, Y., Mason, C.E., Alexander, N., et al. (2016). Extensive sequencing of seven human genomes to characterize benchmark reference materials. Sci. Data *3*, 160025.
18. Landrum, M.J., Lee, J.M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Hoover, J., et al. (2016). ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res. *44* (D1), D862–D868.
19. Cingolani, P., Platts, A., Wang, L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. Fly (Austin) *6*, 80–92.
20. Ding, L.E., Burnett, L., and Chesher, D. (2015). The impact of reporting incidental findings from exome and whole-genome sequencing: predicted frequencies based on modeling. Genet. Med. *17*, 197–204.
21. Whiffin, N., Minikel, E., Walsh, R., O'Donnell-Luria, A.H., Karczewski, K., Ing, A.Y., Barton, P.J.R., Funke, B., Cook, S.A., MacArthur, D., and Ware, J.S. (2017). Using high-resolution variant frequencies to empower clinical genome interpretation. Genet. Med. *19*, 1151–1158.
22. Kobayashi, Y., Yang, S., Nykamp, K., Garcia, J., Lincoln, S.E., and Topper, S.E. (2017). Pathogenic variant burden in the ExAC database: an empirical approach to evaluating population data for clinical variant interpretation. Genome Med. *9*, 13.
23. Song, W., Gardner, S.A., Hovhannisyan, H., Natalizio, A., Weymouth, K.S., Chen, W., Thibodeau, I., Bogdanova, E., Letovsky, S., Willis, A., and Nagan, N. (2016). Exploring the landscape of pathogenic genetic variation in the ExAC population database: insights of relevance to variant classification. Genet. Med. *18*, 850–854.
24. Nelson, H.D., Fu, R., Goddard, K., Mitchell, J.P., Okinaka-Hu, L., Pappas, M., and Zakher, B. (2013). In Risk Assessment, Genetic Counseling, and Genetic Testing for BRCA-Related Cancer: Systematic Review to Update the US Preventive Services Task Force Recommendation (Rockville, MD: Agency for Healthcare Research and Quality).
25. Lincoln, S.E., Yang, S., Cline, M.S., Kobayashi, Y., Zhang, C., Topper, S., Haussler, D., Paten, B., and Nussbaum, R.L. (2017). Consistency of BRCA1 and BRCA2 variant classifications among clinical diagnostic laboratories. JCO Precis Oncol *1*. https://doi.org/10.1200/PO.16.00020.

27. Benn, D.E., Gimenez-Roqueplo, A.P., Reilly, J.R., Bertherat, J., Burgess, J., Byth, K., Croxson, M., Dahia, P.L., Elston, M., Gimm, O., et al. (2006). Clinical presentation and penetrance of pheochromocytoma/paraganglioma syndromes. J. Clin. Endocrinol. Metab. *91*, 827–836.
29. Klein, A., Lillis, S., Munteanu, I., Scoto, M., Zhou, H., Quinlivan, R., Straub, V., Manzur, A.Y., Roper, H., Jeannet, P.Y., et al. (2012). Clinical and genetic findings in a large cohort of patients with ryanodine receptor 1 gene-associated myopathies. Hum. Mutat. *33*, 981–988.
30. Gonsalves, S.G., Ng, D., Johnston, J.J., Teer, J.K., Stenson, P.D., Cooper, D.N., Mullikin, J.C., Biesecker, L.G.; and NISC Comparative Sequencing Program (2013). Using exome data to identify malignant hyperthermia susceptibility mutations. Anesthesiology *119*, 1043–1053.
31. Merritt, A., Booms, P., Shaw, M.A., Miller, D.M., Daly, C., Bilmen, J.G., Stowell, K.M., Allen, P.D., Steele, D.S., and Hopkins, P.M. (2017). Assessing the pathogenicity of RYR1 variants in malignant hyperthermia. Br. J. Anaesth. *118*, 533–543.
32. Lazarin, G.A., Haque, I.S., Nazareth, S., Iori, K., Patterson, A.S., Jacobson, J.L., Marshall, J.R., Seltzer, W.K., Patrizio, P., Evans, E.A., and Srinivasan, B.S. (2013). An empirical estimate of carrier frequencies for 400+ causal Mendelian variants: results from an ethnically diverse clinical sample of 23,453 individuals. Genet. Med. *15*, 178–186.
33. Asselta, R., and Peyvandi, F. (2009). Factor V deficiency. Semin. Thromb. Hemost. *35*, 382–389.
34. Manrai, A.K., Ioannidis, J.P., and Kohane, I.S. (2016). Clinical genomics: from pathogenicity claims to quantitative risk estimates. JAMA *315*, 1233–1234.
35. Yang, S., Lincoln, S.E., Kobayashi, Y., Nykamp, K., Nussbaum, R.L., and Topper, S. (2017). Sources of discordance among germ-line variant classifications in ClinVar. Genet. Med. *19*, 1118–1126.
36. Walsh, R., Thomson, K.L., Ware, J.S., Funke, B.H., Woodley, J., McGuire, K.J., Mazzarotto, F., Blair, E., Seller, A., Taylor, J.C., et al. (2017). Reassessment of Mendelian gene pathogenicity using 7,855 cardiomyopathy cases and 60,706 reference samples. Genet. Med. *19*, 192–203.
37. Manrai, A.K., Funke, B.H., Rehm, H.L., Olesen, M.S., Maron, B.A., Szolovits, P., Margulies, D.M., Loscalzo, J., and Kohane, I.S. (2016). Genetic misdiagnoses and the potential for health disparities. N. Engl. J. Med. *375*, 655–665.
38. Tarailo-Graovac, M., Zhu, J.Y.A., Matthews, A., van Karnebeek, C.D.M., and Wasserman, W.W. (2017). Assessment of the ExAC data set for the presence of individuals with pathogenic genotypes implicated in severe Mendelian pediatric disorders. Genet. Med. *19*, 1300–1308.
39. Bennett, C.A., Petrovski, S., Oliver, K.L., and Berkovic, S.F. (2017). ExACtly zero or once: A clinically helpful guide to assessing genetic variants in mild epilepsies. Neurol. Genet. *3*, e163.
40. Xue, Y., Chen, Y., Ayub, Q., Huang, N., Ball, E.V., Mort, M., Phillips, A.D., Shaw, K., Stenson, P.D., Cooper, D.N., Tyler-Smith, C.; and 1000 Genomes Project Consortium (2012). Deleterious- and disease-allele prevalence in healthy individuals: insights from current predictions, mutation databases, and population-scale resequencing. Am. J. Hum. Genet. *91*, 1022–1032.
41. Rehm, H.L. (2017). A new era in the interpretation of human genomic variation. Genet. Med. *19*, 1092–1095.

