Cascade screening following a polygenic risk score test: what is the disease risk of a sibling conditional on a high score in a proband?

Shai Carmi[1]

[1] Braun School of Public Health and Community Medicine, The Hebrew University of Jerusalem, Jerusalem, Israel

shai.carmi@huji.ac.il

Abstract

Polygenic risk scores (PRSs) for predicting disease risk have become increasingly accurate, leading to increasing popularity of PRS tests. Consider an individual whose PRS test has placed him/her at the top q-quantile of genetic risk. Recently, Reid et al. (Circ Genom Precis Med. 2021;14:e003262) have investigated whether such a finding should motivate cascade screening in the proband's siblings. Specifically, using data from the UK biobank, Reid et al. computed the empirical probability of a sibling of the proband to also have a PRS at the top q-quantile. In this short note, I use the liability threshold model to compute this probability analytically, showing excellent agreement with the empirical results of Reid et al., including that this probability is disease-independent. Further, I compute the probability of a sibling of the proband to be affected, as a function of the quantile threshold q, the proportion of variance explained by the score, and the disease prevalence.

Introduction

Polygenic risk scores show great promise for personalized disease risk prediction. A polygenic risk score (PRS) for a disease is a count of the number of risk alleles carried by an individual, with each allele weighted by its effect size. For complex diseases, a PRS represents the cumulative risk introduced by thousands or more variants, each of a small effect[1]. PRSs were empirically shown to explain a substantial proportion of the variance in disease liability, and individuals at the highest PRS percentiles were shown to have risk that can be even ≈3x higher compared to the population mean[2-4]. This suggests the feasibility personalized prevention and/or intervention in these individuals, whose high risk may not always manifest as traditional risk factors[5].

In clinical genetics, cascade screening refers to the testing of relatives of individuals who were either found to have a disease with a genetic component or were found to carry a disease mutation[6]. Cascade screening is important, as it allows high-specificity identification of individuals at high genetic risk. Cascade screening is currently limited to severe diseases and single mutations of large effects[7]. However, the development of increasingly accurate polygenic risk scores[8], along with the increasing number of individuals receiving PRS results worldwide[9], raise the prospects of performing cascade screening already upon the finding of an extreme PRS.

Recently, Reid et al. have investigated the question of cascade screening following a PRS finding[10]. They empirically studied the following setting. Suppose an individual has taken a PRS test that has placed him/her at the top q-quantile of the PRS distribution (where q can be, e.g., 1%, 5%, 10%, etc.). What is

the probability of a sibling of the proband to also have PRS at the top q-quantile? Using data on ≈23,000 sib-pairs from the UK biobank, Reid et al. have shown that the siblings of those probands have an increased probability to have a high PRS compared to the population average. Further, the probability was about the same across four diseases.

Despite these interesting results, several questions remained open. Specifically, it is unclear (1) whether the empirical results of Reid et al. (and particularly, the independence across diseases) are concordant with predictions of quantitative genetics theory; (2) what is the expected risk for diseases or threshold quantiles not empirically studied; and (3) what is the expected risk of the sibling to become affected, rather than just have a PRS above a cutoff. The latter is important, as the cost-effectiveness of cascade screening depends on the actual disease risk rather than just the PRS. Here, I use the liability threshold model, an established model of disease risk in quantitative genetics, to address these questions.

Methods

Model

The liability threshold model (LTM) is a classic quantitative genetic theory that relates the risk of a disease to underlying genetic and non-genetic factors[11,12]. According to the LTM, a disease has an underlying continuous liability, distributed as a standard normal random variable. The liability can be written as $y = g + \epsilon$, where $g \sim N(0, h^2)$ represents polygenic genetic factors (with variance equal to the heritability h^2) and $\epsilon \sim N(0,1 - h^2)$ represents non-genetic (environmental) factors. An individual is affected whenever his/her liability exceeds a threshold. For a disease with prevalence K, the threshold is z_K, the upper-K quantile of the standard normal distribution. (For example, for $K = 0.01$, $z_K = 2.33$.) The LTM was found to fit well genetic data on complex diseases and is widely applied[13-19].

In our setting, we do not know the precise value of the genetic factors influencing the liability of an individual. Instead, we have an estimate represented by the PRS. We thus write the liability as $y = s + e$, where s is the PRS and e is the residual liability, representing non-modeled genetic factors and non-genetic factors[17,19,20]. We assume that the PRS is normally distributed and has been standardized, such that $s \sim N(0, r^2)$, where r^2 is the variance in liability explained by the PRS. Consequently, $e \sim N(0,1 - r^2)$.

