Setting the Record Straight on Nucleus Origin

All substantive technical claims about the Nucleus Origin preprint made by @sichuan_mala are false.

STEPHAN CORDOGAN

NOV 23, 2025

As the lead author on Nucleus Origin, here is my response to the criticisms outlined in the blog post by sichuan_mala. All of their substantive technical claims are false, and sichuan_mala presumes that standard industry practices are novel developments made by a competitor. It's plausible that an unbiased individual unfamiliar with the field of statistical genetics could have made each of their errors individually, but not collectively.

Unlike our competitors, all of our model weights are public. Sichuan_mala refrained from validating any of the scores, as this would have immediately disproven most of their points. You can download them here, and we encourage the entire StatGen community to independently test our science.

Criticism 1: Identical polygenic score construction methods

First, sichuan_mala claims that the Origin preprint is a copy of a competitor's earlier preprint, in part because we construct PRSs and validate them within-family and across ancestries. This is unreasonable-validations within-family and across ancestries are standard for PRS publications, and best practice for testing PRSs for application to PGT-P data. Additionally, we meta-analyze summary statistics using inverse-variance weighting, and use SBayesRC to construct the PRS. Every new, flagship PRS publication for the past decade has used meta-analysis of summary statistics across different biobanks when available, as well as the best PRS software available, to create their PRS. Nothing about the competitor's approach was remotely novel.

Criticism 2: Identical comparisons to specific papers in the literature

Thompson and Mars are two public groups that have largely overlapping PRS offerings with the ones we trained. There are no other public research groups with recent, largely overlapping PRS offerings. As such, comparison to these two research groups is required to contextualize the relative gain in performance due to increased sample size and newer PRS methods. We refrained from comparing ourselves to direct competitors to avoid animosity within the industry (although this seems to have not been successful). A head-to-head comparison with the competitor would show that our scores overall perform modestly better,

likely due to our utilization of the AllofUs Biobank. We chose identical lifetime prevalence estimates so that individuals could make comparisons between liability R^2 values if they desired.

When lifetime disease prevalence for non-European ancestries were unavailable, they were estimated using European lifetime prevalence and odds ratios or hazard ratios. This was necessary to estimate the liability R^2 of our scores in non-European ancestries. In contrast, the competitor's paper does not attempt to validate any of their scores in non-European ancestries, instead scaling down relative performance proportionally to the genetic distance between the ancestry and Europeans. This is a reasonable approximation, but not as rigorous as direct validation.

Here is how lifetime prevalences were estimated using population prevalences to derive odds ratios.

Here is how lifetime prevalences were estimated using hazard ratios.

It would have been ideal to have good lifetime prevalence data for each ancestry, but this wasn't available for every disease, so approximations were sometimes necessary.

Criticism 3: Identical reference to the same 2025 paper for quality control

Removing low-quality SNPs for the construction of PRS is, again, standard practice. Most serious statistical geneticists have used GenomicSEM, have therefore read the 2025 paper hosted in their wiki on Github, and would revisit it when handling malformed or poor-quality summary statistics [1]. GenomicSEM is one of the most cited statistical genetics software packages, and Elliot Tucker-Drob is one of the most cited statistical geneticists.

Criticism 4: Potentially overlapping training and test sets

There is no overlap between the training and test cohorts. There was a typo in the preprint, although an unbiased reader would have realized this immediately, based on the remainder of the paragraph, and the fact that “training” was stated twice. The testing cohort refers to the sibling cohort, and the training cohort included all individuals unrelated to the sibling cohort (relatedness being defined as a King relatedness coefficient of 0.0442 or higher), who were used for GWAS. This means that there are no first, second, or third degree relations between the training and test cohort. The preprint’s typo will be corrected.

Additionally, the fact that standard errors in our competitor’s whitepaper are of similar sizes for all diseases regardless of their frequency suggests that there was either extreme cherrypicking of individuals to

construct their test cohort, which would introduce enormous bias, or that our competitor created different test cohorts for each disease, which is not standard practice. The effective sample size for binary diseases is highest when both outcomes are equally likely, and lowest when one outcome is much more likely, so the *effective* sample size for a common disease like hypertension is going to be much larger than a rare disease like Alzheimer's disease, with the same *total* sample size. This means that in a properly constructed validation cohort, the confidence interval for hypertension should be much smaller than the confidence interval for Alzheimer's. Unlike Nucleus, our competitor did not explain how their test cohorts were defined, and their weights are not public, so we cannot test either possibility.