The American Journal of Human Genetics, Volume *102*

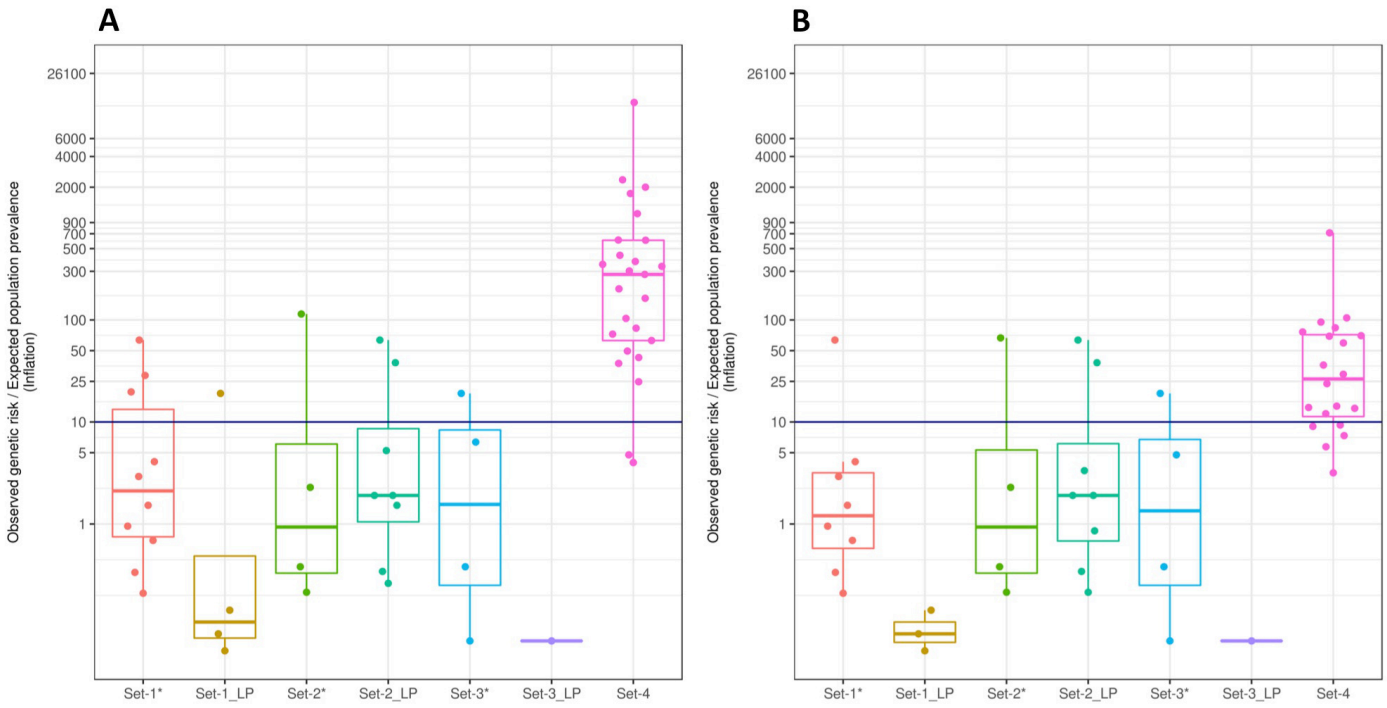
Supplemental Data

Identification of Misclassified ClinVar Variants via Disease Population Prevalence

