

Dylan Lawless@kispi.uzh.ch
Department of Intensive
Care and Neonatology,
University Children's
Hospital Zürich,
University of Zürich.

July 1, 2025

Dear Editors,

From Hardy-Weinberg to Quantifying Variant Causality

Mendelian genetics was rediscovered in 1900, revealing that inheritance operates through discrete units, now known as genes. Yet this insight initially provoked controversy because many traits, such as height, exhibit continuous variation rather than simple categorical patterns. Udney Yule (1902) argued that Mendelian genetics could not explain such continuous traits, fearing that dominant alleles would inevitably increase in frequency **10**.

The problem found its mathematical turning point through a series of pivotal contributions. William E. Castle (1903) demonstrated that without selection, genotype frequencies remain stable over generations **11**. Karl Pearson (1903) identified an equilibrium position for allele frequencies at $p = q = 0.5$ **12**. Reginald Punnett, confronted by Yule's arguments, brought the question to G. H. Hardy, a pure mathematician with little enthusiasm for applied problems. Hardy, famously dismissive of biological applications as "very simple" **13**, wrote his 1908 note in *Science* showing that under random mating, allele frequencies remain constant from generation to generation, provided no selection, mutation, migration, or drift occurs. Independently, Wilhelm Weinberg reached the same conclusion in Germany in 1908 **14**; **15**.

While Hardy may have considered the mathematics straightforward, the biological implications were profound and far-reaching. Hardy-Weinberg equilibrium (HWE) gave population genetics its statistical backbone, allowing genotype frequencies to be calculated from allele frequencies and establishing a baseline expectation for genetic variation in populations. Yet even with HWE, one fundamental problem remained unsolved: knowing how often certain genotypes *should* appear does not answer whether a particular genotype *causes disease* in an individual.

Throughout the 20th century, medical genetics advanced significantly, but the gap between population-level expectations and individual disease causality persisted. Clinicians and researchers developed variant classification systems to label variants as pathogenic, benign, or of uncertain significance. These labels, however, remained categorical and qualitative. They did not quantify the actual probability that a variant causes disease in a patient, nor did they account for causal variants that may be unobserved—either because they lie in regions not sequenced, because sequencing coverage failed, or because they remain undiscovered.

Thus, for more than a century after Hardy and Weinberg, a core question in genetics remained unsolved:

Given a candidate variant in a patient, what is the probability that it truly causes disease—and how certain can we be, considering the possibility of alternative, unobserved causal variants?

Our work addresses this challenge directly. We build upon the foundations laid by Hardy, Weinberg, and the architects of population genetics, and extend these principles into clinical genomics.

- HWE and classical population genetics supply the expected frequencies of genotypes in populations.
- Large-scale variant databases such as gnomAD and ClinVar provide empirical data on how often disease-causing and benign variants occur in the general population.
- Bayesian inference enables us to combine these population-based priors with patient-specific sequencing results, accounting explicitly for both observed variants and the possibility of unobserved pathogenic alleles.

Our framework integrates these elements to calculate *credible intervals (CrIs)* for the probability that any observed variant is causal. It quantifies:

- The likelihood that an observed variant is truly responsible for disease.
- The residual probability that other, unobserved variants could instead be the cause.

This approach moves the field beyond categorical variant labels to a rigorous, quantitative, evidence-aware interpretation. For the first time, clinicians and geneticists can assign a precise probability to the question of variant causality, with an explicit measure of uncertainty, grounded in both population genetics and empirical variant data.

We have demonstrated this framework in inborn errors of immunity (IEI) as a proof of concept, showcasing how it generates variant-level probabilities and integrates them with network-level biological data. However, the method is not limited to IEI; it is applicable to any genetic disease context, offering a quantitative foundation for variant interpretation, clinical decision-making, and the development of future learning systems in genomics.

In retrospect, Hardy’s “very simple” calculation turned out to be only the starting point of a century-long journey. The challenge of quantifying variant causality has required advances in statistical modelling, large-scale genomic databases, and clinical interpretation standards. Our work represents the next step in this progression: resolving a problem that has stood, unaddressed in quantitative terms, since the earliest days of modern genetics.

Yours sincerely,

Dylan Lawless, PhD
on behalf of all co-authors