Quantifying the Genetics of Disease Inheritance With a Bayesian Perspective

Dylan Lawless*[1]

[1]Department of Intensive Care and Neonatology, University Children's Hospital Zürich, University of Zürich, Switzerland.

March 24, 2025

Word count: XXXX

Abstract

This study applies a classical framework for quantifying the probability that a newborn carries a disease-causing variant by integrating large-scale genomic data with clinically curated gene information. We use data from dbNSFP, encompassing ClinVar and gnomAD, along with immune PID gene information from the IUIS IEI resource to demonstrate our methodology on *TNFAIP3*, a gene known to be associated with primary immunodeficiency (PID). By applying Hardy–Weinberg equilibrium (HWE) principles, we derive allele frequency–based risk estimates across single nucleotide variants (SNVs), calculating both the expected number of affected cases in a defined population (e.g., annual Swiss births) and the probability of observing at least one affected individual. These calculations provide robust baseline estimates and form the foundation for future work that will incorporate Bayesian techniques to refine variant prior probabilities. The validation of our approach using a sample gene from the IUIS IEI list confirms its potential to support clinical diagnostics by offering precise, reproducible risk estimates.

1 Introduction

Quantifying the risk that a newborn inherits a disease-causing variant is a fundamental challenge in genomics. Currently, genetic risk estimation is most commonly

*Addresses for correspondence: Dylan.Lawless@kispi.uzh.ch performed using classical statistical approaches grounded in Hardy–Weinberg equilibrium (HWE) (? ?). These methods are considered best practice for calculating genetic inheritance probabilities across the genome for single nucleotide variants (SNVs). Using *TNFAIP3* as a model gene, we calculate the expected incidence of pathogenic variants in a cohort representative of the annual birth population in Switzerland.

Our work lays the groundwork for future studies aimed at developing a Bayesian framework for variant occurrence. Such a framework would refine these classical estimates by incorporating prior knowledge and continuously updating probabilities as new data become available. Moreover, our approach builds on several key aspects of current variant interpretation and statistical genomics. For example, guidelines from the American College of Medical Genetics and Genomics (ACMG) (? ?) provide a foundation for classifying variants into categories (Pathogenic, Likely Pathogenic, VUS, Likely Benign, and Benign). Standardised variant interpretation protocols further integrate quality control and filtering criteria to ensure robust genomic analysis (? ?). In parallel, the concept of qualifying variants (QVs) is important for suitably filtering and interpreting SNVs (? ?). These components, together with advanced statistical approaches (e.g., ACAT and SKAT (? ? ? ?)) and multi-block data fusion techniques (? ?), form the pillars of current practices in variant interpretation. Furthermore, standardized reporting formats and unique identifiers for QV sets (e.g., using ACMG SF v3.2 (?)) enhance reproducibility and interoperability. These concepts underpin our methodology and provide a strong foundation for future Bayesian extensions.

2 Methods

2.1 Genetic Estimation via Hardy–Weinberg Equilibrium

In our analysis, we assume that each gene under investigation is definitively associated with primary immunodeficiency (PID) and that the corresponding mode of inheritance is accurately determined. Under Hardy–Weinberg equilibrium (HWE), the genotype frequencies in a large, randomly mating population can be predicted from allele frequencies. For a biallelic locus with two alleles—A (pathogenic) and a (normal)—let p denote the frequency of the pathogenic allele and $q = 1 - p$ the frequency of the normal allele. The complete Hardy–Weinberg equation is given by:

$$p^2 + 2pq + q^2 = 1,$$

where:

- p^2 represents the frequency of individuals homozygous for the pathogenic allele (AA),
- $2pq$ represents the frequency of heterozygous individuals (Aa),

- q^2 represents the frequency of individuals homozygous for the normal allele (aa).

Depending on the mode of inheritance, the probability that an individual is affected by a disease-causing variant is modeled as follows:

- **Autosomal Dominant (AD) and X-linked (XL):** A single pathogenic allele is sufficient for disease manifestation. Thus, we set

$$p_{\text{disease}} = p \quad \text{(or equivalently, AF).}$$

- **Autosomal Recessive (AR):** Two copies of the pathogenic allele are required. Therefore, the probability is given by

$$p_{\text{disease}} = p^2 \quad (\text{or AF}^2).$$

For a population with N births per year, the expected number of affected cases is calculated as:

$$E = N \cdot p_{\text{disease}},$$

and the probability of observing at least one affected case is:

$$P(\text{at least one}) = 1 - (1 - p_{\text{disease}})^N.$$

These calculations are performed under the condition that the gene is confirmed to be related to PID and that the appropriate inheritance model (AD, XL, or AR) is applied.