Naisha Shah, Ying-Chen Claire Hou, Hung-Chun Yu, Rachana Sainger, C. Thomas Caskey, J. Craig Venter, and Amalio Telenti

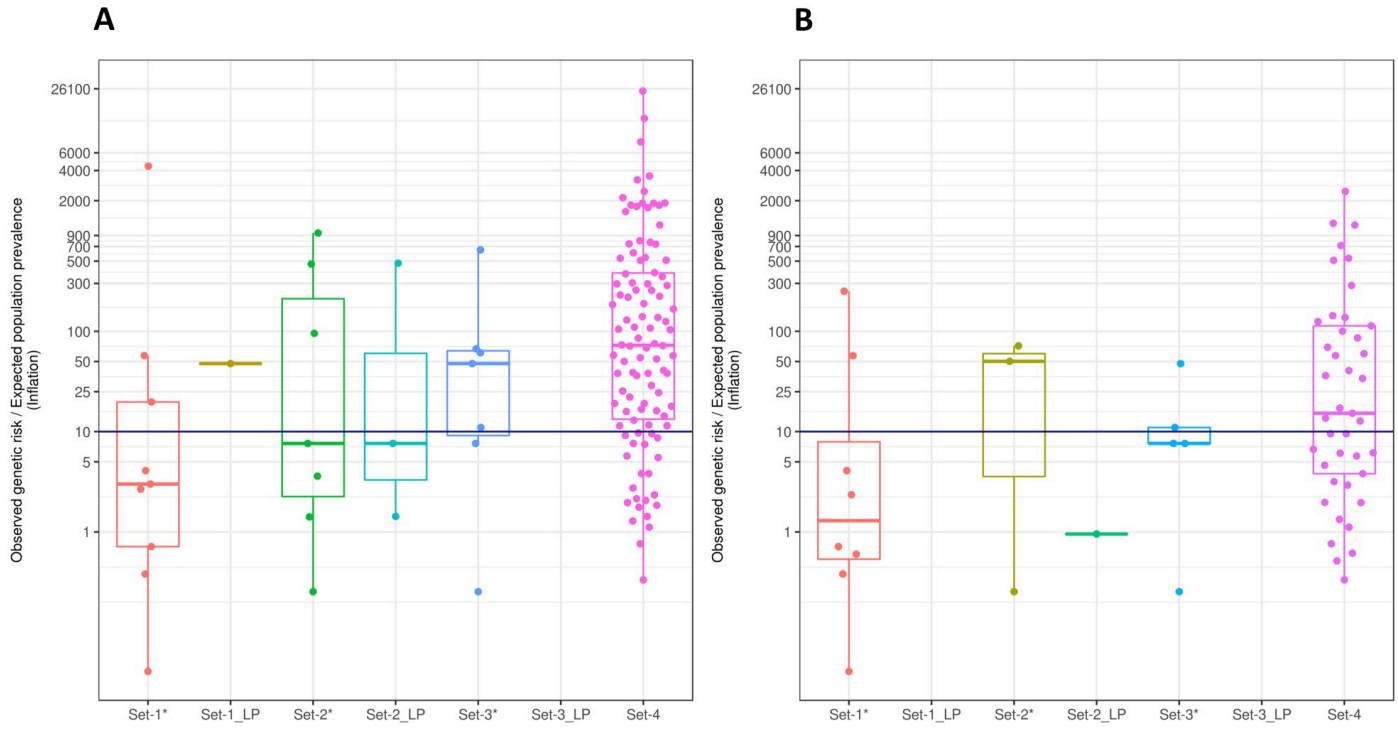
SUPPLEMENTAL FIGURES

Figure S1: Genetic risk in ACMG-59 conditions with additional P and LP sets



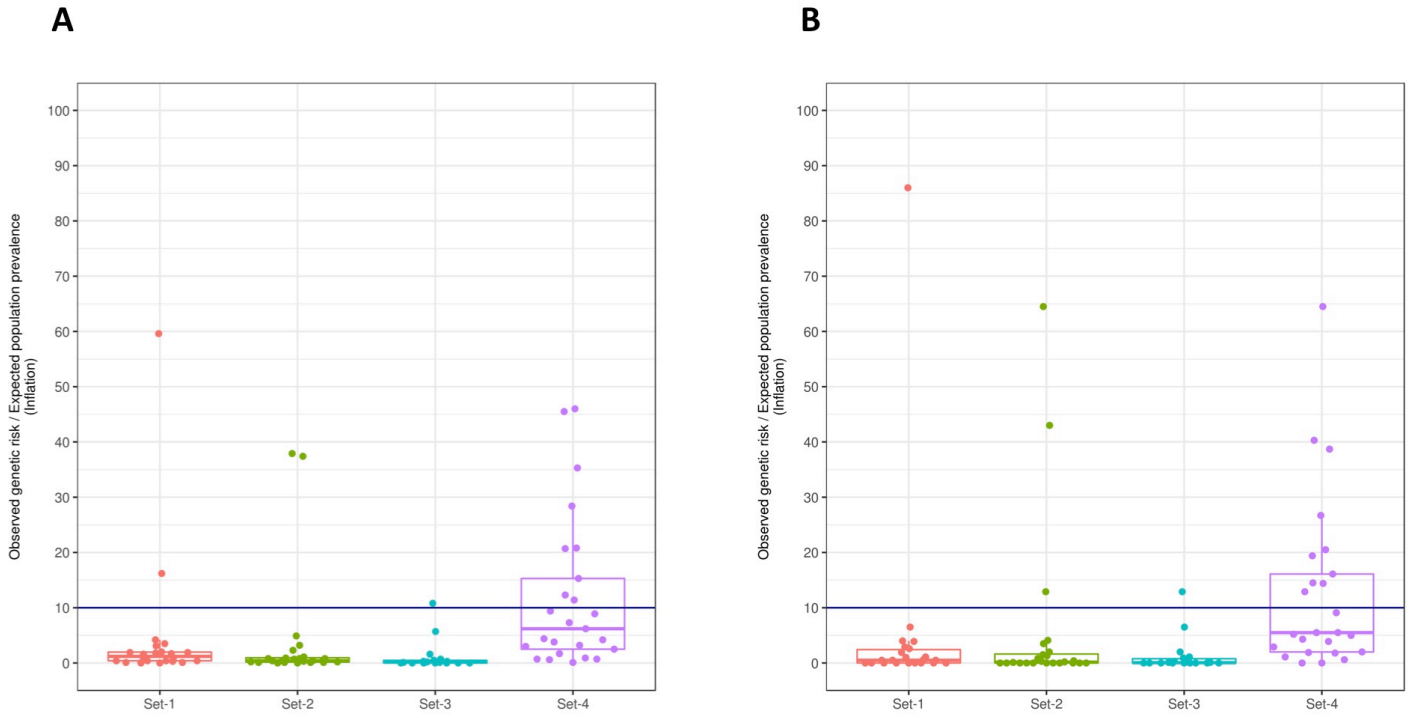
Fold-change of observed genetic risk over expected population prevalence using ClinVar variant sets for the ACMG-59 conditions. The observed genetic risk was calculated using the study population. Each point represents a condition; each condition may be represented in more than one set. The navy-blue line at a fold-change of 10 (i.e. inflation) indicates a theoretical penetrance of 10%. Observations above this line are highly suggestive of misclassified variants. A) Fold-change was calculated using variants per variant set: Set-1* consists of pathogenic (P) variants with 2 or more ClinVar review stars (i.e. two or more submitters with assertion criteria, expert panel and practice guideline); Similarly, Set-1_LP consists of LP variants. Set-2* consists of P variants with 1 star (i.e. one submitter with assertion criteria); Similarly, Set-2_LP consists of LP variants. Set-3* consists of P variants with 0 star (i.e. submitter with no assertion criteria submitted in ClinVar); Similarly, Set-3_LP consists of LP variants. Set-4 consists of variants with conflicting interpretations of pathogenicity. B) Fold-change was re-calculated after variants were filtered for disease-specific minor allele frequency thresholds.

