Liability-scale heritability estimation for biobank studies of low prevalence disease

Sven E. Ojavee^{1,2,*}, Zoltan Kutalik^{1,2,3}, Matthew R. Robinson^{4,*}

1 Department of Computational Biology, University of Lausanne, Lausanne, Switzerland.

2 Swiss Institute of Bioinformatics, Lausanne, Switzerland

3 University Center for Primary Care and Public Health, University of Lausanne, Lausanne, Switzerland

4 Institute of Science and Technology Austria, Klosterneuburg, Austria.

*corresponding author: svenerik.ojavee@unil.ch, matthew.robinson@ist.ac.at

Abstract

Theory for liability-scale models of the underlying genetic basis of complex disease provides an important way to interpret, compare and understand results generated from biological studies. In particular, liability models facilitate an understanding and comparison of the relative importance of genetic and environmental risk factors that shape different clinically important disease outcomes, through estimation of the liability-scale heritability (LSH). Increasingly, large-scale biobank studies that link genetic information to electronic health records are becoming available, containing hundreds of disease diagnosis indicators that mostly occur infrequently within the sample. Here, we propose an extension of the existing liability-scale model theory suitable for estimating LSH in biobank studies of low-prevalence disease. In a simulation study, we find that our derived expression yields lower MSE and is less sensitive to prevalence misspecification as compared to previous transformations, for diseases with $\leq 2\%$ population prevalence and LSH of ≤ 0.45 , especially if the biobank sample prevalence is less than that of the wider population. Applying our expression to 13 diagnostic outcomes of $\leq 3\%$ prevalence in the UK Biobank study, revealed important differences in LSH obtained from the different theoretical expressions, that impact the conclusions made when comparing LSH across disease outcomes. This demonstrates the importance of careful consideration for estimation and prediction of low prevalence disease outcomes, and facilitates improved inference of the underlying genetic basis of $\leq 2\%$ population prevalence diseases, especially where biobanking sample ascertainment results in a healthier sample population.

Introduction

Genetically informed deep-phenotyped biobanks are an increasingly available important research resource. From these data, estimates of SNP heritability, h_{SNP}^2 , can be obtained, a quantity describing the proportion of phenotypic variance attributable to the genetic marker data [1]. Linked electronic health records, provide a large number of binary, presence/absence disease diagnosis indicators and it is important to be able to compare h_{SNP}^2 estimates, in order to infer the relative importance of genetic and environmental risk factors that shape different clinically important disease outcomes.

To better describe the genetics of such binary traits, the notion of liability scale heritability (LSH) has been coined [2] to reflect the underlying continuous nature of additive genetic effects. An initial derivation for LSH was given by Alan Robertson in the Appendix of Dempster and Lerner [3] for the scenario where the case-control ratio was the same in the sample and the population. Lee et al. [4], proposed an extended derivation to account for the fact that, in a case-control study, cases tend to be over-represented compared to the population prevalence, arriving at the following expression for the LSH

$$h_{liab}^2 = h_{obs}^2 \frac{K(1-K)}{\varphi(\Phi^{-1}(K))^2} \frac{K(1-K)}{P(1-P)}, \quad (1)$$

where, h_{obs}^2 is the observed scale heritability, K is the prevalence of the binary trait in the full population, P is the prevalence of the binary trait in the sampled subpopulation, and the denominator of the first fraction is the squared probability density function of the standard normal distribution evaluated at the K th quantile of the inverse cumulative density function of the standard normal distribution. This expression was derived under the assumption that sample prevalence is greater than or equal to population prevalence ($P \geq K$).

The problem of accurately estimating LSH was further investigated by Golan et al. [5] who noted that in the common setting of sample prevalence exceeding population prevalence ($P > K$), Eq.1 applied on REML estimates underestimates LSH. To account for this, they proposed phenotype correlation-genotype correlation (PCGC) regression, which generalises a Haseman-Elston regression and yields unbiased estimates of LSH in simulations. In many settings, especially if there are individual-level data available and the sole goal of the analysis is to estimate the LSH, it is recommended to use PCGC. However, often only summary-level marginal SNP regression coefficients are provided by biobank studies, meaning that h_{SNP}^2 must be indirectly inferred (e.g. using LD score regression [6] or High-definition likelihood [7]), or observed scale heritability estimate is a hyper-parameter embedded into effect size estimation (BayesRR-RC [8]). Therefore, it remains important to facilitate the transformation of observed scaled heritability into LSH.

