Numerical solution of the risk reduction integrals

Liraz Klausner[1] and Shai Carmi[1]

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

September 21, 2022

1 Background

Our goal in developing the calculator is to permit users to estimate the disease risk of a child vs parameters such as the number of embryos or the proportion of variance explained by the PRS. In our previous work (Lencz et al., 2021), we modeled the disease risk using the liability threshold model. We were then able to express the risk as integrals. These integrals had no closed form solution, and we used generic integration methods in R to solve them numerically. However, such methods can be slow or inaccurate. Here, we describe new approaches we used in this paper to solve the integrals numerically.

We denote by K the disease prevalence and by n the number of embryos. We denote by q the polygenic risk score (PRS) quantile above which we exclude embryos. [If all embryos are high risk, we select at random.] We denote by z_K and z_q the K and q upper-quantiles from the standard normal distribution, respectively. Finally, r^2 is variance of the PRS, or equivalently the proportion of the variance in liability explained by the PRS.

2 Direct calculations

The first section includes the integrals that we calculated without resorting to simulations. These include all cases except when conditioning on the parental disease status.

2.1 Lowest risk prioritization

The disease risk when transferring the embryo with the lowest risk is given by Eq. (20) from the appendix of Lencz et al. (2021),

$$P(\text{disease}) = \int_{-\infty}^{\infty} \left[1 - \Phi\left(\frac{z_K - t\sqrt{1 - r^2/2}}{r/\sqrt{2}}\right)\right]^n \phi(t)dt. \tag{1}$$

Above, $\phi(\cdot)$ and $\Phi(\cdot)$ are the standard normal PDF and CDF, respectively. The disease risk conditional on the mean parental score, c, is given by Eq. (23) therein,

$$P(\text{disease} \mid c) = \int_{-\infty}^{\infty} \left[1 - \Phi\left(\frac{z_K - c - t\sqrt{1-r^2}}{r/\sqrt{2}}\right)\right]^n \phi(t)dt \tag{2}$$

We solved these two integrals numerically using R's `integrate` function, as in (Lencz et al., 2021), as this approach was fast and sufficiently accurate.

2.2 High-risk exclusion

The disease risk when excluding high-risk embryos and given the mean parental score is given by Eq. (29) of the appendix of Lencz et al. (2021). We also solved it using R's `integrate`. The unconditional disease risk is given by Eq. (31) therein. The integral is discontinuous, and we rewrite as two separate integrals as

$$\begin{aligned} P(\text{disease}) = \int_{-\infty}^{\infty} \Bigg\{ & f_1(u) \int_{-\infty}^{\sqrt{2}z_q - u} \left[1 - \Phi\left(\frac{z_K - (u+t)r/\sqrt{2}}{\sqrt{1-r^2}}\right)\right] \phi(t)dt + \\ & f_2(u) \int_{\sqrt{2}z_q - u}^{\infty} \left[1 - \Phi\left(\frac{z_k - (u+t)r/\sqrt{2}}{\sqrt{1-r^2}}\right)\right] \phi(t)dt \Bigg\} \phi(u)du, \end{aligned} \tag{3}$$

and $f_1(u)$ and $f_2(u)$ are defined as

$$\begin{aligned} f_1(u) &= \frac{1 - \left[1 - \Phi\left(\sqrt{2}z_q - u\right)\right]^n}{\Phi\left(\sqrt{2}z_q - u\right)} \\ f_2(u) &= \left[1 - \Phi\left(\sqrt{2}z_q - u\right)\right]^{n-1}. \end{aligned}$$

To proceed, we rewrite the inner integral part of Eq. (3) as the more generic form

$$I = \int_h^k \phi(y)\Phi(a + by)dy,$$

by using $1 - \Phi(-y) = \Phi(y)$.