Figure S2: Genetic risk in Orphanet conditions with additional P and LP sets



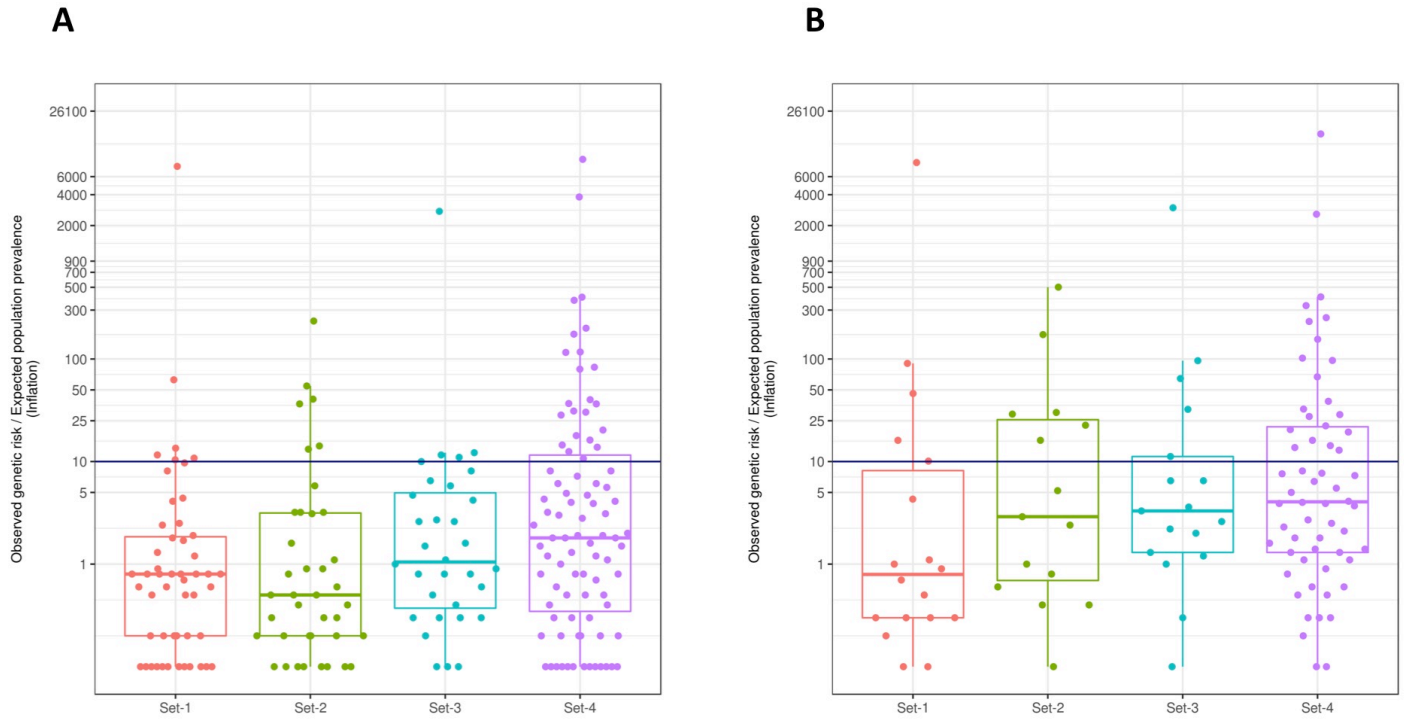
Fold-change of observed genetic risk over expected population prevalence using ClinVar variant sets for the Orphanet conditions. The observed genetic risk was calculated using the study population. Each point represents a condition; each condition may be represented in more than one set. The navy-blue line at a fold-change of 10 (i.e. inflation) indicates a theoretical penetrance of 10%. Observations above this line are highly suggestive of misclassified variants. A) Fold-change was calculated using variants per variant set: Set-1* consists of pathogenic (P) variants with 2 or more ClinVar review stars (i.e. two or more submitters with assertion criteria, expert panel and practice guideline); Similarly, Set-1_LP consists of LP variants. Set-2* consists of P variants with 1 star (i.e. one submitter with assertion criteria); Similarly, Set-2_LP consists of LP variants. Set-3* consists of P variants with 0 star (i.e. submitter with no assertion criteria submitted in ClinVar); Similarly, Set-3_LP consists of LP variants. Set-4 consists of variants with conflicting interpretations of pathogenicity. B) Fold-change was re-calculated after variants were filtered for disease-specific minor allele frequency thresholds.