The structure of the biobank data sets creates additional problems. Firstly, they often represent a subset of the population that often is healthier, younger, or has higher socio-economic background than the average population. Because of this, many if not most binary disease traits have a lower sample prevalence compared to the population prevalence ($P < K$). The classical expression in Eq.1 has been derived and tested for situations where $P \geq K$, and this can result in estimates with inflated variance for situations where $P < K$. Secondly, the disease prevalence in biobank-scale studies can be small and measurement errors of such a small quantity could greatly amplify the variance of the LSH estimate. In practice, as biobanks include people born during different decades, cohorts have usually not reached the end of their lifespan and as disease prevalences change across time, it seems appropriate to accompany the prevalence estimates with error estimates when arriving at an LSH estimate. Furthermore, although Eq.1 has solid theoretical justification, it does not guarantee that the LSH estimate will be lower than or equal to 1, and even if the LSH estimate is bounded, it can still have high variance in the biobank setting, especially if the model is imperfectly specified or there are greater deviations away from the assumption of the existence of latent genetic liability.

Here, we propose an alternative expression to address some of those issues. Firstly, the suggested expression will guarantee that the LSH estimate will be bounded between 0 and 1, and secondly, we demonstrate that for low prevalences ($K \leq 0.02$) our formula results in lower MSE compared to the classical expression. Thirdly, we show that our formula limits the inflation in MSE if we take into account the uncertainty in the prevalence estimation. Although our expression is also derived under the simplifying assumption of $P = K$, we argue that for many biobank-based studies this assumption is preferred as long as $P \leq K$ and the prevalence is small. Finally, we apply our proposed expression on 13 disease outcomes with low sample prevalence in the UK Biobank and we compare the LSH estimates obtained by our expression to those of Eq.1. We also provide a shiny app <https://medical-genomics-group.shinyapps.io/h2liab/>.

Derivation of the expression

Suppose that we have a binary trait with a frequency of K in the population and a frequency of P in the sample is here equal to K . Suppose that a liability model holds meaning that there exists a latent liability l that is defined as a sum of genetic (g) and error (e) components $l = g + e$. We assume that l has a variance of 1 and that g and e come from normal distributions: $g \sim N(0, h_l^2)$, $e \sim N(0, 1 - h_l^2)$, where h_l^2 is the LSH.

Then we assume that the binary disease trait y is associated with l such that $y = 1$ if $l > t$ and $y = 0$ if $l \leq t$, where t is some liability scale threshold defining the required liability value for disease occurrence, and $t = \Phi^{-1}(1 - K)$.

We first write the expression how is the observed scale heritability associated with LSH. As shown by Dempster and Lerner [3]

$$h_o^2 = \frac{h_l^2 \varphi(\Phi^{-1}(K))^2}{K(1 - K)}, \quad (2)$$

where h_o^2 is the observed scale heritability. We recognise that the numerator represents the observed scale genetic variance and the denominator represents total observed scale phenotypic variance. Our idea is to replace the total phenotypic variance estimate $K(1 - K)$ with the sum of genetic and error variances and that by definition will guarantee that the LSH estimate remains bounded. Thus, we need to rewrite the total phenotypic variance using the error variance $E(\text{Var}(y|c + zg))$

$$\text{Var}(y) = h_l^2 \varphi(\Phi^{-1}(K))^2 + E(\text{Var}(y|c + zg)). \quad (3)$$

We find (see Supplementary Information) that the error variance is expressed as

$$E(\text{Var}(y|c + zg)) = K - \tilde{\Phi}(\Phi^{-1}(K), \Phi^{-1}(K), h_l^2), \quad (4)$$

where $\tilde{\Phi}(x_1, x_2, \rho)$ is the cumulative distribution function of a bivariate normal distribution with a mean of 0, variance of 1 and correlation ρ evaluated at (x_1, x_2) . That gives us a the expression for observed scale heritability

$$h_o^2 = \frac{h_l^2 \varphi(\Phi^{-1}(K))^2}{h_l^2 \varphi(\Phi^{-1}(K))^2 + K - \tilde{\Phi}(\Phi^{-1}(K), \Phi^{-1}(K), h_l^2)} = g(h_l^2). \quad (5)$$

It is impossible to find the closed expression for h_l^2 from the last expression, however, we can still plug in values for K and h_o^2 and solve Eq.5 numerically for h_l^2 .