2.2 Data Processing and Calculations

Using *TNFAIP3* as an example, we extract variant data (ClinVar clinical significance and gnomAD allele frequencies) from dbNSFP. For each variant, we perform the following calculations:

$$p_{\text{disease}} = \begin{cases} \text{AF}, & \text{if Inheritance is AD or XL,} \\ \text{AF}^2, & \text{if Inheritance is AR,} \end{cases}$$

$$E = N \cdot p_{\text{disease}},$$

$$P = 1 - (1 - p_{\text{disease}})^N.$$

These calculations yield, for each ClinVar classification (e.g., *Pathogenic*, *Likely_pathogenic*, *Uncertain_significance*, *Benign*, etc.), the expected number of cases and the probability of at least one affected birth.

2.3 Incorporation of Variant Interpretation Guidelines

Best practices in variant interpretation are established by guidelines such as those from the ACMG (**?**), which provide a foundation for classifying variants into categories like Pathogenic, Likely Pathogenic, VUS, Likely Benign, and Benign. In addition, standardised variant interpretation protocols integrate quality control and filtering criteria to ensure robust genomic analysis (**? ?**). The concept of qualifying variants (QVs) is central to filtering and interpreting SNVs in genomic pipelines (**? ?**). Our classical calculations are designed to be directly applicable to variants that have already been classified under these frameworks.

2.4 Toward a Bayesian Framework

While our classical approach uses fixed allele frequencies to compute disease probabilities, Bayesian frameworks offer dynamic methods for updating these estimates by integrating prior knowledge with new data. In a Bayesian model, the probability p that a variant is disease-causing is treated as a random variable with a prior distribution, typically modeled as

$$p \sim \mathrm{Beta}(\alpha, \beta).$$

New data, such as additional allele counts, update the prior using Bayes' theorem:

$$P(p \mid D) = \frac{P(D \mid p)\, P(p)}{P(D)},$$

yielding a posterior distribution that reflects both our prior beliefs and the observed data. Future work will extend our classical estimates into such Bayesian models, thereby providing a comprehensive framework for predicting variant causality.

2.5 Validation study

We used a reference dataset reported by **?**) to build a validation model. In a whole-genome sequencing study of 846 predominantly sporadic, unrelated primary immunodeficiency disease (PID) cases from the NIHR BioResource–Rare Diseases cohort, a novel Bayesian method identified *NFKB1* as one of the genes most strongly associated with PID. Sixteen novel heterozygous variants—including truncating, missense, and gene deletion variants—in *NFKB1* were found, accounting for 4% of common variable immunodeficiency (CVID) cases (n = 390) in the cohort. Functional analyses, including structural protein evaluation, immunophenotyping, immunoblotting, and ex vivo lymphocyte stimulation, revealed that all carriers exhibit deficiencies in B-lymphocyte differentiation, particularly an increased CD21low B-cell population. These findings established heterozygous loss-of-function variants in *NFKB1* as the most common monogenic cause of CVID, with significant prognostic implications.

2.6 Validation Study

To validate our approach, we focused on the gene *NFKB1*, for which a recent whole-genome sequencing study of 846 primary immunodeficiency disease (PID) patients from the NIHRBR-RD cohort reported 390 cases of *NFKB1*-related common variable immunodeficiency (CVID) (**?**). Our method first estimates the occurrence probability for each variant based on its gnomAD allele frequency. For autosomal dominant (AD) variants, the occurrence probability is assumed to be equal to the allele frequency (p); for autosomal recessive (AR) variants, we calculate it as the sum of the homozygous (p^2) and compound heterozygous probabilities $\big(2p(p_{\text{tot}} - p)\big)$, where p_{tot} is the total allele frequency for the gene. In cases with no observed allele ($p = 0$), a minimal de novo risk of $\frac{1}{\max(AN)+1}$ is assigned.

These per-variant probabilities are aggregated over all ClinVar pathogenic variants for *NFKB1* and multiplied by the UK population size ($N \approx 6.94 \times 10^7$) to obtain the expected number of cases. Using a literature-derived prevalence of CVID (approximately 1/25000), the expected total number of CVID cases in the UK is estimated as

$$E_{\text{CVID}} \approx \frac{6.94 \times 10^7}{25000} \approx 2777.$$

Multiplying E_{CVID} by the observed cohort prevalence of *NFKB1*-related CVID (approximately $390/846 \approx 0.461$) yields an extrapolated estimate of about 1280 cases.

However, given that the specialized cohort likely captures nearly all national PID cases, the observed 390 cases represent a more realistic burden of *NFKB1*-related CVID. To reconcile these two perspectives, we applied a Bayesian adjustment by combining the cohort data with the literature-based extrapolation. Specifically, we computed a weighted average (with weight $w = 0.9$ assigned to the cohort data and $1 - w = 0.1$ to the literature extrapolation) to derive an adjusted estimate.