This generic form of the integral is given in Owen (1980), here we show how to derive it. Another way to write the same integral is

$$\begin{cases} I = (\Phi(k) - \Phi(h))P(X < Y), & \text{if} \quad b > 0 \\ I = (\Phi(k) - \Phi(h))P(X > Y), & \text{if} \quad b < 0 \end{cases} \tag{4}$$

with

$$X \sim N\left(-\frac{a}{b}, \frac{1}{|b|}\right), Y \sim \text{TN}(0, 1; h, k),$$

where $\mathrm{TN}(0,1;h,k)$ denotes the truncated standard normal distribution with range (h,k), and X and Y are independent. To see this, we have, for $b > 0$

$$\begin{aligned}P(X < Y) &= \int_h^k P(X < y \mid Y = y) f(Y = y) dy \\ &= \int_h^k \frac{\Phi\left(\frac{y+\frac{a}{b}}{\frac{1}{|b|}}\right)}{\Phi(k) - \Phi(h)} \phi(y) dy = \int_h^k \frac{\Phi(a+by)}{\Phi(k)-\Phi(h)} \phi(y) dy \\ &= \frac{I}{\Phi(k) - \Phi(h)}. \end{aligned} \tag{5}$$

When $b < 0$, we instead calculate

$$\begin{aligned}P(X > Y) &= \int_h^k \frac{1 - \Phi\left(\frac{y+\frac{a}{b}}{\frac{1}{|b|}}\right)}{\Phi(k) - \Phi(h)} \phi(y) dy = \int_h^k \frac{1 - \Phi(-a-by)}{\Phi(k)-\Phi(h)} \phi(y) dy \\ &= \int_h^k \frac{\Phi(a+by)}{\Phi(k)-\Phi(h)} \phi(y) dy = \frac{I}{\Phi(k) - \Phi(h)}, \end{aligned} \tag{6}$$

Given the structure of Eq. (3), we henceforth assume $b > 0$. To do the actual calculation we rewrite the truncated normal as

$$Y = P(Y' = y' \mid h \le Y' \le k), Y' \sim N(0,1).$$

Then, we define $Z = X - Y'$, and note that the joint distribution of Z and Y' is

$$\begin{pmatrix} Y' \\ Z \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ -\frac{a}{b} \end{pmatrix}, \quad \begin{pmatrix} 1 & -1 \\ -1 & 1 + \frac{1}{b^2} \end{pmatrix}\right]$$

Back to the probability in Eq. (5), what we want to calculate is

$$P(X < Y) = P(Z < 0 \mid h \le Y' \le k) = \frac{P(Z < 0, h \le Y' \le k)}{P(h \le Y' \le k)}.$$

Note that $P(h \le Y' \le k) = \Phi(k) - \Phi(h)$, so that $P(Z < 0, h \le Y' \le k) = I$, which means that the integral we want can be defined by the CDF of bivariate normal, that is

$$P(Z < 0, h \le Y' \le k) = P(Z < 0, Y' \le h) - P(Z < 0, Y' \le k).$$

We next define a normalized $Z' = (Z + a/b)/\sqrt{1 + 1/b^2}$, or $Z = \sqrt{1 + 1/b^2}Z' - a/b$. The correlation between Z' and Y' is $\mathrm{cor}(Z', Y') = \mathrm{cor}(Z, Y') \equiv \rho = \frac{-1}{\sqrt{1+1/b^2}}$. This gives

$$P(Z < 0, Y' \le h) = P\left(Z' < \frac{a/b}{\sqrt{1+1/b^2}}, Y' \le h\right) \tag{7}$$

$$= \Phi_2\left(\frac{a/b}{\sqrt{1+1/b^2}}, h; \rho = -\frac{1}{\sqrt{1+1/b^2}}\right)$$

and similarly for $P(Z < 0, Y' \leq k)$. Above, $\Phi_2(\cdot, \cdot; \rho)$ is the CDF of a bivariate standard normal distribution with correlation ρ. This CDF can be calculated through Owen's T function (Owen, 1956). The T function is defined as

$$T(h, \alpha) = \frac{1}{2\pi} \int_0^\alpha \frac{\exp(-\frac{1}{2}h^2(1+x^2))}{1+x^2} dx.$$