We derive the variance for h_l^2 from the last expression using the delta method. Suppose that we know $\text{Var}(h_l^2)$ and we would like to find $\text{Var}(h_o^2)$. For this, we need to differentiate expression 5 with respect to h_l^2 and that results in

$$\begin{aligned} \frac{d}{dh_l^2} g(h_l^2) &= \frac{d}{dh_l^2} \left(1 + \frac{K - \tilde{\Phi}(-t, -t, h_l^2)}{h_l^2 \varphi(-t)^2} \right)^{-1} = \\ &= - \left(1 + \frac{K - \tilde{\Phi}(-t, -t, h_l^2)}{h_l^2 \varphi(-t)^2} \right)^{-2} \frac{-\tilde{\Phi}'(-t, -t, h_l^2) h_l^2 \varphi(-t)^2 - \varphi(-t)^2 (K - \tilde{\Phi}(-t, -t, h_l^2))}{(h_l^2 \varphi(-t)^2)^2} = \\ &= \frac{\tilde{\Phi}'(-t, -t, h_l^2) h_l^2 \varphi(-t)^2 + \varphi(-t)^2 (K - \tilde{\Phi}(-t, -t, h_l^2))}{(h_l^2 \varphi(-t)^2 + K - \tilde{\Phi}(-t, -t, h_l^2))^2}, \end{aligned} \quad (6)$$

where we have denoted $\Phi^{-1}(K) = -t$ (negative of the disease defining threshold), $\tilde{\Phi}'(-t, -t, h_l^2)$ is the partial derivative of $\tilde{\Phi}(-t, -t, h_l^2)$ with respect to h_l^2 and as demonstrated by Drezner and Wesolowsky [9]

$$\tilde{\Phi}'(-t, -t, h_l^2) = \frac{1}{2\pi\sqrt{1 - (h_l^2)^2}} \exp\left(-\frac{t^2}{1 + h_l^2}\right). \quad (7)$$

The variance can thus be expressed from the delta method as

$$\text{Var}(h_o^2) = \text{Var}(g(h_l^2)) \approx \left(\frac{d}{dh_l^2} g(h_l^2) \right)^2 \text{Var}(h_l^2) \quad (8)$$

and conversely for the h_l^2

$$Var(h_l^2) \approx \left(\frac{d}{dh_l^2}g(h_l^2)\right)^{-2}Var(h_o^2). \quad (9)$$

Simulation study

We executed two simulation settings. Simulation 1 followed the strategy of Lee et al. [4] and we used it to demonstrate the implications of the uncertainty of prevalence estimates to the outcome in the absence of ascertainment. There, we used a sample size of 20,000 and we created a low level of relatedness by simulating individuals in independent batches of 100 with genetic values (g) drawn from a multivariate normal distribution given a 100×100 covariance matrix with off-diagonal elements in the covariance matrix were $0.05h_l^2$ and h_l^2 for the diagonals. Error term e was simulated from a normal distribution $N(0, \sigma_e^2)$ such that the variance of the liability $l = g + e$ would be $Var(l) = h_l^2 + \sigma_e^2 = 1$. The corresponding observed binary phenotype was defined as

$$y = \begin{cases} 0, & \text{if } l < \Phi^{-1}(1 - K) \\ 1, & \text{if } l \geq \Phi^{-1}(1 - K) \end{cases} \quad (10)$$

We varied the true LSH from 0.15 to 0.65 with a step of 0.1 and we varied prevalence between $K = 0.001, 0.005, 0.01, 0.05, 0.1, 0.5$. In addition, to mimic the potential uncertainty coming from the heterogeneity across estimates, at each step when estimating the LSH, we drew the value for $\hat{K}$ from a normal distribution $N(K, (0.1K)^2)$ yielding a coefficient of variation of 10%. The number of simulation replicates was 2000.

For low prevalence traits, our proposed formula results in a more accurate result in terms of MSE in many common settings where the LSH is smaller than 0.55 and the prevalence less than 2% (Figure 1). However, we note caution that if the underlying LSH is higher than 0.45, then bounding the LSH estimate below 1 results in a small bias that is observed as the prevalence approaches 1% or greater (Figure 1), but we expect this to be the exception for real-world disease outcomes. Using an inaccurately measured prevalence results in increased MSE, with the new proposed formula being more robust to this kind of misspecification, as it yields lower MSE increase across all scenarios (Figure 3).