Figure S3: Genetic risk in ACMG-59 conditions using gnomAD



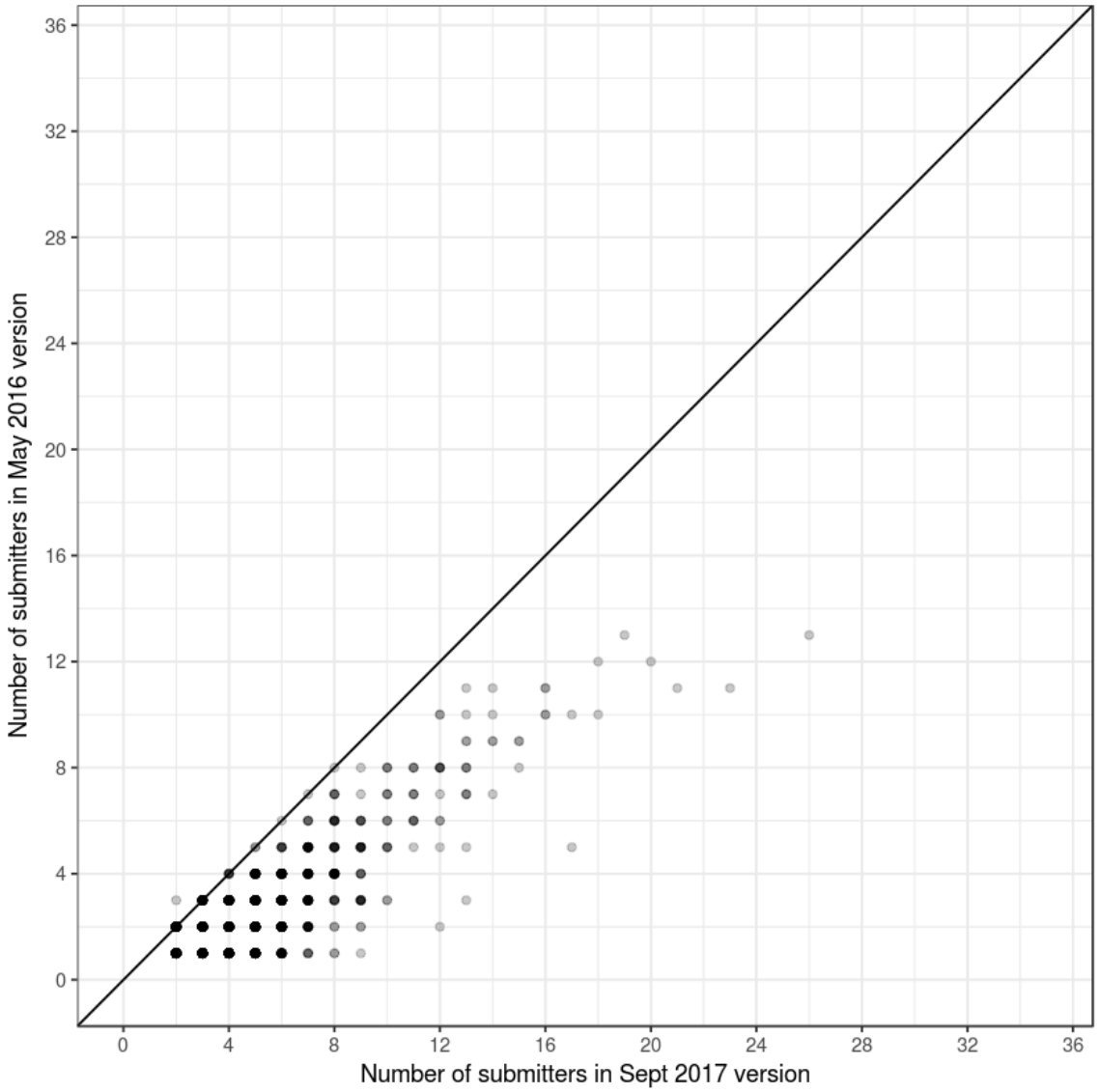
Fold-change of observed genetic risk over expected population prevalence using ClinVar variant sets for the ACMG-59 conditions. Each point represents a condition; each condition may be represented in more than one set. The navy-blue line at a fold-change of 10 (i.e. inflation) indicates a theoretical penetrance of 10%. Observations above this line are highly suggestive of misclassified variants. Fold-change was calculated using variants after disease-specific minor allele frequency filtering per variant set: Set-1 consists of variants with 2 or more ClinVar review stars (i.e. two or more submitters with assertion criteria, expert panel and practice guideline); Set-2 consists of variants with 1 star (i.e. one submitter with assertion criteria); Set-3 consists of variants with 0 star (i.e. submitter with no assertion criteria submitted in ClinVar); Set-4 consists of variants with conflicting interpretations of pathogenicity. A) The observed genetic risk was calculated using gnomAD exome data. B) The observed genetic risk was calculated using gnomAD genome data.

Figure S4: Genetic risk in Orphanet conditions using gnomAD



Fold-change of observed genetic risk over expected population prevalence using ClinVar variant sets for the Orphanet conditions. Each point represents a condition; each condition may be represented in more than one set. The navy-blue line at a fold-change of 10 (i.e. inflation) indicates a theoretical penetrance of 10%. Observations above this line are highly suggestive of misclassified variants. Fold-change was calculated using variants after disease-specific minor allele frequency filtering per variant set: Set-1 consists of variants with 2 or more ClinVar review stars (i.e. two or more submitters with assertion criteria, expert panel and practice guideline); Set-2 consists of variants with 1 star (i.e. one submitter with assertion criteria); Set-3 consists of variants with 0 star (i.e. submitter with no assertion criteria submitted in ClinVar); Set-4 consists of variants with conflicting interpretations of pathogenicity. A) The observed genetic risk was calculated using gnomAD exome data. B) The observed genetic risk was calculated using gnomAD genome data.

Figure S5: Number of ClinVar submitters for variants reclassified to conflicting interpretations of pathogenicity



For 855 P/LP, 2525 VUS, and 2487 B/LB variants that changed its classification to conflicting interpretations of pathogenicity from May 2016 version of ClinVar to September 2017 version, the plot shows the number of ClinVar submitters.

SUPPLEMENTAL TABLES

Table S1: ClinVar variant sets

A list of ClinVar variant sets that was used in the study. The chromosomal positions are in GRCh38 human reference built. Provided as a separate excel file.

Table S2: Genetic risk in ACMG-59 conditions.