Consider next a pair of (full) siblings. Using standard quantitative genetic theory, it can be shown that that the PRSs of the two sibs can be written as $s_1 = c + x_1$ for sib 1, and $s_2 = c + x_2$ for sib 2. In these equations, c is a genetic component shared between the sibs, and is equal to the average maternal and paternal PRSs. Its distribution across the population is $c \sim N(0, r^2/2)$. Then, $x_1 \sim x_2 \sim N(0, r^2/2)$ are two *independent* genetic components. For details on the derivation, see our previous publications[21,22]. For an intuitive explanation, the "segregation" variance, i.e., the variance of any polygenic component across siblings due to the randomness of meiosis, is known to be half the variance in the population[23]. Thus, given the parental PRSs, the PRS of each child has variance $r^2/2$.

We next define a threshold above which an individual is designated as having "high PRS". We define the threshold as the upper-q quantile of the PRS distribution. For example, if $q = 0.01$, an individual is considered to havehigh PRS if his/her PRS is at the top 1% of PRSs across the population. The PRS has zero mean and variance r^2 in the population, and thus, the value of the PRS at the threshold is $z_q r$,

where z_q is the upper-q quantile of the standard normal variable. We are told that sib 1 (the proband) has a high PRS, i.e., $s_1 > z_q r$. We would like to compute the conditional probability that sib 2 either also has high PRS, or is affected.

Our calculations are similar to those of So et al.[17] and Do et al.[19], who have considered the problem of predicting disease risk based on the PRS of an individual and/or a relative, along with the disease status of the relative. Here, we assume the disease status of the proband is unknown (e.g., for a late-onset disease).

The probability that the sibling has high PRS

We would like to compute the probability $P(s_2 > z_q r \mid s_1 > z_q r)$. Using the definition of the conditional probability,

$$(1)\quad P(s_2 > z_q r \mid s_1 > z_q r) = \frac{P(s_1 > z_q r \cap s_2 > z_q r)}{P(s_1 > z_q r)} = \frac{1}{q} P(s_1 > z_q r \ \cap \ s_2 > z_q r).$$

To compute the numerator, we condition on c. Given c, the scores of the two sibs are independent, i.e.,

$$(2)\quad P(s_1 > z_q r \cap s_2 > z_q r \mid c) = P(s_1 > z_q r \mid c) P(s_2 > z_q r \mid c).$$

For $i = 1,2$,

$$(3)\quad P(s_i > z_q r \mid c) = P(c + x_i > z_q r) = P(x_i > z_q r - c) = 1 - \Phi\left(\frac{z_q r - c}{r/\sqrt{2}}\right) = 1 - \Phi\left(\sqrt{2} z_q - \frac{c}{r/\sqrt{2}}\right).$$

In Eq. (3), $\Phi(x)$ is the cumulative distribution of the standard normal variable, and we used the fact that $x_i \sim N(0, r^2/2)$. We can now compute the desired probability (Eq. (1)) by substituting Eq. (3) into Eq. (2), and integrating over all c. Recalling that $c \sim N(0, r^2/2)$,

$$(4)\quad P(s_2 > z_q r \mid s_1 > z_q r) = \frac{1}{q} \int_{-\infty}^{\infty} \frac{\phi\left(\frac{c}{r/\sqrt{2}}\right)}{r/\sqrt{2}} \left[1 - \Phi\left(\sqrt{2} z_q - \frac{c}{r/\sqrt{2}}\right)\right]^2 dc.$$

In Eq. (4), $\phi(x)$ is the density of the standard normal variable. We now change variables, $t = c/(r/\sqrt{2})$, and obtain

$$(5)\quad P(s_2 > z_q r \mid s_1 > z_q r) = \frac{1}{q} \int_{-\infty}^{\infty} \phi(t) \left[1 - \Phi(\sqrt{2} z_q - t)\right]^2 dt.$$

Eq. (5) is our final result for the probability of the sibling to have high PRS. Note that the probability does not depend on r^2, the variance explained by the PRS, and thus is disease- and PRS- independent.