Simulation 2 examined a more realistic scenario similarly to Golan et al. [5]. The final sample size was again 20,000 but here we created the genetic values using SNP data of 10,000 markers. For each of the 10,000 SNPs, we simulated the SNP MAF from a uniform distribution $U(0.05, 0.5)$ and the minor allele counts x_{ij} for individual i at SNP j were simulated from a binomial distribution $B(2, MAF_j)$. The effect size for each SNP j was drawn from a normal distribution such that $\beta_j \sim N(0, \frac{h_l^2}{10,000})$. The genetic value for an individual i g_i was calculated as $g_i = x_i' \beta$ and the error term e_i was simulated from $N(0, \sigma_e^2)$ such that the variance of the liability $l_i = g_i + e_i$ would be $Var(l_i) = h_l^2 + \sigma_e^2 = 1$. The liability scale phenotype l_i was translated into binary phenotype y_i as shown in Eq.10. To create suitably ascertained samples we first simulated a slightly larger population of $\tilde{N}(>20,000)$ that has a case prevalence (in population) of K and then we randomly selected $N \cdot P$ cases and $N \cdot (1 - P)$ controls to achieve case prevalence (in sample) of P . We used GREML [10] from GCTA [11] to estimate the observed scale heritability. Finally, we used the Eq.1 to transform it to the known LSH estimate and we used Eq.5 by using P as the prevalence to yield the new LSH estimate. We used the same values for heritability, we varied the K between 0.01, 0.02 and 0.05, and we varied P between 50%, 75%, 90%, 95%, 100% of the K value. Due to the increased computational complexity, we resorted to 100 simulation replicates.

Here, we find the new expression of Eq.5 gives lower MSE at lower population prevalence ($K \leq 0.02$) and with moderate or low LSH values ($LSH \leq 0.45$) as compared to the classical expression of Eq.1, supporting the results of the first simulation setting (Figure 4). Therefore, we propose that for biobank studies of diseases with $\leq 2\%$ population prevalence, especially where the sample prevalence is less than that of the wider population, Eq.5 should be preferred over Eq.1.

Empirical data analysis

We then analysed 13 ICD-10 binary disease outcomes with $\leq 3\%$ sample prevalence in UK Biobank (Table 1) with our recently proposed GMRM software, to obtain observed scale heritability estimates, accounting for marker effect size differences across SNPs of different minor allele frequencies, linkage disequilibrium and functional annotation [8, 12], using 2,174,071 SNP markers and 382,390 individuals of SNP marker relatedness (≤ 0.05). To attempt to control for environmental confounding before analysis, we adjusted for the leading 20 principle components as supplied by the UK Biobank, sex as a binary factor, age as a linear and quadratic term, east and north coordinate of residence, recruitment centre, and genotype batch. We estimated the posterior mean observed scale heritability from the last 1,500 sampling iterations after stabilisation of the running mean, and then compared the liability scale estimates produced with either using the classical expression of Eq.1 or our new proposed expression of Eq.5.

Liability-scale heritability estimates obtained by Eq.5 were lower than the classical expression of Eq.1 (Figure 4). For disease outcomes, where we assume that the UK Biobank sample prevalence is identical to the wider population prevalence, the estimates from either equation are in agreement (Figure 4). However, once the sample prevalence was $\leq 1\%$, and lower than that estimated in the wider population from which it was drawn, we observe substantially lower liability-scale estimates from Eq.5 (Figure 4). For example, LSH differed by 0.26 for endometriosis and 0.22 for macular degeneration although this difference is mainly driven by the difference in population and sample prevalences. Even though we assumed equal sample and population prevalence for hypertensive renal disease, we still get a difference of 0.14 between the classical and new estimates accompanied by two times narrower 95% CI for the new estimate (Table 1). These differences influence the inference made, as Eq.5 estimates no significant difference of LSH for glaucoma, endometriosis, and macular degeneration, in contrast to the results that would be obtained from Eq.1 (Figure 4). Therefore, the lower MSE of our proposed estimate across many common settings translates to real-world differences in inference for low prevalence diseases within biobank studies.

Discussion

Using biobank-scale data sets to infer the heritability of rare traits is increasingly important, due to the general difficulties of collecting case-control samples for rare diseases. However, our results demonstrate the importance of treating low prevalence traits with extra care. Moreover, the advantage of using our proposed estimate is even greater if we take into account the uncertainty or heterogeneity in the lifetime prevalence estimates. Heterogeneity in the lifetime prevalence estimates is very common and stems from regional differences, differences in the time of the study, methodology, usage of subpopulations, and often the lifetime prevalence estimate is simply unavailable or some proxies are used. All this amounts to a considerable amount of uncertainty and unfortunately, the most common way to handle this is to simply pick one of the prevalence estimates and not take into account the uncertainty. Our proposed estimate limits the error coming uncertainty and we argue that the future analyses using lifetime population prevalence should reflect the uncertainty in the estimates.

