Supplementary Material to
Testing calibration of risk models at extremes of disease-risk

Minsun Song,[1] Peter Kraft,[2] Amit D. Joshi,[2] Myrto Barrdahl,[3] Nilanjan Chatterjee[1]

[1]Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Rockville, Maryland 20850, USA.

[2]Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts 02115, USA.

[3]Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany

1. Asymptotic theory for complete-case analysis

Let $\boldsymbol{\gamma} = (\alpha, \boldsymbol{\beta}^T)^T$ and $\boldsymbol{\vartheta}_i = (1, \frac{\partial m(\boldsymbol{\beta};\mathbf{G}_i)}{\partial \boldsymbol{\beta}}^T)^T$. Let $\boldsymbol{\gamma}_0$ and $\boldsymbol{\beta}_0$ be the true parameter values of $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ under the null model, respectively.

Assumption 1 : Let f and $f^{'}$ be the derivative and second-order derivative of F assuming that $f^{'}$ is continuous. There exists $M < \infty$ s.t. $0< F(x) < 1$ and $f(x) > 0$ for $|x| \leqslant M$.

Assumption 2 : $\boldsymbol{\gamma}_0$ is contained in an open bounded parameter space $\Gamma \in \mathbb{R}^{p+1}$.

Assumption 3 : $\max_i |\{\boldsymbol{\vartheta}_i\}| \leqslant C$ for some $C < \infty$. $\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\vartheta}_i \boldsymbol{\vartheta}_i^T$ is a finite nonsingular matrix. The empirical distribution of $\{\boldsymbol{\vartheta}_i\}$ converges to a (non-degenerate) distribution function.

Assumption 4: $m(.)$ is continuously differentiable.

Proposition 1 *Under the assumptions 1, 2, 3, and 4,*

$$n^{-\frac{1}{2}}(T_{n,c} - n_c) \Rightarrow N(0, \sigma^2_{n,c})$$

where $T_{n,c} = \sum_{i=1}^{n} \frac{(D_i - \hat{\pi}_i)^2}{\hat{\pi}_i \times (1-\hat{\pi}_i)} I(\hat{\pi}_i \in R_c)$, $n_c = \sum_{i=1}^{n} I(\hat{\pi}_i \in R_c)$,

$\sigma^2_{n,c} = \frac{1}{n}\sum_{i=1}^{n} \frac{(1-2\pi_{i0})^2}{\pi_{i0}(1-\pi_{i0})} I(\pi_{i0} \in R_c) - \mathbf{v}^T_{n,c}\Omega^{-1}\mathbf{v}_{n,c}$, $\mathbf{v}_{n,c} = \frac{1}{n}\sum_{i=1}^{n} \frac{(1-2\pi_{i0})}{\pi_{i0}(1-\pi_{i0})} f_{i0}\boldsymbol{\vartheta}_i I(\pi_{i0} \in R_c)$

, and $\Omega = lim_n \sqrt{n}\ Var(\hat{\boldsymbol{\gamma}}_n)$.

Proof. Note that

$$n^{-\frac{1}{2}}(T_{n,c} - n_c) = n^{-\frac{1}{2}} \sum_{i=1}^{n} \frac{(D_i - \hat{\pi}_i)(1 - 2\hat{\pi}_i)}{\hat{\pi}_i(1-\hat{\pi}_i)} I(\hat{\pi}_i \in R_c). \tag{1.1}$$

Define

$$S_{n,c}(\boldsymbol{\gamma}) = \sum_{i=1}^{n} \frac{(D_i - \pi_i)(1-2\pi_i)}{\pi_i(1-\pi_i)} I(\pi_i \in R_c) = \sum_{i=1}^{n} (\frac{D_i}{\pi_i} + \frac{1-D_i}{1-\pi_i} - 2) I(\pi_i \in R_c).$$

To obtain the asymptotic distribution of $T_{n,c}$, we shall first prove that

$$n^{-\frac{1}{2}} S_{n,c}(\hat{\boldsymbol{\gamma}}_n) \Rightarrow N(0, \sigma^2_{n,c}) \tag{1.2}$$

and then show that

$$S_{n,c}(\hat{\boldsymbol{\gamma}}_n) - [T_{n,c} - n_c] = o_{\mathbb{P}}(n^{1/2}). \tag{1.3}$$

To prove (1.2), note that local expansion of $S_{n,c}(\boldsymbol{\gamma})$ around $\boldsymbol{\gamma}_0$ gives

$$n^{-\frac{1}{2}} S_{n,c}(\hat{\boldsymbol{\gamma}}_n) = n^{-\frac{1}{2}} S_{n,c}(\boldsymbol{\gamma}_0) + n^{-1} \frac{\partial S_{n,c}(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}}|_{\boldsymbol{\gamma}_0} n^{\frac{1}{2}}(\hat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_0) + o_p(1).$$