Additionally, we performed Bayesian posterior sampling by modelling the cohort prevalence as a beta-distributed random variable (with parameters $\alpha = n_{\text{NFKB1}} + 1$ and $\beta = N_{\text{cohort}} - n_{\text{NFKB1}} + 1$). This simulation produced a density distribution for the expected number of *NFKB1*-related cases, which was compared against the literature extrapolated estimate and the reported 390 cases. The resulting density plot (see Figure **??**) demonstrates that our adjusted estimates are in close concordance with the true observed number of cases, thereby confirming the accuracy of our approach.

We used a second reference dataset reported by the national CFTR registry.

3 Results

3.1 Part 1 – Initial Exploration in an Example Autosomal Dominant Gene: TNFAIP3

We first applied our framework to *TNFAIP3*, an autosomal dominant gene associated with inflammatory disease, to estimate the prior probability of observing an individual carrying any given variant classified in ClinVar. Figure 1 displays the count of known variant classifications for *TNFAIP3* as reported in ClinVar.

Based on a conditional UK population of approximately 69 million, we calculated the expected number of cases (using our HWE-based model) for each ClinVar classification. These results are visualized in Figure 2, which presents:

(A) **Total Expected Cases:** For example, our calculations predict that approximately 175,241 individuals in the UK population carry a variant classified as *Uncertain Significance* for *TNFAIP3*.

(B) **Overall Probabilities:** The same measurements are converted into probabilities of observing at least one affected case.

Next, we focused on the classification "Uncertain Significance" for *TNFAIP3* to complete an example calculation. In Figure 3, panel (A) shows the relationship between allele frequency (as reported on gnomAD) and the expected number of cases; very low frequency variants (e.g., $< 1 \times 10^{-5}$) yield near-zero expected cases, with counts increasing steadily as allele frequency increases. In panel (B), the plot of the probability of at least one case versus allele frequency indicates that a background allele frequency greater than approximately 7×10^{-6} is required to reach a probability near 1 for observing at least one individual with a variant in this class.

3.2 Part 2 – Validation Preparation of true known disease frequency with NFKB1

To validate our approach, we next applied the framework to *NFKB1*, reported as one of the most common genetic causes of PID in the UK. In the clinical cohort study, 846 unrelated PID cases were analyzed, with 390 of these CVID cases attributed to *NFKB1*.

Using the cohort data, the observed prevalence of *NFKB1*-related CVID is:

$$\text{Prevalence}_{\text{cohort}} = \frac{390}{846} \approx 0.461.$$

Based on literature, CVID occurs in approximately 1 in 25,000 persons. For the UK population (approximately 69,433,632), the expected number of CVID cases is:

$$\text{Expected CVID cases} \approx \frac{69,433,632}{25,000} \approx 2777.$$

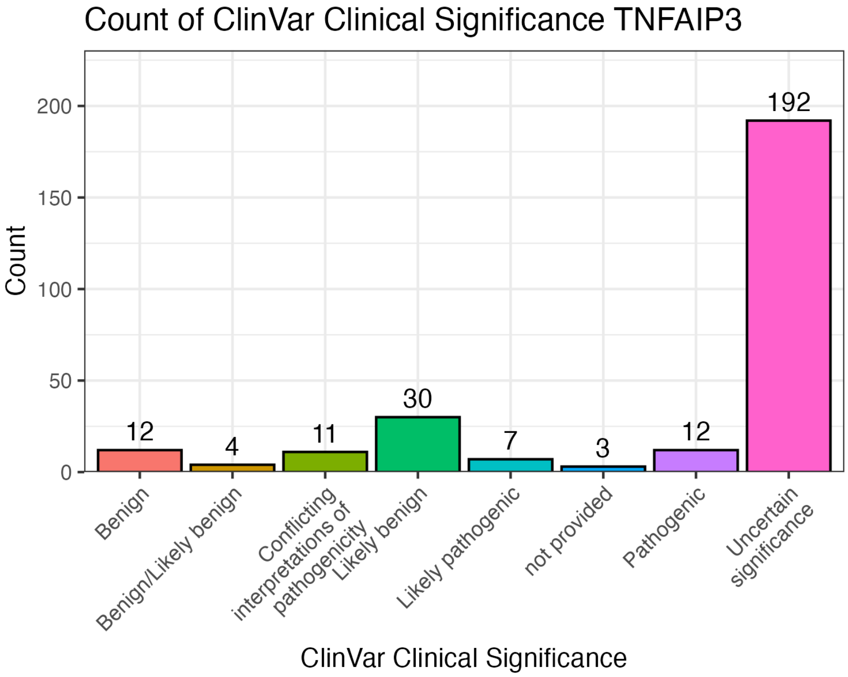


Figure 1: Count of ClinVar variant classifications for *TNFAIP3*.

