Background

Participating in a study is rarely random and is often driven by traits that predispose individuals to volunteer. Such ascertainment can bias associations detected in the study between various exposures and outcomes. Existing methods to detect ascertainment typically require phenotype records, individual-level genotypes, or large samples of related individuals. Here, we introduce a summary-statistic based method to detect and quantify ascertainment in large-scale biobanks.

Methods and Results

Our approach estimates a parameter, θ, which captures the deviation in the mean polygenic score of an ascertained sample relative to its expectation in a non-ascertained reference. We propose two estimators, $\hat{\theta}_1$ and $\hat{\theta}_2$, and evaluated their performance through extensive simulations. We show that $\hat{\theta}_1$ is more powerful but sensitive to residual stratification, and reference misspecification, whereas $\hat{\theta}_2$ is robust to these factors, albeit with slightly reduced power. Applying $\hat{\theta}_2$ across 21 complex traits and diseases in 7 large-scale biobanks from Europe, the Middle East, East Asia, and Latin America, we found significant evidence of ascertainment that are consistent with known recruitment strategies of some biobanks. For example, participation in the Australian Genomics of Depression Study was driven by individuals with higher genetic susceptibility for major depression disorder $\left(\text{MDD: } \hat{\theta}_2 = 0.43 \pm 0.07, P = 2.9 \times 10^{-09}\right)$ and post-traumatic stress disorder $\left(\text{PTSD: } \hat{\theta}_2 = 0.23 \pm 0.08, P = 1.8 \times 10^{-03}\right)$, reflecting recruitment based on psychiatric traits. We also uncovered previously unrecognized patterns in the Mexican (MDD, PTSD, asthma), Taiwan (asthma), and Qatar (anxiety) biobanks.

Conclusions

Our findings demonstrate that ascertainment can be reliably quantified from summary data, providing scalable way to characterise participation bias across biobanks. This approach can be readily integrated into large-scale genetic studies to improve interpretation of biobank findings and mitigate bias in downstream analyses.