Meanwhile,

$$n^{-1} \frac{\partial S_{n,c}(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}}|_{\boldsymbol{\gamma}_0} = n^{-1} \sum_{i=1}^{n} (\frac{(1-D_i)}{(1-\pi_{i0})^2} - \frac{D_i}{\pi^2_{i0}}) f_{i0} \boldsymbol{\vartheta}_i I(\pi_{i0} \in R_c) = \mathbf{U}_{n,c} - \mathbf{v}_{n,c}$$

with

$$\mathbf{U}_{n,c} = -n^{-1} \sum_{i=1}^{n} \frac{(D_i - \pi_{i0})(1 - 2\pi_{i0} + 2\pi^2_{i0})}{\pi^2_{i0}(1-\pi_{i0})^2} f_{i0} \boldsymbol{\vartheta}_i I(\pi_{i0} \in R_c)$$

and

$$\mathbf{v}_{n,c} = n^{-1} \sum_{i=1}^{n} \frac{1 - 2\pi_{i0}}{\pi_{i0}(1-\pi_{i0})} f_{i0} \boldsymbol{\vartheta}_i I(\pi_{i0} \in R_c).$$

Under the assumptions 1-4, $\lim_{n\to\infty} \mathbf{U}_{n,c} = \mathbf{0}$ and it follows that the limiting distribution of $n^{-\frac{1}{2}} S_{n,c}(\hat{\boldsymbol{\gamma}}_n)$ is the same as the limiting distribution of

$$n^{-\frac{1}{2}} S_{n,c}(\boldsymbol{\gamma}_0) - \mathbf{v}^T_{n,c} n^{\frac{1}{2}} (\hat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_0).$$

Since $\hat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_0 = -[\frac{\partial^2 l_n}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}^T}|_{\boldsymbol{\gamma}_n^*}]^{-1} \frac{\partial l_n}{\partial \boldsymbol{\gamma}}|_{\boldsymbol{\gamma}_0}$ where l_n is the log likelihood for obtaining $\hat{\boldsymbol{\gamma}}_n$ and $\boldsymbol{\gamma}_n^*$ lies between $\hat{\boldsymbol{\gamma}}_n$ and $\boldsymbol{\gamma}_0$,

$$\mathbf{v}_{n,c}^T n^{\frac{1}{2}} (\hat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_0) = \mathbf{v}_{n,c}^T (A_n^*)^{-1} n^{-\frac{1}{2}} \frac{\partial l}{\partial \boldsymbol{\gamma}} |_{\boldsymbol{\gamma}_0},$$

where $A_n^* = -n^{-1} \frac{\partial^2 l}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}^T}|_{\boldsymbol{\gamma}_n^*}$. Since $\lim_n A_n^* = \Omega$, the limiting distribution of $n^{-\frac{1}{2}}(T_{n,c} - n_c)$ is the same as the limiting distribution of

$$n^{-\frac{1}{2}} S_{n,c}(\boldsymbol{\gamma}_0) - n^{-\frac{1}{2}} \mathbf{v}_{n,c}^T \Omega^{-1} \frac{\partial l}{\partial \boldsymbol{\gamma}} |_{\boldsymbol{\gamma}_0} = n^{-\frac{1}{2}} \sum_{i=1}^n \frac{D_i - \pi_{i0}}{\pi_{i0}(1 - \pi_{i0})} \{(1 - 2\pi_{i0}) I(\pi_{i0} \in R_c) - \mathbf{v}_{n,c}^T \Omega^{-1} f_{i0} \boldsymbol{\vartheta}_i\}.$$

Define

$$Z_{n,c,i} = \frac{D_i - \pi_{i0}}{\pi_{i0}(1 - \pi_{i0})} \{(1 - 2\pi_{i0}) I(\pi_{i0} \in R_c) - \mathbf{v}_{n,c}^T \Omega^{-1} \boldsymbol{\vartheta}_i f_{i0}\}.$$

Then $Z_{n,c,1}, ..., Z_{n,c,n}$ are independent random variables such that

$$E(Z_{n,c,i}) = 0$$

$$Var(Z_{n,c,i}) = \frac{((1 - 2\pi_{i0}) I(\pi_{i0} \in R_c) - \mathbf{v}_{n,c}^T \Omega^{-1} \boldsymbol{\vartheta}_i f_{i0})^2}{\pi_{i0}(1 - \pi_{i0})} = \sigma_{n,c,i}^2$$

$$E(|Z_{n,c,i}|^3) = \frac{(1 - 2\pi_{i0} + 2\pi_{i0}^2)}{\pi_{i0}^2(1 - \pi_{i0}^2)} |(1 - 2\pi_{i0}) I(\pi_{i0} \in R_c) - \mathbf{v}_{n,c}^T \Omega^{-1} \boldsymbol{\vartheta}_i f_{i0}|^3 = m_{n,c,i}.$$