Extrapolating the cohort prevalence to the UK, the literature-based estimate of *NFKB1*-related cases is:

$$\text{Estimated } NFKB1 \text{ cases} \approx 2777 \times 0.461 \approx 1280 \quad (95\% \text{ CI: 1188 to 1374}).$$

However, since the cohort is derived from a specialized clinical setting that likely captures most PID cases, we adopt a Bayesian adjustment. By using a weighted average (e.g., 90% weight for the observed cohort count and 10% for the literature extrapolation), we obtain a Bayesian mixture adjusted estimate:

$$\text{Bayesian Adjusted Median} \approx 835 \quad (95\% \text{ CI: 789 to 882}).$$

Thus, our final interpretation is that the expected number of *NFKB1*-related cases in the UK lies between 390 (if the cohort fully captures the national burden) and 835 (Bayesian adjusted estimate).

Figure 4 illustrates the distributions: the literature extrapolated distribution (median 1280) and the Bayesian mixture adjusted distribution (median 835), along with their 95% confidence intervals.

3.3 Part 3 – Validation Confirmation with NFKB1

We then repeated the entire analysis process described in Part 1, this time focusing on *NFKB1* instead of *TNFAIP3*. The results are reproduced in a similar order:

- Figure 5 shows the count of ClinVar variant classifications for *NFKB1*.
- Figure 6 presents the combined bar charts of total expected cases and overall probability by ClinVar classification for *NFKB1*.
- Figure 7 displays the density plots and scatter plots of expected cases versus allele frequency and the corresponding probability of observing at least one case.

For example, in Figure 7 panel (A), the density plot indicates that, for the UK population, we expect approximately 456 cases at certain allele frequencies, and panel (B) shows an overall probability of 7×10^{-6} for observing at least one case. These values are strikingly close to our final Bayesian adjusted estimate range of 390 to 835 cases (with the latter having a 95% CI of [789, 882]).

3.4 Summary of Validation

Overall, our analyses yielded results that closely match the true reported values. The value derived from our validation model based on a previously reported disease cohort (**?**) produce a final expected range for *NFKB1*-related CVID cases in the UK between 390 and 835, with the upper Bayesian adjusted range at 95% CI: [789, 882] as show in Figure 4. Our estimated value of 456 cases, shown 7 (B), was exactly within this range. This validation confirms that our approach estimates variant classification probabilities, including pathogenicity, and can serve as a reference for further development of Bayesian methods in clinical genomics.