The probability that the sibling is affected

Denote by y_1 and y_2 the liabilities of the two sibs, and recall that sib 2 will be affected if $y_2 > z_K$, where z_K is the upper-K quantile of the standard normal distribution and K is the prevalence. We would like to compute the following probability,

(6) $P(y_2 > z_K \mid s_1 > z_q r) = \frac{P(y_2 > z_K \cap s_1 > z_q r)}{P(s_1 > z_q r)} = \frac{1}{q} P(y_2 > z_K \cap s_1 > z_q r)$.

As above, we write $s_i = c + x_i$ for $i = 1,2$, and condition on c. First, we note that

(7) $P(y_2 > z_K \cap s_1 > z_q r \mid c) = P(y_2 > z_K \mid c) P(s_1 > z_q r \mid c)$,

because knowledge of the score of sib 1, given c, is not informative on the remaining liability components of sib 2. As above,

(8) $P(s_1 > z_q r \mid c) = 1 - \Phi\left(\sqrt{2} z_q - \frac{c}{r/\sqrt{2}}\right)$.

Next,

(9) $P(y_2 > z_K \mid c) = P(s_2 + e_2 > z_K \mid c) = P(c + x_2 + e_2 > z_K) = P(x_2 + e_2 > z_K - c) \equiv P(\tilde{e}_2 > z_K - c) = 1 - \Phi\left(\frac{z_K - c}{\sqrt{1 - r^2/2}}\right)$.

We defined $\tilde{e}_2 \equiv x_2 + e_2$, and used the fact that x_2 and e_2 are independent normals, such that $\tilde{e}_2$ is normal with zero mean and variance $\mathrm{Var}(\tilde{e}_2) = r^2/2 + (1 - r^2) = 1 - r^2/2$. We can compute the desired probability (Eq. (6)) by substituting Eqs. (8) and (9) into Eq. (7), and integrating over all c,

(10) $P(y_2 > z_K \mid s_1 > z_q r) = \frac{1}{q} \int_{-\infty}^{\infty} \frac{\phi\left(\frac{c}{r/\sqrt{2}}\right)}{r/\sqrt{2}} \left[1 - \Phi\left(\sqrt{2} z_q - \frac{c}{r/\sqrt{2}}\right)\right] \left[1 - \Phi\left(\frac{z_K - c}{\sqrt{1 - r^2/2}}\right)\right] dc$.

We can again change variables, $t = c/(r/\sqrt{2})$, and obtain

(11) $P(y_2 > z_K \mid s_1 > z_q r) = \frac{1}{q} \int_{-\infty}^{\infty} \phi(t) \left[1 - \Phi(\sqrt{2} z_q - t)\right] \left[1 - \Phi\left(\frac{z_K - tr/\sqrt{2}}{\sqrt{1 - r^2/2}}\right)\right] dt$.

This is our final expression for the probability of the sibling of the proband to be affected. Here, the probability depends on r^2, as well as on the prevalence K.

R implementation

We solved the integrals in Eqs. (5) and (11) numerically using the function `integrate` in R. Our code is as follows.

```
risk_sib_high_prs = function(q) {
  integrand = function(t)
    return(dnorm(t)*pnorm(qnorm(1-q)*sqrt(2)-t, lower.tail=F)^2 / q)
  return(integrate(integrand,-Inf,Inf)$value)
}

risk_sib_high_affected = function(q,K,r) {
  zq = qnorm(1-q)
  zK = qnorm(1-K)
  integrand = function(t) {
    arg1 = zq*sqrt(2)-t
```

```
    arg2 = (zK-t*r/sqrt(2)) / (sqrt(1-r^2/2))
    return(dnorm(t)*pnorm(arg1,lower.tail=F)*pnorm(arg2,lower.tail=F) / q)
  }
  return(integrate(integrand,-Inf,Inf)$value)
}
```

Results and discussion

Reid et al.[10] have first studied the correlation between the PRSs of relatives. They found that the correlation between the PRSs of siblings and second-degree relatives was ≈0.5 and ≈0.25, respectively. These correlations are naturally expected: based on standard quantitative genetic theory, the correlation between the genetic values of relatives is equal to their coefficient of relatedness[12,24].

Next, Reid et al. have computed the empirical probability (across sib pairs in the UK biobank), that, given that the proband has PRS at the top q quantile, that a sibling of the proband also has PRS at the top q quantile. Specifically, they considered four diseases (atrial fibrillation, coronary artery disease, diabetes, and severe obesity) and four values of q (1%, 5%, 10%, and 20%).