Under the assumptions 1-4, we have that

$$\lim_{n \to \infty} (\sum_{i=1}^n \sigma_{n,c,i}^2)^{-\frac{1}{2}} (\sum_{i=1}^n m_{n,c,i})^{\frac{1}{3}} = 0,$$

so by Liapounov's Central Limit Theorem, for large n,

$$n^{-\frac{1}{2}} \sum_{i=1}^n \frac{(D_i - \pi_{i0}) \{(1 - 2\pi_{i0}) I(\pi_{i0} \in R_c) - \mathbf{v}_{n,c}^T \Omega^{-1} f_{i0} \boldsymbol{\vartheta}_i\}}{\pi_{i0}(1 - \pi_{i0})} \sim N(0, \frac{1}{n} \sum_{i=1}^n \sigma_{n,c,i}^2)$$

where $\frac{1}{n} \sum_{i=1}^n \sigma_{n,c,i}^2$ can be written as follows $\frac{1}{n} \sum_{i=1}^n \frac{(1 - 2\pi_{i0}^2)^2}{\pi_{i0}(1 - \pi_{i0})} I(\pi_{i0} \in R_c) - \mathbf{v}_{n,c}^T \Omega^{-1} \mathbf{v}_{n,c}$.

We now prove (1.3). Observe that

$$S_{n,c}(\hat{\boldsymbol{\gamma}}_n) - [T_{n,c} - n_c] = \sum_{i=1}^n \frac{(D_i - \hat{\pi}_i)(1 - 2\hat{\pi}_i)}{\hat{\pi}_i(1 - \hat{\pi}_i)} [I(\pi_{i0} \in R_c) - I(\hat{\pi}_i \in R_c)] := \sum_{i=1}^n \Xi_i \tau_i.$$

Let $\Upsilon_i = 1/\sqrt{n}\sum_{j=1}^{i}\Xi_j$. Then standard arguments show that $\{\Upsilon_i\}_{i=1}^{n}$ converges weakly to a centered Gaussian process on $D[0,1]$ with the Skorohod topology. Meanwhile, it is easy to see that

$$\max_{1\leqslant i\leqslant n}|\tau_i| = o_{\mathbb{P}}(1) \text{ for continuously differentiable } m(\cdot).$$

Hence by a classic random change of time argument (see for instance Billingsley 1999, p.151 or Lemma 4.1 of Moore and Spruill, 1975), we have (1.3) follows.

Note : Covariance between a pair of thresholds c and d has an analytical form as

$$\sigma_{n,c,d}^2 := Cov(n^{-\frac{1}{2}}T_{n,c}, n^{-\frac{1}{2}}T_{n,d}) = \frac{1}{n}\sum_{i=1}^{n}\frac{(1-2\pi_{i0})^2}{\pi_{i0}(1-\pi_{i0})}I(\pi_{i0}\in R_c\bigcap R_d) - \mathbf{v}_{n,c}^{'}\Omega^{-1}\mathbf{v}_{n,d}.$$

2. Additive model

The additive model for the risk of the disease in the underlying population given the SNP genotype data for p loci is given by

$$\text{pr}(D=1|\mathbf{G}) = b_0 + \sum_{j=1}^{p} b_j G_j, \tag{2.4}$$

where $\mathbf{G}$ and $\mathbf{G_{-j}}$ denotes the vector of genotypes across p SNPs and the vector of genotypes across $p{-}1$ SNPs except SNP j, respectively,

$$b_0 = \text{pr}(D=1|\mathbf{G}=\mathbf{0}),$$

and

$$\begin{aligned} b_j &= \text{pr}(D=1|G_j=1,\mathbf{G_{-j}}=\mathbf{0}) - \text{pr}(D=1|\mathbf{G}=\mathbf{0}) \\ &= \text{pr}(D=1|G_j=2,\mathbf{G_{-j}}=\mathbf{0}) - \text{pr}(D=1|G_j=1,\mathbf{G}_{-j}=\mathbf{0}). \end{aligned}$$

In words, for any SNP j, b_j denotes risk differences associated with the genotype status for the SNP j while holding the genotype status for all the other SNPs at the reference level 0.