The analysis of empirical data in UK Biobank calculating SNP heritability demonstrated that in many settings the classical expression could lead to likely, if not clear overestimates. For example for hypertensive renal disease, the classic formula gave an LSH estimate of 0.82 (95% CI 0.65, 0.96) whereas the new formula estimated LSH of 0.68 (95% CI 0.58, 0.74). For a similar trait of chronic kidney disease, the heritability values have been estimated between 0.30 to 0.75 [13] suggesting that the classical estimate might be unrealistically high. We also observe a similar effect for macular degeneration where the classical formula also yields a likely overestimate of 0.62 (95% CI 0.49, 0.78) whereas the literature suggests SNP heritability of [14] of 0.47 that had been calculated on a larger case count.

In general, there seems to be a switch point in prevalence after which the classical expression (Eq.1) tends to become more effective. That is probably because $K(1 - K)$ is a good estimator for the total phenotypic variance with high K values (as in Eq.2) but with small K values that product $K(1 - K)$ becomes tiny and the expressions using the inverse $\frac{1}{K(1-K)}$ will become highly sensitive to the observed scale heritability estimation error. Our proposed expression makes the total phenotypic variance dependent on the estimated observed genetic variance and thus controls better for the mismatch between the total phenotypic and genotypic variances in the classical expression.

There are important caveats to our proposed formula. Even though we manage to effectively constrain the heritability between 0 and 1, it also introduces a small downwards bias that becomes more visible with higher values of prevalence and true liability scale heritability (>0.6). Nevertheless, we argue that this will likely be the exception for real-world disease outcomes. Any transformation of scale will be an approximation, made under a set of theoretical assumptions and our aim here is to simply provide an approach that facilitates comparisons of the proportion of variance attributable to the SNP markers for low prevalence diseases with as low MSE as possible. Secondly, our formula assumes that the biobank prevalence is approximately reflecting the population prevalence. In most low prevalence scenarios we believe this approach is justified since the sample prevalence is similar or slightly smaller compared to the population prevalence, the latter case resulting in more conservative or lower MSE estimates (Figure 4). In principle, it is possible to extend the Eq 5 to account for ascertainment. However, the outcome when expressing total phenotypic variance as a sum of error and genetic variances would still depend on products of $K(1 - K)$ and $P(1 - P)$ and thus this approach would not have any benefit in terms of reducing MSE.

Here, we have proposed a new expression for calculating LSH suitable for traits with low prevalence. We have shown that our proposed formula results in a more accurate LSH estimator in terms of MSE in many common settings and in general results in a slightly more conservative estimate that can result in more accurate estimates of liability scale heritability. Hopefully, it can lead to a more realistic quantification of rare trait heritabilities, many of which are still yet to be explored.

Acknowledgements

This project was funded by an SNSF Eccellenza Grant to MRR (PCEGP3-181181), and by core funding from the Institute of Science and Technology Austria. This research was supported by the Scientific Service Units (SSU) of IST Austria through resources provided by Scientific Computing (SciComp).

Author contributions

SEO and MRR conceived and designed the study. SEO derived the equations. ZK provided study oversight. SEO and MRR analysed the data and wrote the paper. All authors approved the final manuscript prior to submission.

Author competing interests

The authors declare no competing interests.

Figure 1. MSE of previous and new formula in the case of no ascertainment and no measurement error for the prevalence across different values for the liability-scale heritability. The new estimate yields lower MSE in the settings where the prevalence is low ($K < 0.01$). The benefit of using the new formula decreases as the underlying liability-scale heritability increases, stemming from the small bias introduced in the new estimate that is also bounding the estimate below 1. Data sets created under simulation scenario 1.

Figure 2. Increase in MSE due to measurement error of prevalence. Using prevalence that has not been completely accurately measured results in increase in the mean squared error. However, the new proposed formula is more robust to this kind of misspecification as it yields lower MSE increase across all scenarios. Data sets created under simulation scenario 1.

Figure 3. Differences in MSE if sample prevalence is lower than the population prevalence. The new expression works better at lower population prevalences ($K \leq 0.02$) and with moderate or low LSH values ($h_l^2 \leq 0.45$). Data sets created under simulation scenario 2.

