blood **commentary**

CLINICAL TRIALS AND OBSERVATIONS

Comment on Stefanucci et al, *page 2055*

From black and white to fifty shades of grey

Andrew D. Paterson | The Hospital for Sick Children and University of Toronto

In this issue of *Blood,* **Stefanucci et al[1] used whole-exome sequencing generated from blood from UK Biobank participants to describe the magnitude of effects on both laboratory and clinical phenotypes for selected "pathogenic" or "likely pathogenic" rare coding variants in genes, of which some variants have previously been claimed to be causal for bleeding, thrombotic, or platelet disorders (BTPDs).**

First, some historical background as to how variants and genes were believed to be causal in individuals with specific BTPDs. Initially, genetic linkage analysis was performed in families with multiple affected individuals. These families were commonly referred to tertiary centers, likely had severe clinical disease, and were often selected to have a large number of affected relatives, and patterns of disease consistent with simple modes of inheritance (ie, autosomal dominant or recessive or X-linked recessive). After regions of the genome that segregate with the phenotype were identified, sequencing of coding regions of genes within the linked region(s) was performed, and coding variants that were bioinformatically predicted to change the amino acid sequence of the protein were identified and claimed to be causal for the disease. Sometimes, the frequency of such variants (minor allele frequency [MAF]) was examined in population cohorts, but often the sample size was limited.

Now, fast forward to the current era where next-generation sequencing, either whole-exome sequencing (all coding regions of all genes) or whole-genome sequencing, is performed to make a molecular genetic diagnosis in individuals affected with BTPDs. This is phenotype-first analysis:

individuals and families with a phenotype are studied genetically. Variants identified in probands are passed through bioinformatic pipelines, and community guidelines[2] are used to assess whether they are likely "pathogenic" for the presenting phenotype(s). It has been shown that when rare variants are identified in a modest number of small families with a disorder, even if they play a causal role, their penetrance estimates are grossly inflated.[3]

Although there are databases with MAFs obtained from multiple large cohorts (eg, Genome Aggregation Database[4,5]), they however contain no phenotypic information, severely limiting application to understand the relationships of specific variants to traits and diseases. Parallel community databases of variants identified from phenotype-first analysis, ClinVar (https://www.ncbi.nlm.nih.gov/clinvar),[6] are typically based on unrelated individuals, have sparse phenotype data, have no statistical analysis, and predictably result in inconsistencies in classifications for the same variant from different submitters.

So, how do these approaches compare with what is now possible from population cohorts, like the UK Biobank?[7]

Participants were invited and included without regard to phenotype, were extensively phenotyped across many domains, and were subjected to multiple detailed molecular analyses, specifically whole-exome sequencing. Variants present in subjects can be tested for association with numerous phenotypes using robust statistical methods.[1,8] This approach is in sharp contrast to the phenotype-first analysis (above) where the number of individuals with each variant is typically small, commonly just a few individuals from a single family, which results in limited power for statistical association analysis between each variant with phenotypes.

In the current work, variants in genes thought to be responsible for BTPDs were classified by multidisciplinary teams of clinicians, geneticists, and bioinformaticians, blind to clinical information, into 1 of the following: pathogenic/likely pathogenic, undecided, or rejected. Variants with MAF >0.1% were excluded. Those variants present in >5 individuals were subjected to statistical analysis for association with continuous platelet traits (count and mean volume), measured centrally at baseline. This analysis was restricted to 213 variants present in 46 genes (including *VWF, F8,* and *F9* for bleeding; *PROC, PROS1,* and *SERPINC1* for thrombosis; and 41 genes for platelet traits) present in 7432 participants.

The effects of 2 specific rare missense variants (1 each in *RUNX1* and *MECOM*), which both have >100 individuals with the minor allele, are shown (see figure). These results show that they both are not significantly associated with either platelet count or volume, contrary to previous claims of pathogenicity. In contrast, numerous other variants met statistical significance criteria, with the magnitude of effects on platelet and clinical traits shown, which are important for variant interpretation and clinical translation. This work provides both useful positive and negative data; some
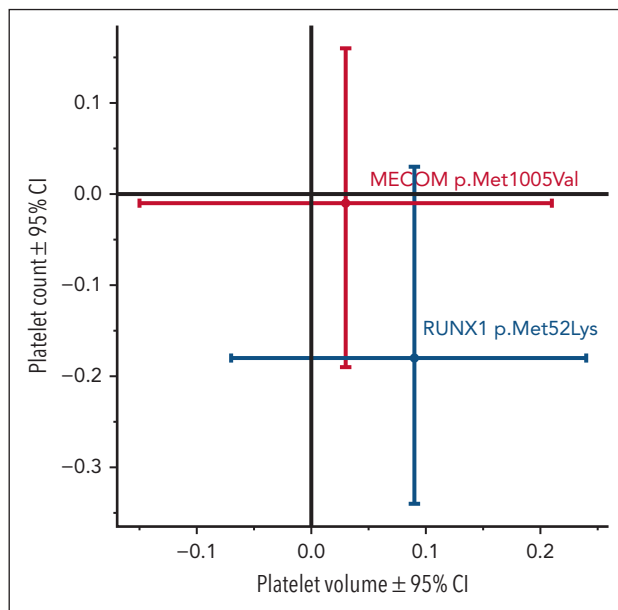
Effects of 2 variants on platelet count (y-axis) and mean platelet volume (x-axis). *RUNX1* rs200431130 p.Met52Lys (n = 164 individuals with minor allele) and *MECOM* rs200049869 p.Met1005Val (n = 127). Values are mean ± 95% confidence interval (CI), and units are SD. The figure has been adapted from Figure 3B in the article by Stefanucci et al that begins on page 2055.