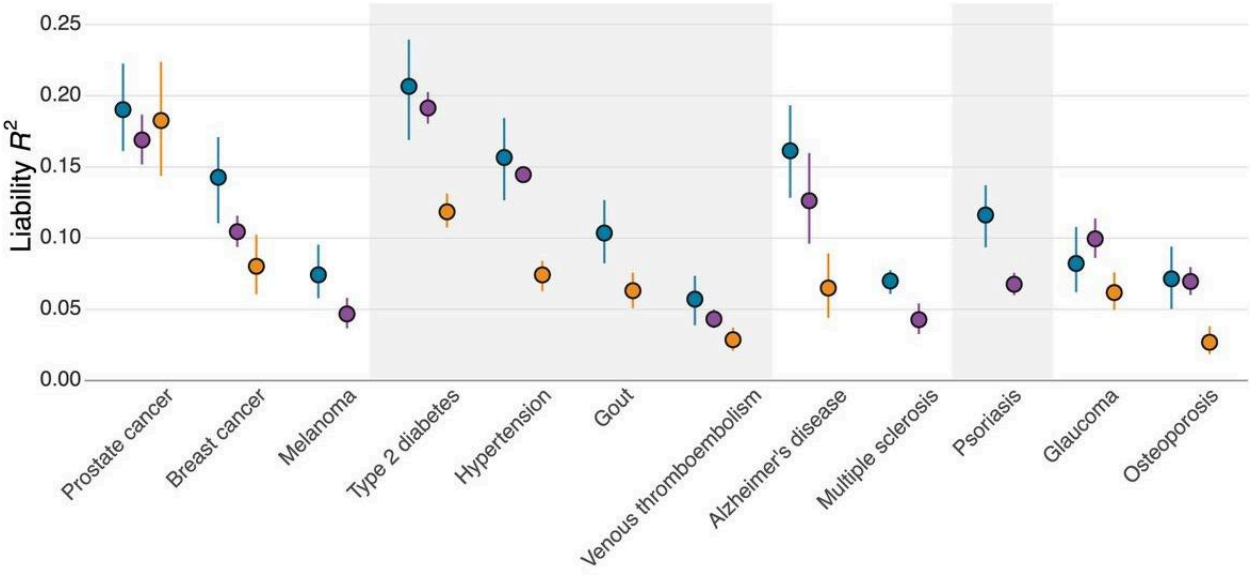

Figure 1. Our competitor's validation (blue), compared to two research groups. The size of the standard error bars vary with the frequency of the

disease across the research groups (purple and gold), but our competitor's don't, implying nonstandard validation practices [2].

Liability R^2 plots using points for estimates, and bars for 95% confidence intervals, are also standard within the industry, as shown below [3,4]. GGplot is a commonly used package for this purpose.

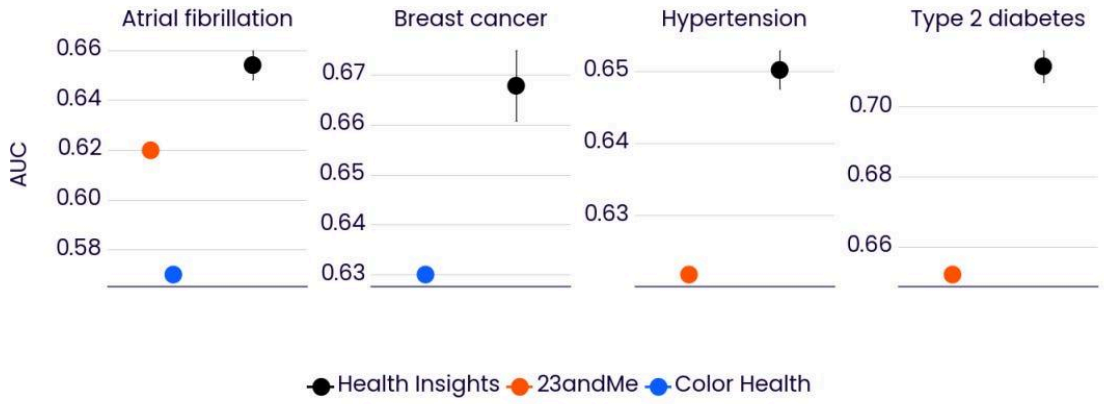

Figure 2. Plot from PRS publication **Derivation and validation of Health Insights polygenic risk scores and integrated risk tools.**