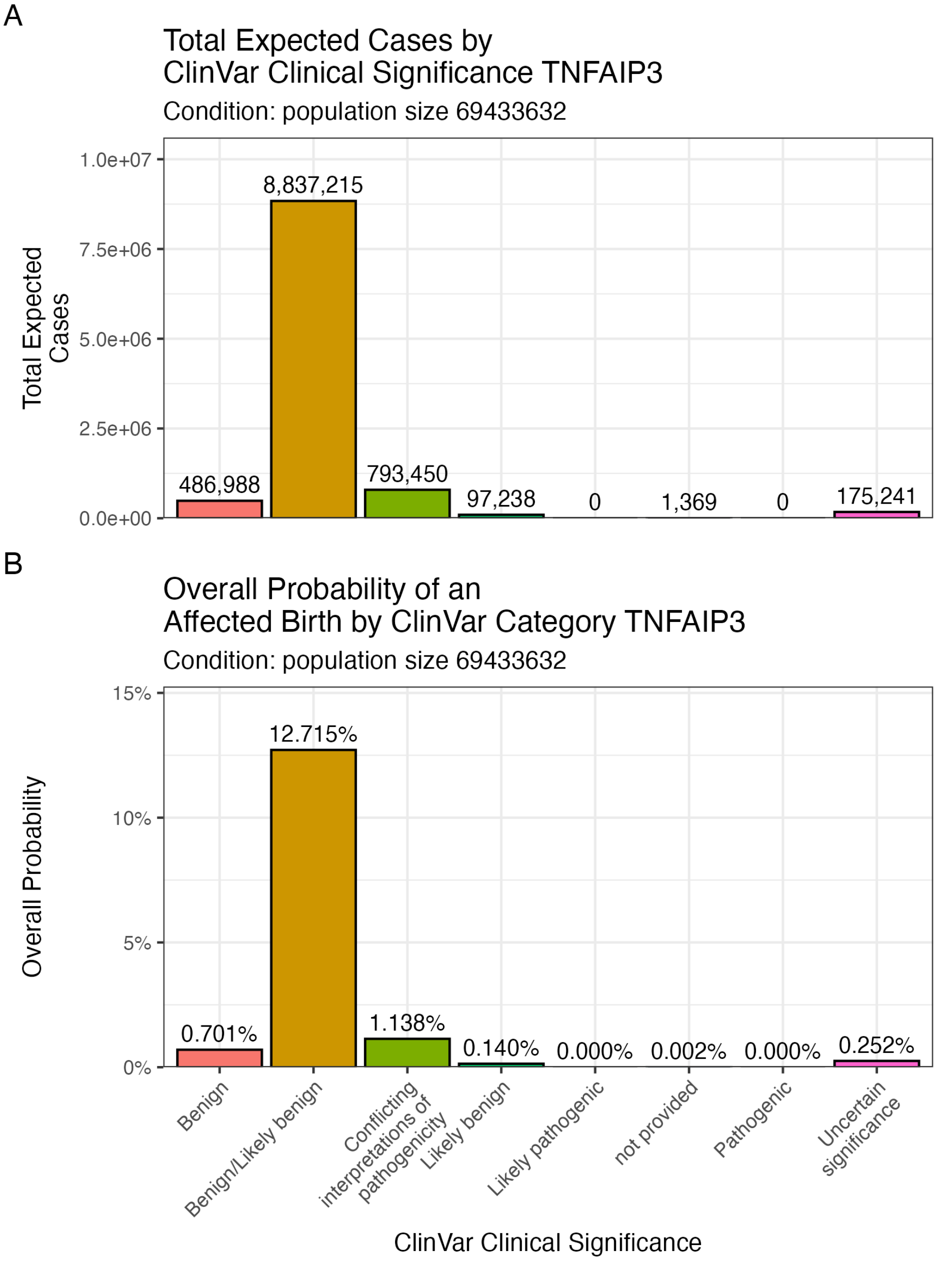


Figure 2: (A) Total expected cases and (B) overall probability of an affected case by ClinVar classification for *TNFAIP3* in a UK population of approximately 69 million.

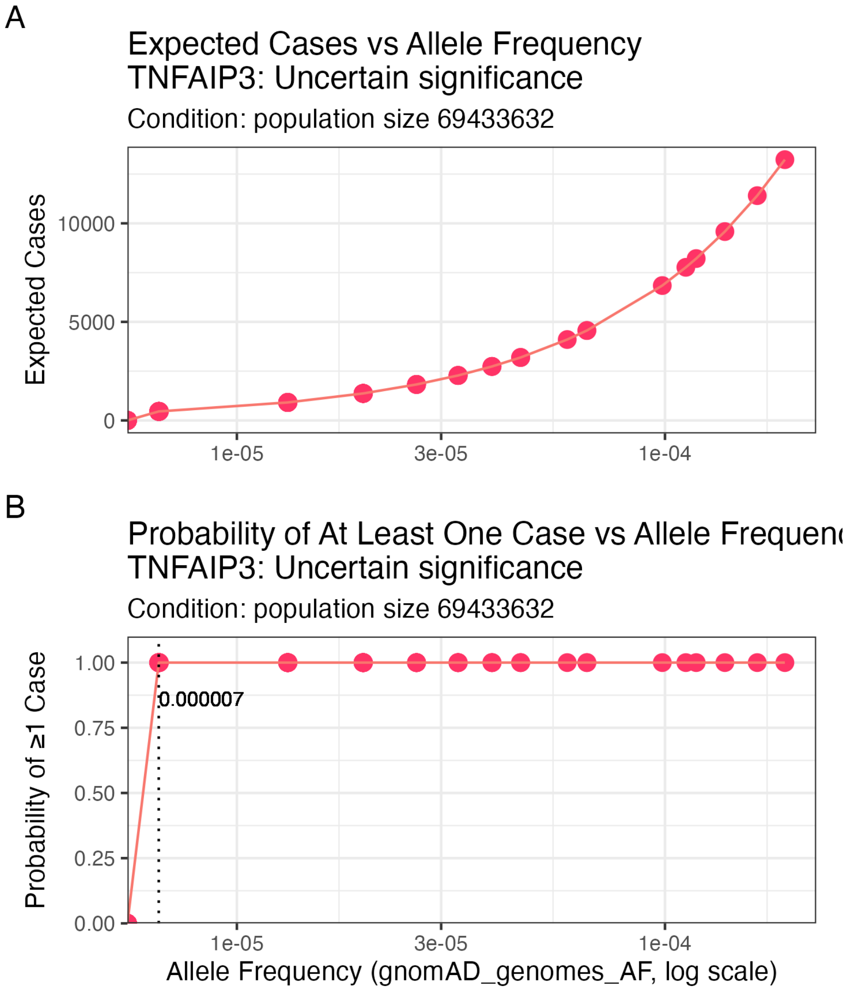


Figure 3: Scatter plots of (A) Expected Cases vs. Allele Frequency and (B) Probability of 1 Case vs. Allele Frequency for variants in *TNFAIP3* (classification "Uncertain Significance"). Dotted vertical lines indicate threshold allele frequencies corresponding to a probability of 1.