Figure 4. Comparison of previous and new liability-scale heritability estimates with 95% CI for 13 ICD-10 binary disease outcomes with $\leq 3\%$ sample prevalence in the UK Biobank. Table 1 gives the full disease names, codes, sample prevalence, population prevalence, observed-scale estimates, and liability-scale conversions. We analysed the 13 traits with our recently proposed GMRM software, to obtain observed scale heritability estimates, accounting for marker effect size differences across SNPs of different minor allele frequency, linkage disequilibrium and functional annotation in 382,390 unrelated UK Biobank individuals. We estimated the posterior mean observed scale heritability and plot the liability scale estimates produced with either using the classical expression of Eq.1 on the x-axis or our new proposed expression of Eq.5 on the y-axis.

Disease	ICD-10 code	Sample prevalence	Estimated population prevalence	h^2_{obs} 95% CI	New SNP- h^2_{liab} 95% CI	Classic SNP- h^2_{liab} 95% CI
Carpal Tunnel Syndrome	G56	2.8%	5.0% [15]	0.038 (0.031, 0.044)	0.242 (0.201, 0.279)	0.291 (0.240, 0.338)
Chronic obstructive pulmonary disease	J44	2.6%	4.6% [16]	0.036 (0.029, 0.043)	0.244 (0.198, 0.286)	0.291 (0.235, 0.344)
Oesophagitis	K20	2.5%	2.5%*	0.026 (0.020, 0.032)	0.180 (0.138, 0.220)	0.182 (0.139, 0.223)
Iron deficient anaemia	D50	2.4%	5.5% [17]	0.021 (0.015, 0.027)	0.151 (0.112, 0.191)	0.192 (0.141, 0.244)
Atherosclerosis	I70-I79	2.3%	2.3%*	0.025 (0.020, 0.031)	0.191 (0.153, 0.234)	0.193 (0.154, 0.239)
Osteoporosis	M80-M82	2.1%	5.1% [18]	0.028 (0.022, 0.034)	0.219 (0.175, 0.263)	0.284 (0.227, 0.344)
Cellulitis	L03	2.1%	2.1%*	0.023 (0.017, 0.029)	0.186 (0.141, 0.229)	0.188 (0.141, 0.233)
Endometriosis	N80	1.8%	10.0% [19]	0.043 (0.034, 0.051)	0.367 (0.297, 0.430)	0.634 (0.504, 0.758)
Acute Renal Failure	N17	1.8%	1.8%*	0.022 (0.018, 0.028)	0.197 (0.156, 0.248)	0.199 (0.157, 0.253)
Glaucoma	H40	1.4%	2.5% [20]	0.030 (0.025, 0.037)	0.311 (0.256, 0.371)	0.370 (0.302, 0.448)
Macular degeneration	H35.3	0.8%	3.4% [21]	0.025 (0.020, 0.032)	0.408 (0.327, 0.495)	0.624 (0.491, 0.783)
Hyperthyroidism	E05	0.5%	0.8% [22]	0.027 (0.022, 0.032)	0.550 (0.466, 0.626)	0.656 (0.535, 0.786)
Hypertensive renal disease	I12	0.4%	0.4%*	0.027 (0.021, 0.032)	0.679 (0.581, 0.743)	0.817 (0.650, 0.961)

Table 1. 13 selected disease outcomes recorded in the UK Biobank of <3% sample prevalence with higher prevalence in the general UK population. Columns of the table give the commonly used disease name, the ICD-10 code, the UK Biobank sample prevalence, and the estimated UK population prevalence with corresponding reference (* denotes that we used the UK Biobank due to unavailability or vast heterogeneity in the estimates), the posterior mean 0/1 observed scale single nucleotide polymorphism (SNP) heritability with 95% credible interval.