Then, the additive model corresponds to

$$\begin{aligned}\mathrm{pr}(D=1|\mathbf{G}) &= \mathrm{pr}(D=1|\mathbf{G}=\mathbf{0}) + \sum_{j=1}^{p}\{\mathrm{pr}(D=1|G_j=1,\mathbf{G_{-j}}=\mathbf{0}) - \mathrm{pr}(D=1|\mathbf{G}=\mathbf{0})\}G_j \\ &= \sum_{j=1}^{p}\mathrm{pr}(D=1|G_j=1,\mathbf{G_{-j}}=\mathbf{0})G_j - (\sum_{j=1}^{p}G_j-1)\mathrm{pr}(D=1|\mathbf{G}=\mathbf{0})\end{aligned}$$

If we define $RR(\mathbf{G}) = \mathrm{pr}(D=1|\mathbf{G})/\mathrm{pr}(D=1|\mathbf{G}=\mathbf{0})$ to be relative risk for the groups compared to the reference group with $\mathbf{G}=\mathbf{0}$, then the constraint can be expressed as

$$RR(\mathbf{G}) = \sum_{j=1}^{p} RR(G_j=1,\mathbf{G_{-j}}=\mathbf{0}) \times G_j - (\sum_{j=1}^{p}G_j - 1).$$

Under the assumption of rare diseases, that allows us to approximate odds-ratios by relative risks, the above constraint can be written in the form

$$\exp(m(\boldsymbol{\beta};G)) = OR(\mathbf{G}) = \sum_{j=1}^{p} OR(G_j=1,\mathbf{G_{-j}}=\mathbf{0})G_j - (\sum_{j=1}^{p}G_j-1). \qquad (2.5)$$

Thus the additive model corresponds to a logistic model with $m(\boldsymbol{\beta};G) = \log(\sum_{j=1}^{p}\beta_j \times G_j + 1)$. where $OR(G_j=1,\mathbf{G_{-j}}=\mathbf{0}) = 1+\beta_j$ where $\beta_j = \frac{b_j}{b_0}$.

3. COMPUTATION OF $\hat{\pi}_i^*(\mathbf{G}_{i,obs})$ FOR SUBJECTS WITH MISSING GENOTYPE DATA

We propose incorporating individuals with missing genotype data using a modification of the test-statistics in the form

$$T_{n,c} = \sum_{i=1}^{n} \frac{(D_i - \hat{\pi}_i^*(\mathbf{G}_{i,obs}))^2}{\hat{\pi}_i^*(\mathbf{G}_{i,obs})(1-\hat{\pi}_i^*(\mathbf{G}_{i,obs}))} I(\hat{\pi}_i^*(\mathbf{G}_{i,obs}) \in R_c)$$

where $\mathbf{G}_{i,obs}$ denotes the observed genotype data for the i-th subject and

$$\hat{\pi}_i^*(\mathbf{G}_{i,obs}) = \hat{\pi}_i^* = \mathrm{pr}(D_i=1|\mathbf{G}_{i,obs},R_i=1).$$

We note that if we write $\mathbf{G}_i = (\mathbf{G}_{i,obs}, \mathbf{G}_{i,miss})$, i.e. the total genotype vector into the observed and unobserved genotypes, then we can write

$$\mathrm{pr}(D_i = 1|\mathbf{G}_{i,obs}) = \sum_{\mathbf{G}_{i,miss}} \mathrm{pr}(D_i = 1|\mathbf{G}_{i,miss}, \mathbf{G}_{i,obs})\mathrm{pr}(\mathbf{G}_{i,miss}|\mathbf{G}_{i,obs}).$$

If the original model for probability of the disease given all SNPs has logistic form, then, under the assumption that the disease under study is rare in the general population, the probability model for the disease given the observable genotypes can be also written in the multiplicative form

$$\mathrm{pr}(D_i = 1|\mathbf{G}_{i,obs}) = \exp(\alpha + m^*(\boldsymbol{\beta}, \mathbf{q}; \mathbf{G}_{i,obs}))$$

where $m^*(\boldsymbol{\beta}, \mathbf{q}; \mathbf{G}_{obs}) = \log(\sum_{\mathbf{G}_{miss}} \exp(m(\boldsymbol{\beta}; \mathbf{G}_{miss}, \mathbf{G}_{obs}))\mathrm{pr}(\mathbf{G}_{miss}|\mathbf{G}_{obs})$, $\mathbf{q} = (q_1,, ..., q_{K_{miss}})$ denotes $\mathrm{pr}(\mathbf{G}_{miss}|\mathbf{G}_{obs})$ and K_{miss} denotes the number of SNPs with missing genotype data. Finally, following arguments similar as before, under the case-control sampling, the probability of the disease given observable genotype can be approximated as

$$\mathrm{pr}(D|\mathbf{G}_{obs}, R = 1) \approx \frac{\exp(\alpha^* + m^*(\boldsymbol{\beta}, \mathbf{q}; \mathbf{G}_{obs}))}{1 + \exp(\alpha^* + m^*(\boldsymbol{\beta}, \mathbf{q}; \mathbf{G}_{obs}))}.$$

For our application, since breast cancer patients were incident cases arising in retrospective underlying cohorts, the assumption of rare disease is quite reasonable for simplification of the calculations. Further, since the 19 SNPs we are studying represent independent susceptibility loci, we can write

$$m^*(\boldsymbol{\beta}, \mathbf{q}; \mathbf{G}_{obs}) = \log(\sum_{\mathbf{G}_{miss}} \exp(m(\boldsymbol{\beta}; \mathbf{G}_{miss}, \mathbf{G}_{obs}))\mathrm{pr}(G_{miss,1}) \times ... \times \mathrm{pr}(G_{miss,K_{miss}})).$$