Conditions	Estimated Population Prevalence	Mode of Inheritance	Genes	Observed Genetic risk	Fold Change	Observed Genetic Risk (dMAF)	Fold Change (dMAF)
Lynch Syndrome	227.27	Autosomal dominant	MLH1 MSH2 MSH6 PMS2	105	0.5	105	0.5
Familial hypercholesterolemia	500	Autosomal dominant	APOB LDLR PCSK9	314	0.6	314	0.6
Ehlers-Danlos syndrome, vascular type	1	Autosomal dominant	COL3A1	0	-	0	-
Familial adenomatous polyposis	3.2	Autosomal dominant	APC	0	-	0	-
Catecholaminergic polymorphic ventricular tachycardia	10	Autosomal dominant	RYR2	10	-	10	-
Ornithine transcarbamylase deficiency	7.14	X-linked recessive	OTC	10	-	10	-
Hypertrophic cardiomyopathy, Dilated cardiomyopathy	400	Autosomal dominant; X-linked recessive	MYBPC3 MYH7 TNNT2 TNNI3 TPM1 MYL3 ACTC1 PRKAG2 MYL2 LMNA GLA	581	1.5	562	1.4
WT1-related Wilms tumor	10	Autosomal dominant	WT1	19	1.9	19	1.9
Arrhythmogenic right ventricular cardiomyopathy	100	Autosomal dominant	DSC2 DSG2 DSP PKP2 TMEM43	295	3.0	229	2.3
Juvenile polyposis	6.25	Autosomal dominant	BMPR1A SMAD4	0	-	0	-
Hereditary Breast and Ovarian Cancer	250	Autosomal dominant	BRCA1 BRCA2	448	1.8	448	1.8

MYH-Associated Polyposis; Adenomas, multiple colorectal, FAP type 2; Colorectal adenomatous polyposis,autosomal recessive,with pilomatricomas	2.5	Autosomal recessive	MUTYH	10	-	10	-
Marfan Syndrome, Loeys-Dietz Syndromes, and Familial Thoracic Aortic Aneurysms and Dissections	20	Autosomal dominant	ACTA2 FBN1 MYH11 SMAD3 TGFBR1 TGFBR2	114	5.7	76	3.8
Romano-Ward Long QT Syndromes Types 1, 2, and 3, Brugada Syndrome	50	Autosomal dominant	KCNH2 KCNQ1 SCN5A	305	6.1	305	6.1
Multiple Endocrine Neoplasia Type 1	3.3	Autosomal dominant	MEN1	0	-	0	-
Li-Fraumeni Syndrome	7	Autosomal dominant	TP53	48	6.8	48	6.8
Retinoblastoma	6	Autosomal dominant	RB1	48	7.9	38	6.4
Neurofibromatosis type 2	1.78	Autosomal dominant	NF2	0	-	0	-
Familial Medullary Thyroid Cancer (FMTC);Multiple Endocrine Neoplasia Type 2	2.9	Autosomal dominant	RET	57	19.7	10	3.3
Peutz-Jeghers Syndrome	2.2	Autosomal dominant	STK11	0	-	0	-
PTEN Hamartoma Tumor Syndrome	0.5	Autosomal dominant	PTEN	0	-	0	-
Hereditary Paraganglioma-Pheochromocytoma Syndrome	0.3	Autosomal dominant	SDHAF2 SDHB SDHC SDHD	48	158.8	48	158.8
Tuberous Sclerosis Complex	17.2	Autosomal dominant	TSC1 TSC2	0	-	0	-
Von Hippel Lindau syndrome	2.3	Autosomal dominant	VHL	0	-	0	-
Wilson disease	10	Autosomal recessive	ATP7B	0	-	0	-
Malignant hyperthermia susceptibility	1	Autosomal dominant	CACNA1S RYR1	219	219.2	133	133.4

A list of ACMG-59 conditions with at least one P/LP variant from set-1, set-2 or set-3 observed in the study. The last two columns with “dMAF” suffix (“Observed Genetic risk (dMAF)” and “Fold Change (dMAF)”) are observed genetic risk and fold change calculated after filtering variants using disease-specific minor allele frequency (dMAF) filter. The population prevalence and genetic risk are calculated per 100,000. Fold change was not calculated if only one individual of the 10,495 samples was identified with genetic risk of the disease condition.

Table S3: Genetic risk in Orphanet conditions.