Fast and accurate methods exist for computing the T function numerically (OwenQ R package (Laurent, 2022; Patefield, 2000)). Given the T function, the bivariate CDF can be computed as

$$\begin{aligned} \Phi_2(x, y; \rho) &= 0.5\Phi(x) + 0.5\Phi(y) - T(x, a_x) - T(y, a_y) \\ &\quad - \begin{cases} 0 & \text{if } xy > 0 \text{ or } xy = 0, x + y \geq 0 \\ 0.5 & \text{otherwise} \end{cases} \end{aligned} \tag{8}$$

with

$$\begin{aligned} a_x &= \frac{y}{x\sqrt{1-\rho^2}} - \frac{\rho}{\sqrt{1-\rho^2}}, \\ a_y &= \frac{x}{y\sqrt{1-\rho^2}} - \frac{\rho}{\sqrt{1-\rho^2}}. \end{aligned}$$

In Eq. (3), the inner integral was calculated using the T function, where some of the terms were simplified when we substituted $x = -\infty$ or $y = \infty$. We computed the outer integral with R's `integrate` function.

3 Monte Carlo simulations

When conditioning on the parental disease status (Section 6 in the appendix of Lencz et al. (2021)), we need to calculate multiple integrals for both the exclude and lowest risk strategies. The resulting integrals are more complicated and tedious to write. However, each one of them can be written using the integral $\int_{-\infty}^{\infty} f(x)\phi(x; \sigma)dx$ (i.e., where x is normal with zero mean and standard deviation σ). This allows us to use Monte Carlo simulation by sampling $X \sim N(0, \sigma)$, and then calculating

$$\frac{1}{M} \sum_{i=1}^{M} f(x_i).$$

Based on the law of large numbers, this converges to the desired expected value

$$\mathbb{E}[f(X)] = \int_{-\infty}^{\infty} f(x)\phi(x; \sigma)dx.$$

The same approach works for higher dimensional integrals, with the expected value of the joint distribution. This can be written as

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_i, \ldots, x_m)\phi_m(x_1, \ldots, x_m; \mathbf{\Sigma})dx_1 \cdots dx_m = \mathbb{E}_{\boldsymbol{x}}(f(x_1, ..., x_m)), \tag{9}$$

where $\phi_m(\boldsymbol{x}; \boldsymbol{\Sigma})$ is the PDF of a multivariate normal distribution with zero means and variance matrix $\boldsymbol{\Sigma}$. Again from the law of large numbers, this can be estimated as

$$\frac{1}{M} \sum_{i=1}^{M} f(x_{i1}, \ldots, x_{im}),$$

by sampling x_{ij}, $(i = 1, \ldots, M,\ j = 1, \ldots, m)$ from multivariate normal distribution.

To estimate the standard deviation of $\int_{-\infty}^{\infty} f(x)\phi(x)dx = \mathbb{E}[f(X)]$, we computed the empirical standard deviation of $f(x)$ over the M draws and divided by $\sqrt{M}$.

3.1 Control-variate based Monte Carlo integration

Monte-carlo based methods can be noisy and require a large sample size to ensure low error. We therefore used control variates to reduce the variance.

The idea is as follows. Suppose that we want to estimate $\mathbb{E}(f(X))$ for some function $f(x)$, and some random variable X. We define a new random variable Y,

$$Y = f(X) + \alpha(g(X) - \mathbb{E}(g(X))),$$

where g is some function with known expectation, which we call the control-variate. Taking expectation from both sides, we have

$$\mathbb{E}(Y) = \mathbb{E}(f(X)) + \alpha(\underbrace{\mathbb{E}(g(x)) - \mathbb{E}(g(x))}_{=0})) = \mathbb{E}(f(X)).$$

So the new random variable has the same expected value as $f(X)$. Its variance is

$$\text{Var}(Y) = \text{Var}(f(X)) + \alpha^2 \text{Var}(g(X)) + 2\alpha \text{Cov}(f(X), g(X)).$$