We used the liability threshold model to derive an analytical expression for the probability of the sibling of the proband to have PRS at the top q quantile (Methods; Eq. (5)). We compare our theoretical predictions to the empirical observations of Reid et al. in Figure 1A, showing excellent agreement. Our theory also predicts that the risk of the sibling is independent of the disease and the accuracy of the PRS, as empirically observed by Reid et al. Our figure also provides the expected risk generally for values of q in the range 0-25%.

For studying the cost-effectiveness of cascade screening, it is necessary to estimate the risk of the sibling in the case screening is *not* applied. To this end, we need to compute the risk of the sibling *to* be *affected*, conditional on the proband having a high PRS. We used the liability threshold model to derive an analytical expression for this probability (Methods; Eq. (11)). We plot the results in Figure 1B, for a representative value of the prevalence (5%), and for four values of the proportion of variance explained by the PRS. The results show that, as expected, the risk of the sibling is always higher than compared to a random individual from the population. Also as expected, the risk increases with increasing accuracy of the score, and with a higher PRS of the proband (i.e., a smaller percentile used to define the top of the PRS distribution).

Our simple R code will allow researchers to substitute any value for the top-PRS threshold, the accuracy of the score, and the disease prevalence in order to compute the expected outcomes of cascade screening in any setting of interest. A similar analytical approach could be used to compute the risk given other types of relatives, such as a child or a cousin of the proband. Further conditioning on the disease status of the proband can be incorporated as in [17,19]. We expect our results to form a necessary building block for future studies of the cost-effectiveness of cascade screening following a PRS test.

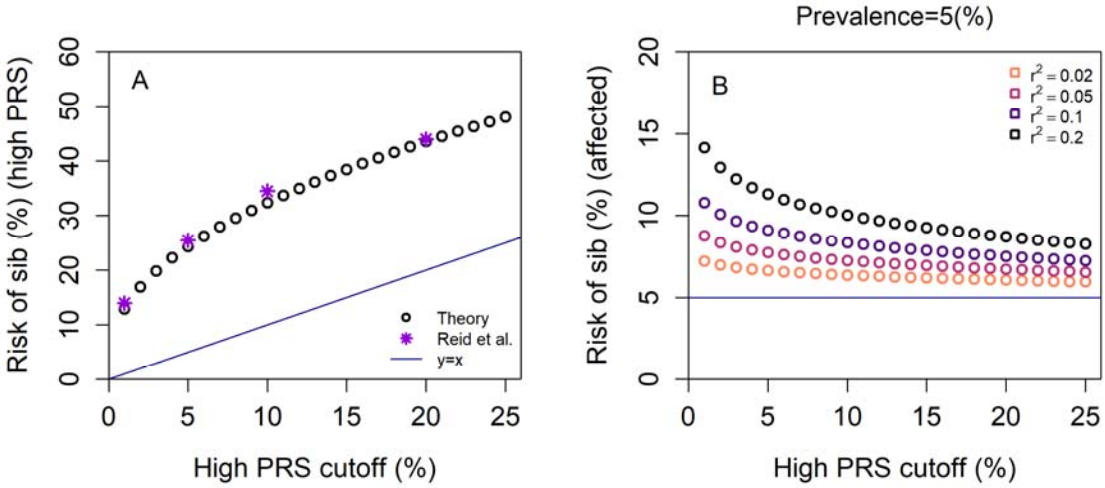

Figure 1. The expected risk of a sibling conditional on a proband having a PRS above a cutoff. We assume that the proband is known to have a PRS at the top q quantile of the PRS distribution. The cutoff percentile varies along the x-axis. In (A), we plot the risk of the sib of the proband to have a high PRS, defined using the same cutoff. The diagonal blue line is $y = x$, which is the risk for an unrelated individual. The circles are the theoretical probabilities we derived based on the liability threshold model, obtained by numerically evaluating Eq. (5). The violet stars correspond to the empirical values (mean across diseases) obtained by Reid et al. In (B), we plot the risk of the sib to be affected. We assume a prevalence of $K = 5\%$, and show results for multiple values of r^2, a measure of PRS accuracy equal to the proportion of variance in liability explained by the PRS (legend). The horizontal blue line represents the risk of an unrelated individual (equal to K). The circles are the theoretical probabilities we derived, obtained by numerically evaluating Eq. (11).