However, when K_{miss} is large, computation of $m^*(\boldsymbol{\beta}, \mathbf{q}; \mathbf{G}_{obs})$ can still be quite complex due to the large number of possible configuration for $\mathbf{G}_{miss}$. One solution is that one could estimate the sum by stochastic simulation where one simulates a relatively large number of samples by simulating the joint genotype data for missing SNPs. Thus one can estimate the sum by summing over only such "imputed" value of genotype instead of many theoretical combinations. For the multiplicative model, however, the computation can be remarkably simplified as we can write

$$m^*(\boldsymbol{\beta}, \mathbf{q}; \mathbf{G}_{obs}) = \log(\prod_{j: observed\ G_j} \exp(\beta_j G_{obs,j}) \prod_{j: missing\ G_j} \sum_{G_{miss,j}} \exp(\beta_j G_{miss,j}) \text{pr}(G_{miss,j}).$$

In other words, under the multiplicative model, the multivariate sum over all possible combinations of different missing SNPs can be simply computed as the product over terms that only involve univariate sum of possible genotype configuration for individual missing SNPs .

4. Asymptotic theory in Missing genotype data

For missing genotype data, we have more additional parameters, which are allele frequencies for controls, $\mathbf{q} = (q_1, ..., q_p)$. We denote E be the samples who have no missing genotype data in G and E_0 be the subset of E and be from controls. Let $\boldsymbol{\gamma} = (\alpha, \boldsymbol{\beta}^T)^T$, $\boldsymbol{\eta}_i = (\frac{\partial m(\boldsymbol{\beta}, \mathbf{q}; \mathbf{G}_{i,obs})}{\partial \mathbf{q}}^T)^T$, and $\boldsymbol{\vartheta}_i = (1, \frac{\partial m(\boldsymbol{\beta}, \mathbf{q}; \mathbf{G}_{i,obs})}{\partial \boldsymbol{\beta}}^T)^T$. Let $\boldsymbol{\gamma}_0$ and $\boldsymbol{\beta}_0$ be the true parameter values of $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ under the null model, respectively. Let n_{co} be the number of samples who have complete genotype data and $n_{co,0}$ be the number of samples who have complete genotype data and are from controls. $\hat{\mathbf{q}}_{n_{co,0}} = (\hat{q}_1, ..., \hat{q}_p)$ denotes the estimators for $\mathbf{q}$ based on samples as many as $n_{co,0}$ and $\hat{\boldsymbol{\beta}}_{n_{co}}$ denotes the estimator for $\boldsymbol{\beta}$ using samples as many as n_{co}. Let $\hat{\pi}_i^* = \pi_i^*(\hat{\boldsymbol{\beta}}_{n_{co}}, \hat{\mathbf{q}}_{n_{co,0}}; \mathbf{G}_{i,obs})$.

Lemma 1 *Under the regularity conditions similar to the assumptions in the Section 1,*

$$n^{-\frac{1}{2}}(T_{n,c} - n_c) \Rightarrow N(0, \sigma_{n,c}^2)$$

where $T_{n,c} = n^{-1/2} \sum_{i=1}^n \frac{(D_i - \hat{\pi}_i^*)(1-2\hat{\pi}_i^*)}{\hat{\pi}_i^*(1-\hat{\pi}_i^*)} I(\hat{\pi}_i^* \in R_c), n_c = \sum_{i=1}^n I(\hat{\pi}_i^* \in R_c)$,

$$Z_{n,c,i,1} = \frac{(D_i - \hat{\pi}_{i0}^*)(1 - 2\hat{\pi}_{i0}^*)}{\hat{\pi}_{i0}^*(1 - \hat{\pi}_{i0}^*)} I(\hat{\pi}_{i0}^* \in R_c), Z_{n,c,i,2} = \frac{n}{n_{co}} \frac{(D_i - \hat{\pi}_{i0}^*)}{\hat{\pi}_{i0}^*(1 - \hat{\pi}_{i0}^*)} \mathbf{v}_{n,c}^T \Omega^{-1} f_i \boldsymbol{\vartheta}_i I(i \in E),$$

$$Z_{n,c,i,3} = \frac{n}{2n_{co,0}} (G_i - 2q_{i0}) \boldsymbol{\varepsilon}_{n,c}^T \Omega^{-1} I(i \in E_0), Z_{n,c,i} = Z_{n,c,i,1} + Z_{n,c,i,2} + Z_{n,c,i,3}, \sigma_{n,c}^2 = Var(Z_{n,c,i,}),$$

$$\mathbf{v}_{n,c} = \frac{1}{n} \sum_{i=1}^n \frac{(1 - 2\pi_{i0})}{\pi_{i0}(1 - \pi_{i0})} f_{i0} \boldsymbol{\vartheta}_i I(\pi_{i0}^* \in R_c), \boldsymbol{\varepsilon}_{n,c} = \frac{1}{n} \sum_{i=1}^n \frac{(1 - 2\pi_{i0})}{\pi_{i0}(1 - \pi_{i0})} f_{i0} \boldsymbol{\eta}_i I(\pi_{i0}^* \in R_c),$$

and $\Omega = lim_n \sqrt{n}\ Var(\hat{\boldsymbol{\gamma}}_n)$.