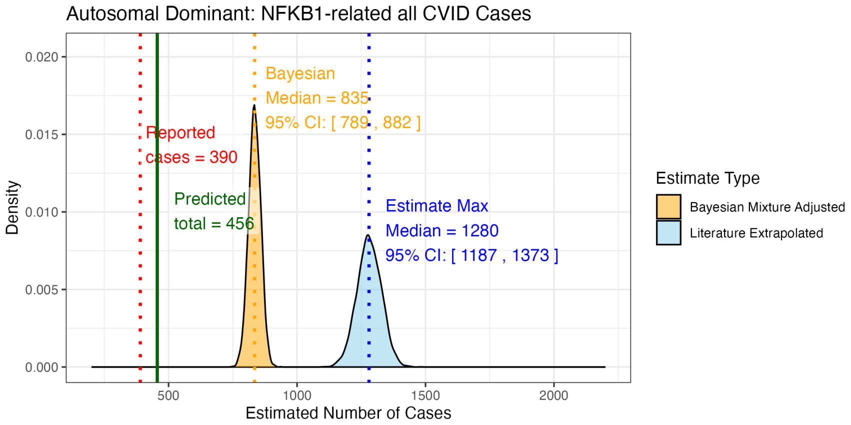


Figure 4: Density distributions for the estimated number of *NFKB1*-related CVID cases in the UK. The blue distribution represents the literature extrapolated estimates (median 1280), and the orange distribution shows the Bayesian mixture adjusted estimates (median 835, 95% CI: [789, 882]).

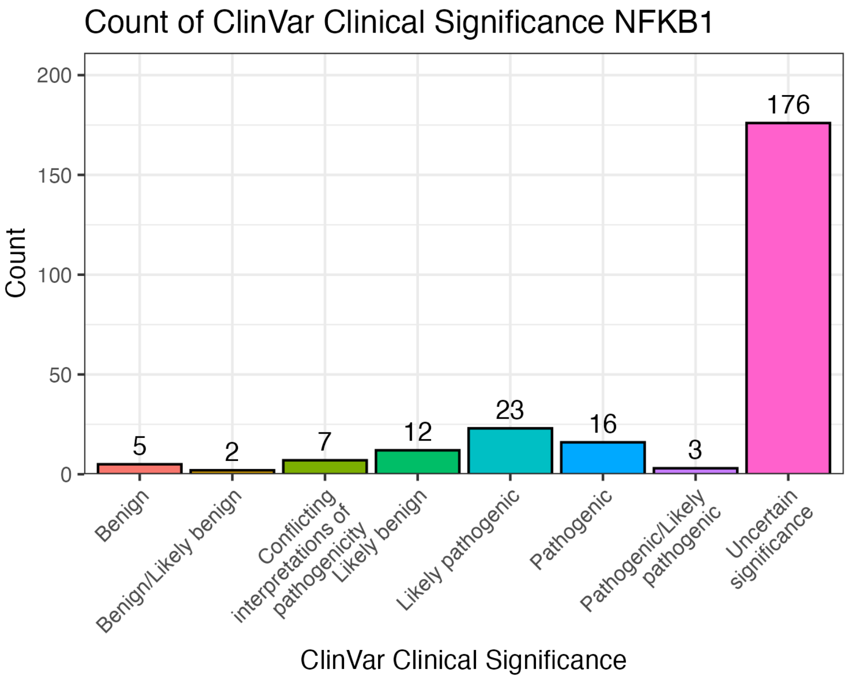


Figure 5: Count of ClinVar variant classifications for *NFKB1*.