variants that have previously been found in affected individuals may now have their role in the disease questioned, whereas for other variants, their role is strengthened. MacArthur et al noted, "Unambiguous assignment of disease causality for sequence variants is often impossible, particularly for the very low-frequency variants underlying many cases of rare, severe diseases."[9(p 469)]

Other questions that are addressed in this work include the following: What is the true mode of inheritance for specific variants? Loss-of-function variants in *MPL* have been reported to have an autosomal recessive mode of inheritance for congenital amegakaryocytic thrombocytopenia. However, they are paradoxically associated with *higher* platelet count than controls in (presumably) simple heterozygotes (carriers), consistent with overdominant inheritance. Also, the authors showed the additive effects of (poly)genetic (risk) scores on platelet traits in combination with some of the rare variants that they examined.

Limitations of the work include that UK Biobank subjects were aged 40 to 69 years at baseline and had a participation rate of 5.5%, with participants significantly healthier and wealthier and more likely women than the general population.[10] It is likely that individuals with BTPDs are underrepresented in the UK Biobank, and those who did participate have milder disease than the general population, which could tend to attenuate phenotypic associations.

A further limitation is that the authors restricted analysis to variants with MAF <0.1%, because they are more likely to be responsible for rare Mendelian diseases. This unfortunately resulted in the exclusion of several well-known variants (eg, *VWF* p.Thr1034del [rs368366214, MAF in African ancestry individuals=1.5%] and p.Tyr1584Cys [rs1800386, MAF in UK Biobank=0.43%]) for which insights about the association with traits would provide important information. Platelet function tests were absent, as were relevant protein levels (F8, F9, and VWF), but the latter are being generated from large-scale proteomics in the UK Biobank.

Are the findings from the UK Biobank specific to just hematology? No, similar findings have been previously reported for monogenic diabetes and developmental disorders,[8] among others, and are expected to be expanded to other disorders. But the data published here provide valuable insights into DNA variants in coagulation and platelet disorders.

Conflict-of-interest disclosure: The author declares no competing financial interests. ■

REFERENCES

1. Stefanucci L, Collins J, Sims MC, et al. The effects of pathogenic and likely pathogenic variants for inherited hemostasis disorders in 140 214 UK Biobank participants. *Blood.* 2023;142(24):2055-2068.

2. Richards S, Aziz N, Bale S, et al. 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.* 2015; 17(5):405-424.

3. Paterson AD, Seok S-C, Vieland VJ. The effect of ascertainment on penetrance estimates for rare variants: implications for establishing pathogenicity and for genetic counselling. *PLoS One.* 2023;18(9):e0290336.

4. Karczewski KJ, Francioli LC, Tiao G, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature.* 2020;581(7809):434-443.

5. Gudmundsson S, Singer-Berk M, Watts NA, et al. Variant interpretation using population databases: lessons from gnomAD. *Hum Mutat.* 2022;43(8):1012-1030.

6. Landrum MJ, Chitipiralla S, Brown GR, et al. ClinVar: improvements to accessing data. *Nucleic Acids Res.* 2020;48(D1): D835-D844.

7. Bycroft C, Freeman C, Petkova D, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature.* 2018;562(7726): 203-209.

8. Wright CF, West B, Tuke M, et al. Assessing the pathogenicity, penetrance, and expressivity of putative disease-causing variants in a population setting. *Am J Hum Genet.* 2019;104(2):275-286.

9. MacArthur DG, Manolio TA, Dimmock DP, et al. Guidelines for investigating causality of sequence variants in human disease. *Nature.* 2014;508(7497):469-476.

10. Fry A, Littlejohns TJ, Sudlow C, et al. Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population. *Am J Epidemiol.* 2017;186(9): 1026-1034.

https://doi.org/10.1182/blood.2023022171

© 2023 by The American Society of Hematology