References

1. Yang, J., Zeng, J., Goddard, M. E., Wray, N. R. & Visscher, P. M. Concepts, estimation and interpretation of snp-based heritability. *Nature genetics* **49**, 1304–1310 (2017).
2. Falconer, D. S. The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Annals of human genetics* **29**, 51–76 (1965).
3. Dempster, E. R. & Lerner, I. M. Heritability of threshold characters. *Genetics* **35**, 212 (1950).
4. Lee, S. H., Wray, N. R., Goddard, M. E. & Visscher, P. M. Estimating missing heritability for disease from genome-wide association studies. *The American Journal of Human Genetics* **88**, 294–305 (2011).
5. Golan, D., Lander, E. S. & Rosset, S. Measuring missing heritability: Inferring the contribution of common variants. *Proceedings of the National Academy of Sciences* **111**, E5272–E5281 (2014).
6. Bulik-Sullivan, B. K. *et al.* Ld score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature genetics* **47**, 291–295 (2015).
7. Ning, Z., Pawitan, Y. & Shen, X. High-definition likelihood inference of genetic correlations across human complex traits. *Nature genetics* **52**, 859–864 (2020).
8. Patxot, M. *et al.* Probabilistic inference of the genetic architecture underlying functional enrichment of complex traits. *Nature Communications* **12**, 1–16 (2021).
9. Drezner, Z. & Wesolowsky, G. O. On the computation of the bivariate normal integral. *Journal of Statistical Computation and Simulation* **35**, 101–107 (1990).
10. Yang, J. *et al.* Common snps explain a large proportion of the heritability for human height. *Nature genetics* **42**, 565–569 (2010).
11. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. Gcta: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics* **88**, 76–82 (2011).
12. Orliac, E. J. *et al.* Maximizing gwas discovery and genomic prediction accuracy in biobank data. *bioRxiv* (2021). <https://www.biorxiv.org/content/early/2021/08/12/2021.08.12.456099.full.pdf>.
13. Cañadas-Garre, M. *et al.* Genetic susceptibility to chronic kidney disease – some more pieces for the heritability puzzle. *Frontiers in Genetics* **10** (2019).
14. Fritsche, L. G. *et al.* A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants. *Nature genetics* **48**, 134–143 (2016).
15. Middleton, S. D. & Anakwe, R. E. Carpal tunnel syndrome. *BMJ* **349** (2014).
16. Rayner, L., Sherlock, J., Creagh-Brown, B., Williams, J. & deLusignan, S. The prevalence of copd in england: An ontological approach to case detection in primary care. *Respiratory Medicine* **132**, 217–225 (2017).
17. Fairweather-Tait, S. Iron nutrition in the uk: Getting the balance right. *The Proceedings of the Nutrition Society* **63**, 519–28 (2004).
18. Svedbom, A. *et al.* Osteoporosis in the european union: a compendium of country-specific reports. *Archives of osteoporosis* **8**, 1–218 (2013).

19. Zondervan, K. T., Becker, C. M. & Missmer, S. A. Endometriosis. *New England Journal of Medicine* **382**, 1244–1256 (2020). PMID: 32212520.
20. Allison, K., Patel, D. & Alabi, O. Epidemiology of glaucoma: The past, present, and predictions for the future. *Cureus* **12** (2020).
21. Pennington, K. L. & DeAngelis, M. M. Epidemiology of age-related macular degeneration (amd): associations with cardiovascular disease phenotypes and lipid factors. *Eye and vision* **3**, 1–20 (2016).
22. Garmendia Madariaga, A., Santos Palacios, S., Guillén-Grima, F. & Galofré, J. C. The Incidence and Prevalence of Thyroid Dysfunction in Europe: A Meta-Analysis. *The Journal of Clinical Endocrinology & Metabolism* **99**, 923–931 (2014).

Supplementary Information

Derivation of the error variance term

Suppose that the genetic value (of an individual) g and error term e are from normal distributions $g \sim N(0, h_l^2)$ and $e \sim N(0, 1 - h_l^2)$, where h_l^2 is the underlying liability scale heritability. Then the underlying liability $l = g + e$ and the binary trait y with a prevalence of K is defined as

$$y = \begin{cases} 0, & \text{if } l < \Phi^{-1}(1 - K) \\ 1, & \text{if } l \geq \Phi^{-1}(1 - K) \end{cases} \quad (11)$$

From this we will derive the error variance term $E(\text{Var}(y|c + zg))$ where c is some constant and z is the standard Gaussian density evaluated at $\Phi^{-1}(1 - K)$ as shown in [3]. As $c + zg$ is a linear combination of g we can equivalently find $E(\text{Var}(y|g))$. First, we note the conditional distribution of y given g

$$\begin{array}{c|cc} y & 0 & 1 \\ \hline P(y|g) & P\left(\frac{e}{\sqrt{1-h_l^2}} < \frac{\Phi^{-1}(1-K)-g}{\sqrt{1-h_l^2}}\right) = \Phi\left(\frac{\Phi^{-1}(1-K)-g}{\sqrt{1-h_l^2}}\right) & P\left(\frac{e}{\sqrt{1-h_l^2}} \geq \frac{\Phi^{-1}(1-K)-g}{\sqrt{1-h_l^2}}\right) = \Phi\left(\frac{g-\Phi^{-1}(1-K)}{\sqrt{1-h_l^2}}\right) \end{array} \quad (12)$$