Orphanet ID	Condition	Estimated Population Prevalence	Mode of Inheritance	Genes	Observed Genetic risk	Fold Change	Observed Genetic Risk (dMAF)	Fold Change (dMAF)
48	Congenital bilateral absence of vas deferens	50	Autosomal recessive	CFTR	10	-	10	-
55	Oculocutaneous albinism	45	Autosomal recessive	OCA2	10	-	10	-
60	Alpha-1-antitrypsin deficiency	63.5	Autosomal recessive	SERPINA1	191	3.00	38	0.60
122	Birt-Hogg-Dubé syndrome	0.5	Autosomal dominant	FLCN	29	57.17	29	57.17
130	Brugada syndrome	75	Autosomal dominant	CACNA1C-AS1;CACNA1C;CACNA1C-AS2;CACNA1C;CACNA1C;CACNA2D1;CACNB2;GPD1L;KCNE3;NSUN6;CACNB2;SCN10A;SCN3B;SCN5A;TRPM4	38	0.51	38	0.51
145	Hereditary breast and ovarian cancer syndrome	250	Autosomal dominant	BRCA1;BRCA2	400	1.60	353	1.41
212	Cystathioninuria	7.1	Autosomal recessive	CTH	10	-	0	-
232	Sickle cell anemia	467.3	Autosomal recessive	HBB	19	0.04	19	0.04
268	Autosomal recessive limb-girdle muscular dystrophy type 2B	0.13	Autosomal recessive	DYSF	10	-	0	-
282	Frontotemporal dementia	3	Autosomal dominant	CHMP2B;POU1F1;MAPT	10	-	10	-
287	Ehlers-Danlos syndrome, classic type	5	Autosomal dominant	COL5A1;COL5A2;LOC101448202;COL5A1	10	-	10	-
324	Fabry disease	1.11	X-linked recessive	RPL36A-HNRNPH2;GLA	10	-	10	-
325	Congenital factor II deficiency	0.05	Autosomal recessive	F2	48	952.83	0	-
377	Gorlin syndrome	5.3	Autosomal dominant	LOC100507346;PTCH1;PTCH1;PTCH2;SUFU	19	3.60	0	-

429	Hypochondroplasia	3.3	Autosomal dominant	FGFR3	10	-	10	-
524	Li-Fraumeni syndrome	7	Autosomal dominant	TP53	29	4.08	29	4.08
558	Marfan syndrome	20	Autosomal dominant	FBN1;TGFBR2	29	1.43	19	0.95
565	Menkes disease	2.5	X-linked recessive	ATP7A	19	7.62	0	-
586	Cystic fibrosis	111	Autosomal recessive	CFTR	10	-	10	-
597	Central core disease	0.4	Autosomal dominant	RYR1	86	214.39	57	142.93
636	Neurofibromatosis type 1	50	Autosomal dominant	NF1	19	0.38	19	0.38
652	Multiple endocrine neoplasia type 1	3.3	Autosomal dominant	MEN1;RET	10	-	10	-
653	Multiple endocrine neoplasia type 2	2.9	Autosomal dominant	RET	57	19.71	10	-
661	Ondine syndrome	0.5	Autosomal dominant	BDNF-AS;BDNF;GDNF;RET	324	647.93	0	-
676	Hereditary chronic pancreatitis	0.57	Autosomal dominant	CFTR;CTRC;SPINK1	2820	4948.05	172	300.90
758	Pseudoxanthoma elasticum	2.5	Autosomal recessive	ABCC6	152	60.98	19	7.62
759	Central precocious puberty	20	Autosomal dominant	KISS1R	152	7.62	0	-
790	Retinoblastoma	6	Autosomal dominant	RB1	10	-	10	-
882	Tyrosinemia type 1	54	Autosomal recessive	FAH	76	1.41	0	-
1243	Best vitelliform macular dystrophy	20	Autosomal dominant	BEST1	19	0.95	19	0.95
2152	Mowat-Wilson syndrome	1.7	Autosomal dominant	ZEB2	10	-	10	-
2337	Non-epidermolytic palmoplantar keratoderma	2.5	Autosomal dominant	AQP5	19	7.62	19	7.62
2686	Cyclic neutropenia	0.1	Autosomal dominant	ELANE	10	-	10	-
3193	Supravalvular aortic stenosis	13.3	Autosomal dominant	ELN	10	-	0	-
32960	Tumor necrosis factor receptor 1 associated periodic syndrome	0.1	Autosomal dominant	TNFRSF1A	48	476.42	0	-
44890	Gastrointestinal stromal tumor	14.5	Autosomal dominant	KIT;PDGFRA;SDHB;SDHC	19	1.31	19	1.31
79241	Biotinidase deficiency	5	Autosomal recessive	BTD	10	-	10	-

79432	Oculocutaneous albinism type 2	46.15	Autosomal recessive	OCA2	10	-	10	-
98672	Autosomal dominant optic atrophy	83	Autosomal dominant	OPA1	10	-	10	-
98878	Hemophilia A	19.3	X-linked recessive	F8	10	-	10	-
98879	Hemophilia B	4	X-linked recessive	F9	10	-	10	-
100985	Autosomal dominant spastic paraplegia type 4	0.91	Autosomal dominant	SPAST	10	-	10	-
101016	Romano-Ward syndrome	40	Autosomal dominant	KCNQ1	29	0.71	29	0.71
182090	Pulmonary arterial hypertension	5.2	Autosomal dominant	BMPR2;ENG;LOC102723566;ENG;SMAD9	57	10.99	57	10.99

A list of Orphanet conditions with at least one P/LP variant from set-1, set-2 or set-3 observed in the study. The last two columns with "dMAF" suffix ("Observed Genetic risk (dMAF)" and "Fold Change (dMAF)") are observed genetic risk and fold change calculated after filtering variants using disease-specific minor allele frequency (dMAF) filter. The population prevalence and genetic risk are calculated per 100,000. Fold change was not calculated if only one individual of the 10,495 samples was identified with genetic risk of the disease condition.