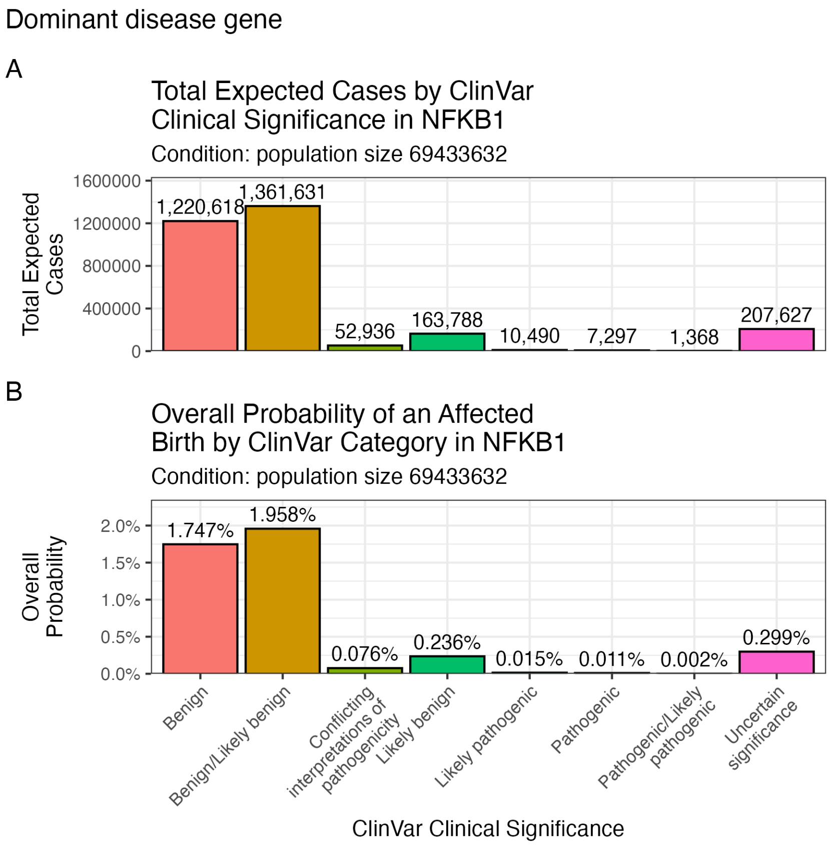


Figure 6: Combined bar charts of total expected cases and overall probability by ClinVar classification for *NFKB1* in the UK population.

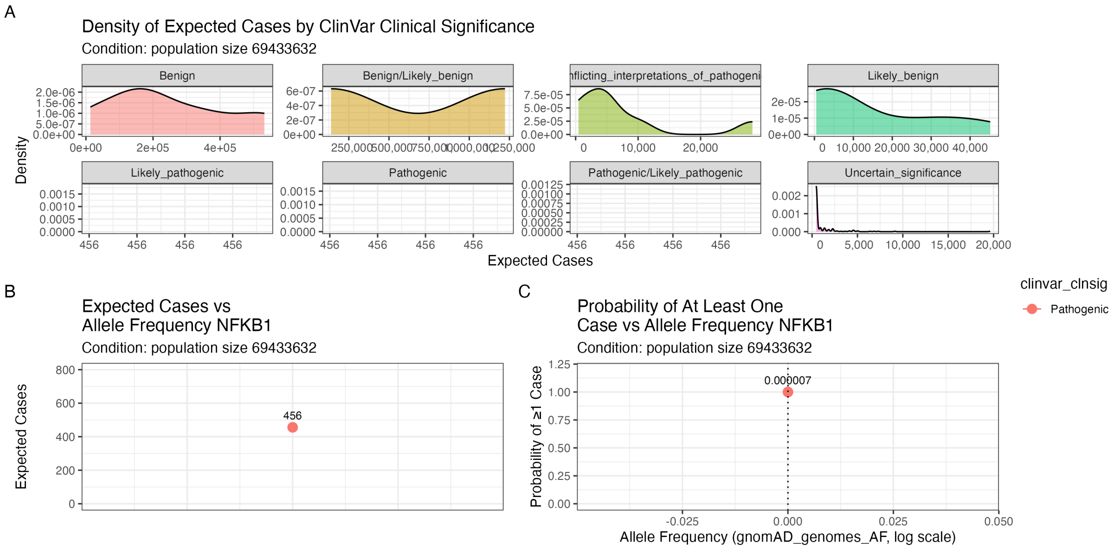


Figure 7: Density and scatter plots for *NFKB1*: (A) Expected Cases vs. Allele Frequency, and (B) Probability of 1 Case vs. Allele Frequency for the UK population.

4 Discussion

Our study demonstrates that classical genetic estimation using HWE yields accurate, genome-wide probabilities of disease occurrence for SNVs. These estimates form robust priors for Bayesian models currently under development. The integration of variant interpretation guidelines, such as those provided by the ACMG (**?**) and complementary frameworks (**? ?**), ensures that our estimates are clinically relevant.

In addition, standardised variant interpretation protocols that incorporate quality control and filtering criteria (**? ?**) are essential for robust genomic analysis. The concept of qualifying variants (QVs) (**? ?**) plays a central role in filtering and interpreting SNVs, while statistical approaches such as ACAT and SKAT (**? ? ? ?**) offer robust tools for aggregating variant effects in association studies. Moreover, multi-block data fusion techniques enable the integration of DNA, RNA, and proteomic data for comprehensive variant interpretation (**? ?**). Standardised reporting formats and unique identifiers for QV sets (e.g., using ACMG SF v3.2) enhance reproducibility and interoperability (**?**).