Proof. Note that

$$n^{-\frac{1}{2}}(T_{n,c}-n_c)=n^{-\frac{1}{2}}\sum_{i=1}^{n}\frac{(D_i-\hat{\pi}_i^*)(1-2\hat{\pi}_i^*)}{\hat{\pi}_i^*(1-\hat{\pi}_i^*)}I(\hat{\pi}_i^*\in R_c). \tag{4.6}$$

Define

$$S_{n,c}(\boldsymbol{\gamma},\mathbf{q})=\sum_{i=1}^{n}\frac{(D_i-\pi_i^*)(1-2\pi_i^*)}{\pi_i^*(1-\pi_i^*)}I(\pi_{i0}^*\in R_c)=\sum_{i=1}^{n}(\frac{D_i}{\pi_i^*}+\frac{1-D_i}{1-\pi_i^*}-2)I(\pi_{i0}^*\in R_c).$$

To obtain the asymptotic distribution of $T_{n,c}$, we shall first prove that

$$n^{-\frac{1}{2}}S_{n,c}(\hat{\boldsymbol{\gamma}}_{n_{co}},\hat{\mathbf{q}}_{n_{co,0}})\Rightarrow N(0,\sigma_{n,c}^2) \tag{4.7}$$

and then show that

$$S_{n,c}(\hat{\boldsymbol{\gamma}}_{n_{co}},\hat{\mathbf{q}}_{n_{co,0}})-[T_{n,c}-n_c]=o_{\mathbb{P}}(n^{1/2}). \tag{4.8}$$

To prove (4.7), note that local expansion of $S_{n,c}(\boldsymbol{\gamma},\mathbf{q})$ around $\boldsymbol{\gamma}_0$ and $\mathbf{q}_0$ gives

$$n^{-\frac{1}{2}}S_{n,c}(\hat{\boldsymbol{\gamma}}_{n_{co}},\hat{\mathbf{q}}_{n_{co,0}})=n^{-\frac{1}{2}}S_{n,c}(\boldsymbol{\gamma}_0,\mathbf{q}_0)+n^{-1}\frac{\partial S_{n,c}(\boldsymbol{\gamma},\mathbf{q})}{\partial\boldsymbol{\gamma}}|_{\boldsymbol{\gamma}_0}n^{\frac{1}{2}}(\hat{\boldsymbol{\gamma}}_{n_{co}}-\boldsymbol{\gamma}_0)+n^{-1}\frac{\partial S_{n,c}(\boldsymbol{\gamma},\mathbf{q})}{\partial\mathbf{q}}|_{\mathbf{q}_0}n^{\frac{1}{2}}(\hat{\mathbf{q}}_{n_{co,0}}-\mathbf{q}_0)+o_p(1).$$

Meanwhile,

$$n^{-1}\frac{\partial S_{n,c}(\boldsymbol{\gamma},\mathbf{q})}{\partial\mathbf{q}}|_{\mathbf{q}_0}=n^{-1}\sum_{i=1}^{n}(\frac{(1-D_i)}{(1-\pi_{i0}^*)^2}-\frac{D_i}{\pi_{i0}^2})f_{i0}\boldsymbol{\eta}_iI(\pi_{i0}^*\in R_c)=\boldsymbol{\delta}_{n,c}-\boldsymbol{\varepsilon}_{n,c}$$

with

$$\boldsymbol{\delta}_{n,c}=-n^{-1}\sum_{i=1}^{n}\frac{(D_i-\pi_{i0}^*)(1-2\pi_{i0}^*+2{\pi_{i0}^*}^2)}{{\pi_{i0}^*}^2(1-\pi_{i0}^*)^2}f_{i0}\boldsymbol{\eta}_iI(\pi_{i0}^*\in R_c)$$

and

$$\boldsymbol{\varepsilon}_{n,c}=n^{-1}\sum_{i=1}^{n}\frac{1-2\pi_{i0}^*}{\pi_{i0}^*(1-\pi_{i0}^*)}f_{i0}\boldsymbol{\eta}_iI(\pi_{i0}^*\in R_c).$$

It can be shown that we have $\lim_{n\to\infty}\boldsymbol{\delta}_{n,c}=\mathbf{0}$.