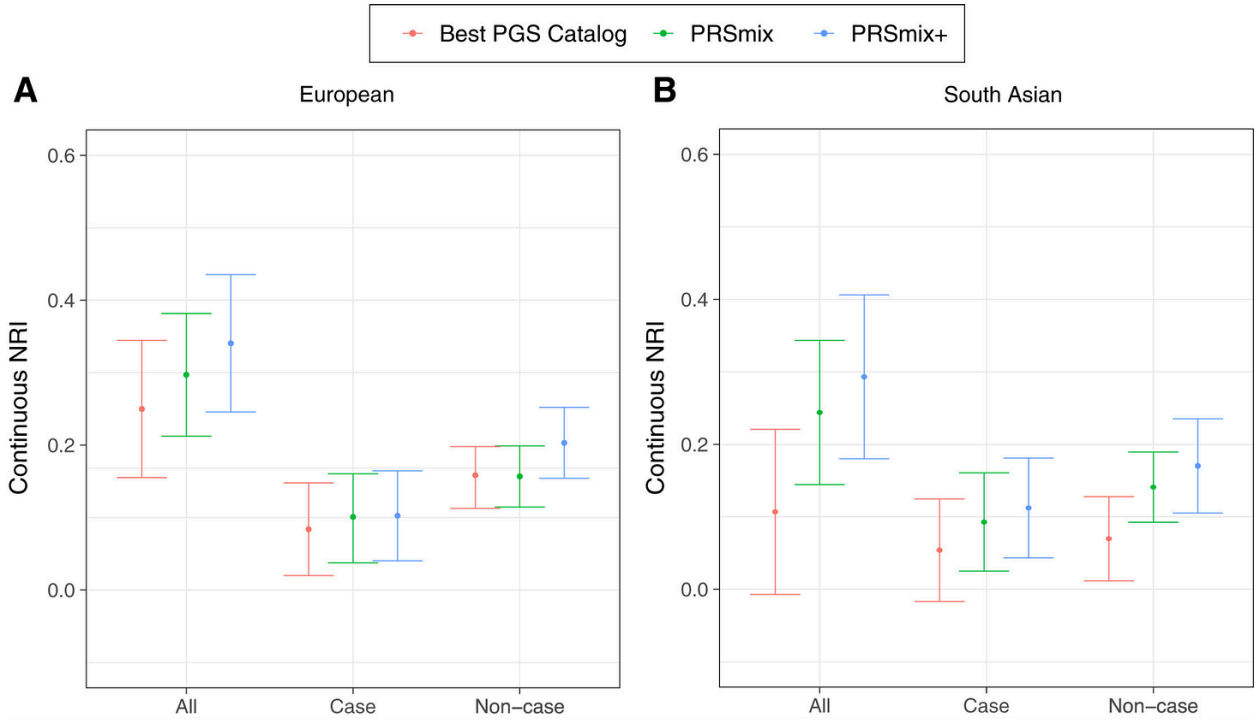

Figure 3. Plot from PRS publication **Integrative polygenic risk score improves the prediction accuracy of complex traits and diseases.**

Finally, I think sichuan_mala's claim about the non-independence between the sibling halves is valid, but this doesn't bias the liability R^2 estimates, only expands the confidence interval of the estimates slightly. We'll change this in the preprint, perhaps using the correlation coefficient with the phenotype and corresponding standard error from our population model of PRS performance.

Criticism 5: Nearly identical cohorts for within-family validation

We did not identify parent-offspring pairs as siblings. Although there are ~460,000 individuals of majority European ancestry within the UK Biobank, only ~409,000 of them self-reported as “Caucasian/White British”. Our competitor’s choice to restrict their analysis to self-reported Caucasians is unusual, but sichuan_mala assumed that it was standard. We found 40,862 siblings within the UK Biobank, and other analyses have found ~40,000 individuals within the UK Biobank sibling cohort when restricting to European ancestry as well [5]. Our competitor’s unusual choice to restrict their analysis to self-reported Caucasian/White British apparently decreased the size of their sibling cohort to 35,197. This is negligent by sichuan_mala- a cursory look at the self-reported Caucasian/White British field in the UK Biobank showcase would have explained this [6].

Criticism 6: Bizarre usage of “total blood pressure”

First, the use of total blood pressure instead of SDP or DBP separately is entirely logical considering that either can be used to make a hypertension diagnosis. This increased the total genetic signal in our AllofUs GWAS more than the use of SBP and DBP separately, due to highly overlapping genetic etiology between the two phenotypes. Additionally, offsetting measured blood pressure for medication usage is

standard practice, both for GWAS and non-genetic studies [7,8]. Here's an example from [7]:

"After calculating the mean SBP and DBP values from the two BP measurements, we adjust for medication use by adding 15 and 10 mmHg to SBP and DBP, respectively, for individuals reported to be taking BP-lowering medication (21.4% of individuals)"

Even if this was a novel approach, it is completely intuitive. Had sichuan_mala searched GWAS literature more diligently, they would have discovered that this was standard practice.

Criticism 7: Incorrect ICD9 code used for prostate cancer

Incorrect ICD codes were not used in the actual validation, and if they were, they would have *underestimated* the power of the PRS. Typos were made in the supplementary table, and will be corrected.

Conclusion

The substantive criticisms of the Origin paper by sichuan_mala are without merit. Considering the existing tensions within the industry, I'm not surprised that much of X jumped to amplify this article. However, I'm particularly disappointed in several individuals who have downloaded the scores for personal or research use, but still joined the dogpile. They

could have tested the scores to investigate sichuan_mala's claims of incorrect case/control designations or training/testing sample overlap, and disproved them themselves. Instead, they took these claims at face value. I will be happy to field any further questions about Origin models, both on Substack or X.