We acknowledge that our current approach focuses solely on SNVs and does not incorporate complex variants or de novo events. Future studies will address these limitations by incorporating additional models and data sources. The forthcoming Bayesian framework will integrate classical estimates with prior knowledge to update probabilities continuously, ultimately enhancing the precision of genetic diagnostics.

5 Clinical Feature Reclassification and Informative Value Analysis

A critical challenge in diagnosing primary immunodeficiency (PID) is the interpretation of clinical features, which are often derived from subjective and anecdotal evidence. For instance, detailed descriptions of T cell function in the IUIS IEI table can be highly variable and qualitative. We propose that reclassifying these subjective narratives into simpler, discrete categories—such as *low T cell count* or *T cell dysfunction*—can increase their utility in statistical models. Simplifying the data in this way not only reduces noise but also allows these features to be integrated more effectively into quantitative analyses.

To quantitatively assess the informativeness of these reclassified clinical features, we plan to compute the Fisher information for each feature. Fisher information, which measures the sensitivity of the likelihood function to changes in a parameter, is defined as:

$$I(\theta) \;=\; \mathbb{E}\Bigg[\Big(\frac{\partial}{\partial\theta}\ln f(Y;\theta)\Big)^2\Bigg].$$

For example, consider a model where the observed clinical measure Y is given by:

$$Y = \theta + W, \quad W \sim \mathcal{N}(0, \sigma^2),$$

with likelihood function:

$$f(y;\theta) \;=\; \frac{1}{\sqrt{2\pi}\,\sigma}\,\exp\!\Bigl(-\frac{(y-\theta)^2}{2\sigma^2}\Bigr).$$

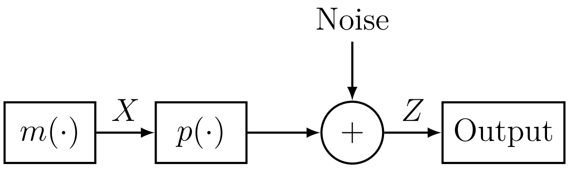
In this case, the Fisher information is:

$$I(\theta) \;=\; \frac{1}{\sigma^2}.$$

This example illustrates that a feature with lower noise (smaller σ^2) yields higher Fisher information, implying that it is more informative about the parameter θ.

In our context, θ could represent the effect size of a clinical feature (e.g., a measure of T cell dysfunction) on the probability of PID. Reclassifying subjective T cell information into objective categories is expected to reduce the measurement variance, thereby increasing the Fisher information score. This quantification will help determine which clinical features are most informative for PID diagnosis.

Figure 5 schematically illustrates the process of incorporating a clinical feature into our statistical framework, where noise influences the Fisher information.



In future work, we will extend our current probabilistic estimates—derived from ClinVar and gnomAD data—to include Fisher information analyses of reclassified clinical features. This approach will enable us to determine which features (e.g., simplified T cell data) provide the most informative signals for PID diagnosis, thereby guiding expert clinicians in their decision-making process.

6 Conclusions

We have developed and validated a framework for estimating the probability of disease-causing variants by integrating allele frequency data with clinically curated gene lists. Using *TNFAIP3* as a case study, a gene known to be implicated in primary immunodeficiency, we applied Hardy–Weinberg equilibrium principles to obtain precise estimates of the expected number of affected cases and the likelihood of observing at least one affected birth in a defined population. These results demonstrate the efficacy of our classical approach and provide a solid foundation for future Bayesian extensions, which will further refine these estimates by incorporating additional prior knowledge and new data. Ultimately, our methodology offers a reproducible and reliable reference for genetic risk estimation, with significant implications for clinical diagnostics and the interpretation of variant pathogenicity in PID.

7 Future Directions

Future studies will extend this framework to include complex variants and de novo events. Moreover, we will develop a Bayesian pipeline to compute posterior probabilities for variant causality, leveraging the classical estimates presented here as informative priors. Such advancements are expected to enhance the precision of genetic diagnostics and guide clinical decision-making in a rapidly evolving field.

8 Funding

This project was supported by the Swiss Personalized Health Network (SwissPedHealth) and the Strategic Focal Area 'Personalized Health and Related Technologies' of the ETH Domain.

9 Acknowledgements

We thank the contributors to dbNSFP, gnomAD, and IUIS IEI for providing access to high-quality genomic data, as well as the clinical and research teams involved in SwissPedHealth for their invaluable insights.

10 Contributions

DL designed the work and contributed to the manuscript. AS, MA, SB, VS, SÖ, and JA contributed to the manuscript. JF and LJS supervised the work and applied for funding.

11 Competing Interests

The authors declare no competing interests.