In the proof of Proposition 1, we show that asymptotically, $n^{-1}\frac{\partial S_{n,c}(\boldsymbol{\gamma},\mathbf{q})}{\partial\boldsymbol{\gamma}}|_{\boldsymbol{\gamma}_0}\approx-\mathbf{v}_{n,c}$. Thus the limiting distribution of $n^{-\frac{1}{2}}S_{n,c}(\hat{\boldsymbol{\gamma}}_{n_{co}},\hat{\mathbf{q}}_{n_{co,0}})$ is the same as the limiting distribution of

$$n^{-\frac{1}{2}}S_{n,c}(\boldsymbol{\gamma}_0,\mathbf{q}_0)-\mathbf{v}_{n,c}^Tn^{\frac{1}{2}}(\hat{\boldsymbol{\gamma}}_{n_{co}}-\boldsymbol{\gamma}_0)-\boldsymbol{\varepsilon}_{n,c}^Tn^{\frac{1}{2}}(\hat{\mathbf{q}}_{n_{co,0}}-\hat{\mathbf{q}}_0).$$

Since $\hat{\boldsymbol{\gamma}}_{n_{co}} - \boldsymbol{\gamma}_0 = -[\frac{\partial^2 l_{n_{co}}}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}^T}|\boldsymbol{\gamma}^*_{n_{co}}]^{-1}\frac{\partial l_{n_{co}}}{\partial\boldsymbol{\gamma}}|_{\boldsymbol{\gamma}_0}$ where $\boldsymbol{\gamma}^*_{n_{co}}$ lies between $\hat{\boldsymbol{\gamma}}_{n_{co}}$ and $\boldsymbol{\gamma}_0$,

$$\mathbf{v}^T_{n,c} n^{\frac{1}{2}}(\hat{\boldsymbol{\gamma}}_{n_{co}} - \boldsymbol{\gamma}_0) = \mathbf{v}^T_{n,c}(B^*_{n_{co}})^{-1}\frac{n^{\frac{1}{2}}}{n_{co}}\frac{\partial l_{n_{co}}}{\partial\boldsymbol{\gamma}}|_{\boldsymbol{\gamma}_0},$$

where $B^*_{n_{co}} = -n^{-1}_{co}\frac{\partial^2 l_{n_{co}}}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}^T}|\boldsymbol{\gamma}^*_{n_{co}}$. Since $\lim_{n_{co}} B^*_{n_{co}} = \Omega$, the limiting distribution of $n^{-\frac{1}{2}}(T_{n,c} - n_c)$ is the same as the limiting distribution of

$$n^{-\frac{1}{2}} S_{n,c}(\boldsymbol{\gamma}_0, \mathbf{q}_0) - \frac{\sqrt{n}}{n_{co}}\mathbf{v}^T_{n,c}\Omega^{-1}\frac{\partial l_{n_{co}}}{\partial\boldsymbol{\gamma}}|_{\boldsymbol{\gamma}_0} - \boldsymbol{\varepsilon}^T_{n,c} n^{\frac{1}{2}}(\hat{\mathbf{q}}_{n_{co,0}} - \hat{\mathbf{q}}_0) = n^{-\frac{1}{2}}\sum_{i=1}^{n}(Z_{n,c,i,1} + Z_{n,c,i,2} + Z_{n,c,i,3})$$

where

$$Z_{n,c,i,1} = \frac{(D_i - \pi^*_{i0})}{\pi^*_{i0}(1 - \pi^*_{i0})}(1 - 2\pi^*_{i0})I(\pi^*_{i0} \in R_c)$$

$$Z_{n,c,i,2} = \frac{n}{n_{co}}\frac{(D_i - \pi^*_{i0})}{\pi^*_{i0}(1 - \pi^*_{i0})}\mathbf{v}^T_{n,c}\Omega^{-1} f_i \boldsymbol{\vartheta}_i I(i \in E)$$

$$Z_{n,c,i,3} = \frac{n}{2n_{co,0}}(G_i - 2q_{i0})\boldsymbol{\varepsilon}^T_{n,c} I(i \in E_0).$$

Define $Z_{n,c,i} = Z_{n,c,i,1} + Z_{n,c,i,2} + Z_{n,c,i,3}$. Then $Z_{n,c,1}, ..., Z_{n,c,n}$ are independent random variables. Thus with the similar arguments shown in the proof of Proposition 1, Liapounouv's Central Limit Theorem is applied. Thus, the asymptotic distribution of $n^{-\frac{1}{2}}(T_{n,c} - n_c)$ is $N(0, \frac{1}{n}\sum_{i=1}^{n}\sigma^2_{n,c,i})$ where $\sigma^2_{n,c,i} = Var(Z_{n,c,i})$.