Competing interests

S. C. is paid consultant at MyHeritage.

Bibliography

1. A. R. Martin, M. J. Daly, E. B. Robinson *et al.* Predicting Polygenic Risk of Psychiatric Disorders. *Biol Psychiatry* **86**, 97 (2019).
2. N. Mars, J. T. Koskela, P. Ripatti *et al.* Polygenic and clinical risk scores and their impact on age at onset and prediction of cardiometabolic diseases and common cancers. *Nat Med* **26**, 549 (2020).
3. A. V. Khera, M. Chaffin, K. G. Aragam *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet* **50**, 1219 (2018).
4. L. Kachuri, R. E. Graff, K. Smith-Byrne *et al.* Pan-cancer analysis demonstrates that integrating polygenic risk scores with modifiable risk factors improves risk prediction. *Nat Commun* **11**, 6084 (2020).
5. L. Sun, L. Pennells, S. Kaptoge *et al.* Polygenic risk scores in cardiovascular risk prediction: A cohort study and modelling analyses. *PLoS Med* **18**, e1003498 (2021).
6. J. W. Knowles, D. J. Rader & M. J. Khoury. Cascade Screening for Familial Hypercholesterolemia and the Use of Genetic Testing. *JAMA* **318**, 381 (2017).
7. P. J. Talmud, S. Shah, R. Whittall *et al.* Use of low-density lipoprotein cholesterol gene score to distinguish patients with polygenic and monogenic familial hypercholesterolaemia: a case-control study. *Lancet* **381**, 1293 (2013).

8. S. A. Lambert, G. Abraham & M. Inouye. Towards clinical utility of polygenic risk scores. *Hum Mol Genet* **28**, R133 (2019).
9. E. Widén, N. Junna, S. Ruotsalainen *et al.* Communicating polygenic and non-genetic risk for atherosclerotic cardiovascular disease - An observational follow-up study. *medRxiv*, 2020.09.18.20197137 (2020).
10. N. J. Reid, D. G. Brockman, C. Elisabeth Leonard *et al.* Concordance of a High Polygenic Score Among Relatives: Implications for Genetic Counseling and Cascade Screening. *Circ Genom Precis Med*, CIRCGEN120003262 (2021).
11. E. R. Dempster & I. M. Lerner. Heritability of Threshold Characters. *Genetics* **35**, 212 (1950).
12. M. Lynch & B. Walsh. *Genetics and analysis of quantitative traits*. (Sinauer Associates, 1998).
13. P. M. Visscher & N. R. Wray. Concepts and Misconceptions about the Polygenic Additive Model Applied to Disease. *Hum Hered* **80**, 165 (2015).
14. N. R. Wray & M. E. Goddard. Multi-locus models of genetic risk of disease. *Genome Med* **2**, 10 (2010).
15. S. H. Lee, N. R. Wray, M. E. Goddard *et al.* Estimating missing heritability for disease from genome-wide association studies. *Am J Hum Genet* **88**, 294 (2011).
16. O. Weissbrod, J. Flint & S. Rosset. Estimating SNP-Based Heritability and Genetic Correlation in Case-Control Studies Directly and with Summary Statistics. *Am J Hum Genet* **103**, 89 (2018).
17. H. C. So, J. S. Kwan, S. S. Cherny *et al.* Risk prediction of complex diseases from family history and known susceptibility loci, with applications for cancer screening. *Am J Hum Genet* **88**, 548 (2011).
18. D. S. Falconer. The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Ann Hum Genet* **29**, 51 (1965).
19. C. B. Do, D. A. Hinds, U. Francke *et al.* Comparison of family history and SNPs for predicting risk of complex disease. *PLoS Genet* **8**, e1002973 (2012).
20. S. H. Lee, M. E. Goddard, N. R. Wray *et al.* A better coefficient of determination for genetic profile analysis. *Genet Epidemiol* **36**, 214 (2012).
21. E. Karavani, O. Zuk, D. Zeevi *et al.* Screening Human Embryos for Polygenic Traits Has Limited Utility. *Cell* **179**, 1424 (2019).
22. T. Lencz, D. Backenroth, A. Green *et al.* Utility of polygenic embryo screening for disease depends on the selection strategy. *bioRxiv*, 2020.11.05.370478 (2020).
23. N. R. Wray, K. E. Kemper, B. J. Hayes *et al.* Complex Trait Prediction from Genome Data: Contrasting EBV in Livestock to PRS in Humans: Genomic Prediction. *Genetics* **211**, 1131 (2019).
24. P. M. Visscher & M. E. Goddard. From R.A. Fisher's 1918 Paper to GWAS a Century Later. *Genetics* **211**, 1125 (2019).