Taking the derivative with respect to α in order to minimize the variance, we can find a^* that minimizes $\text{Var}(Y)$,

$$\alpha^* = -\frac{\text{Cov}(f(X), g(X))}{\text{Var}(g(X))},$$

which is a minimum since the original function is convex (the second derivative is $2\text{Var}(g(X)) \geq 0$). The new variance after plugging in α^* is

$$\begin{aligned} \text{Var}(Y) &= \text{Var}(f(X)) + \frac{\text{Cov}^2(f(X), g(X))}{\text{Var}(g(X))} - 2\frac{\text{Cov}^2(f(X), g(X))}{\text{Var}(g(X))} \qquad (10) \\ &= \text{Var}(f(X)) - \underbrace{\frac{\text{Cov}^2(f(X), g(X))}{\text{Var}(g(X))}}_{\geq 0} \leq \text{Var}(f(X)). \end{aligned}$$

This variance is lower than the original variance. We can therefore use the empirical mean of Y in order to get better estimates of $\mathbb{E}(f(X))$. Usually it is difficult to calculate the the variance of $g(X)$ or the covariance $\mathrm{Cov}(f(X), g(X))$. We thus used the data to estimate both of them and then $\hat{\alpha}^*$. In practice, it can be shown that $\mathbb{E}(f(X))$ can be estimated as the intercept of the least squares linear regression of $f(x)$ on $g(x) - \mathbb{E}(g(x))$.

The method can be extended to multiple control variates:

$$Y = f(X) + \alpha_1(g_1(X) - \mathbb{E}(g_1(X))) + \alpha_2(g_2(X) - \mathbb{E}(g_2(X))) + \cdots \tag{11}$$

Defining the row vector $\boldsymbol{g}(x) = (g_1(x), g_2(x), \ldots)$, the optimal coefficients are given by $(\alpha_1, \alpha_2, \ldots) = \mathbb{E}(\boldsymbol{g}(x)^\top \boldsymbol{g}(x))^{-1}\mathbb{E}(\boldsymbol{g}(x)^\top y)$. The method can be similarly extended to multiple variables.

In our problem, we used the control variate method only for the outermost integral in the exclusion strategy (over the variable g_f (see (Lencz et al., 2021))). For the lowest risk strategy we used the method for all variables. We picked $g_1(x) = x$ and $g_2(x) = \tanh(x)$ as the control variates, with known expected values

$$\begin{aligned} &\mathbb{E}(X) = 0, \\ &\mathbb{E}[\tanh(X)] = 0. \end{aligned} \tag{12}$$

The second equality is since tanh is an odd function, and the PDF of a normal variable with zero mean is an even function, so the product is odd and the integral from $-\infty$ to ∞ is zero.

References

Stéphane Laurent. *OwenQ: Owen Q-Function*, 2022. URL https://CRAN.R-project.org/package=OwenQ. R package version 1.0.5.

Todd Lencz, Daniel Backenroth, Einat Granot-Hershkovitz, Adam Green, Kyle Gettler, Judy H Cho, Omer Weissbrod, Or Zuk, and Shai Carmi. Utility of polygenic embryo screening for disease depends on the selection strategy. *eLife*, 10:e64716, 2021. ISSN 2050-084X. doi: 10.7554/eLife.64716. URL https://doi.org/10.7554/eLife.64716.

D. B. Owen. A table of normal integrals. *Communications in Statistics - Simulation and Computation*, 9(4):389–419, 1980. ISSN 0361-0918. doi: 10.1080/03610918008812164. URL https://doi.org/10.1080/03610918008812164.

Donald B. Owen. Tables for Computing Bivariate Normal Probabilities. *The Annals of Mathematical Statistics*, 27(4):1075–1090, 1956. ISSN 0003-4851, 2168-8990. doi: 10.1214/aoms/1177728074. URL https://projecteuclid.org/journals/annals-of-mathematical-statistics/volume-27/issue-4/Tables-for-Computing-Bivariate-Normal-Probabilities/10.1214/aoms/1177728074.full.

Mike Patefield. Fast and accurate calculation of owen's t function. *Journal of Statistical Software*, 5(5):1–25, 2000. doi: 10.18637/jss.v005.i05. URL https://www.jstatsoft.org/index.php/jss/article/view/v005i05.