As y can be equal to only 0 or 1, we can write

$$\begin{aligned} \text{Var}(y|g) &= E(y^2|g) - E(y|g)^2 = E(y|g) - E(y|g)^2 = P(y = 1|g) - P(y = 1|g)^2 = \\ &= \Phi\left(\frac{g - \Phi^{-1}(1 - K)}{\sqrt{1 - h_l^2}}\right) - \Phi\left(\frac{g - \Phi^{-1}(1 - K)}{\sqrt{1 - h_l^2}}\right)^2. \end{aligned} \quad (13)$$

To find $E(\text{Var}(y|g))$ we need to find $E(\Phi\left(\frac{g - \Phi^{-1}(1 - K)}{\sqrt{1 - h_l^2}}\right))$ and $E(\Phi\left(\frac{g - \Phi^{-1}(1 - K)}{\sqrt{1 - h_l^2}}\right)^2)$. For this, we use auxiliary standardised Gaussian random variables X , X_1 and X_2 that are independent of g and X_1 is independent of X_2 . From this it follows that $\text{Var}(X\sqrt{1 - h_l^2} - g) = 1$ and using the law of total probability we get

$$\begin{aligned} E\left(\Phi\left(\frac{g - \Phi^{-1}(1 - K)}{\sqrt{1 - h_l^2}}\right)\right) &= P\left(X \leq \frac{g - \Phi^{-1}(1 - K)}{\sqrt{1 - h_l^2}}\right) = P(X\sqrt{1 - h_l^2} - g \leq -\Phi^{-1}(1 - K)) = \\ &= \Phi(-\Phi^{-1}(1 - K)) = 1 - \Phi(\Phi^{-1}(1 - K)) = K. \end{aligned} \quad (14)$$

Secondly, we see that we can analogously use X_1 and X_2 to find the second moment of $\Phi\left(\frac{g - \Phi^{-1}(1 - K)}{\sqrt{1 - h_l^2}}\right)$. For this we need to find the following correlation

$$\text{cor}(X_1\sqrt{1 - h_l^2} - g, X_2\sqrt{1 - h_l^2} - g) = E((X_1\sqrt{1 - h_l^2} - g)(X_2\sqrt{1 - h_l^2} - g)) = E(g^2) = h_l^2. \quad (15)$$

Now we express the expectation using a cumulative distribution function of a bivariate Gaussian distribution of two random variables that have a correlation of h_l^2

$$\begin{aligned} E\left(\Phi\left(\frac{g - \Phi^{-1}(1-K)}{\sqrt{1-h_l^2}}\right)^2\right) &= E\left(P\left(X_1 \leq \frac{g - \Phi^{-1}(1-K)}{\sqrt{1-h_l^2}}\right)P\left(X_2 \leq \frac{g - \Phi^{-1}(1-K)}{\sqrt{1-h_l^2}}\right)\right) = \\ E\left(P\left(X_1 \leq \frac{g - \Phi^{-1}(1-K)}{\sqrt{1-h_l^2}}, X_2 \leq \frac{g - \Phi^{-1}(1-K)}{\sqrt{1-h_l^2}}\right)\right) &= P\left(X_1 \leq \frac{g - \Phi^{-1}(1-K)}{\sqrt{1-h_l^2}}, X_2 \leq \frac{g - \Phi^{-1}(1-K)}{\sqrt{1-h_l^2}}\right) \\ P\left(X_1 \sqrt{1-h_l^2} - g \leq -\Phi^{-1}(1-K), X_2 \sqrt{1-h_l^2} - g \leq -\Phi^{-1}(1-K)\right) &= \\ \tilde{\Phi}(-\Phi^{-1}(1-K), -\Phi^{-1}(1-K), h_l^2) &= \tilde{\Phi}(\Phi^{-1}(K), \Phi^{-1}(K), h_l^2), \quad (16) \end{aligned}$$

where $\tilde{\Phi}(x_1, x_2, \rho)$ is the cumulative distribution function of a standardised bivariate Gaussian distribution with a correlation of ρ . The first equation follows from the definition of cumulative distribution function, second from the independence of X_1 and X_2 , third from the law of total probability. Thus, by combining the two last results, we get the final expression for the error variance

$$E(\text{Var}(y|c + zg)) = K - \tilde{\Phi}(\Phi^{-1}(K), \Phi^{-1}(K), h_l^2). \quad (17)$$