We now prove (4.8). Observe that

$$S_{n,c}(\hat{\boldsymbol{\gamma}}_{n_{co}}, \hat{\mathbf{q}}_{n_{co,0}}) - [T_{n,c} - n_c] = \sum_{i=1}^{n}\frac{(D_i - \hat{\pi}^*_i)(1 - 2\hat{\pi}^*_i)}{\hat{\pi}^*_i(1 - \hat{\pi}^*_i)}[I(\pi^*_{i0} \in R_c) - I(\hat{\pi}^*_i \in R_c)] := \sum_{i=1}^{n}\Xi_i\tau_i.$$

Let $\Upsilon_i = 1/\sqrt{n}\sum_{j=1}^{i}\Xi_j$. Then standard arguments show that $\{\Upsilon_i\}_{i=1}^{n}$ converges weakly to a centered Gaussian process on $D[0,1]$ with the Skorohod topology. Meanwhile, it is easy to see that

$$\max_{1\leqslant i\leqslant n}|\tau_i| = o_{\mathbb{P}}(1) \text{ for continuously differentiable } m(\cdot).$$

Hence by a classic random change of time argument (see for instance Billingsley 1999, p.151 or Lemma 4.1 of Moore and Spruill, 1975), we have (4.8) follows.

Note : It is expected that $Cov(Z_{n,c,i,2}, Z_{n,c,i,3})$ is non-zero since the samples for estimating $\mathbf{q}$ are the subset of samples for estimating $\boldsymbol{\beta}$. Thus we use the sandwich estimator to estimate $Var(Z_{n,c,i})$.

5. Adjusting covariates

When there are covariates $\mathbf{X}$ which affect disease risk but are not used for constructing a risk prediction model, we are interested in testing adequacy of a risk model of $\mathbf{G}$ adjusting $\mathbf{X}$. Let $\mathbf{X}$ be categorical covariate having K patterns. First, we could model that the odds ratio of $\mathbf{G}$ is homogeneous across the configurations of covariate $\mathbf{X}$. Then we modify model (1) in the main document as follows

$$\pi = \text{pr}(D = 1|\mathbf{G}) = F(\alpha_k + m(\boldsymbol{\beta}; \mathbf{G}))$$

for $k = 1, ..., K$ and apply our proposed procedures.

Second, we could model that the odds ratios of $\mathbf{G}$ are heterogeneous across the configurations of covariate $\mathbf{X}$. For samples such that $\mathbf{X} = \mathbf{x}_k$, we fit model (1) in the main document, obtain $T_{n^{(k)},c}$ from $\hat{\boldsymbol{\beta}}_{n^{(k)}}$, $n_c^{(k)}$, and $\sigma_{n^{(k)},c}$ for $c = 1, ..., C$ where $n^{(k)}$ is the sample size for which $\mathbf{X} = \mathbf{x}_k$. Then T_n could be modified as

$$T_n = \max_c \left| \frac{\sum_{k=1}^{K} {n^{(k)}}^{-\frac{1}{2}} (T_{n^{(k)},c} - n_c^{(k)})}{\sum_{k=1}^{K} \sigma_{n^{(k)},c}} \right|$$

where $T_{n^{(k)},c} = \sum_{i \text{ s.t. } \mathrm{x}_i = \mathbf{x}_k} \frac{(D_i - \hat{\pi}_i)^2}{\hat{\pi}_i \times (1 - \hat{\pi}_i)} I(m(\hat{\boldsymbol{\beta}}_{n^{(k)}}; \mathbf{G}_i) \in R_c)$. Note that we construct risk region based on $m(.)$ function rather than $\hat{\pi}$ since we want to construct risk regions such that for given c, the distribution of samples which fall in the risk region is as uniform as possible across the values of $\mathbf{X}$. $\frac{\sum_{k=1}^{K} {n^{(k)}}^{-\frac{1}{2}} (T_{n^{(k)},c} - n_c^{(k)})}{\sum_{k=1}^{K} \sigma_{n^{(k)},c}}$ is a Gaussian stochastic process. Therefore, we can evaluate p-value using the similar way.

6. Parametric bootstrap

We describe how to analyze BPC3 dataset using parametric bootstrap.

1. Form estimates $\hat{\boldsymbol{\beta}}_{n_{co}}$ and $\hat{\mathbf{q}}_{n_{co,0}}$ and therefore $\hat{\pi}^*$ by fitting a null model on $\mathbf{G}_{obs}$ and D where $\hat{\boldsymbol{\beta}}_{n_{co}}$ and $\hat{\mathbf{q}}_{n_{co,0}}$ denote the estimators for $\boldsymbol{\beta}$ using samples as many as n_{co} which is the size of the samples who have complete genotype data and $\mathbf{q}$ using samples as many as $n_{co,0}$, the size of samples who have complete genotype data and are from controls, respectively.

2. For each c, calculate $T_{n,c} = \sum_{i=1}^{n} \frac{(D_i - \hat{\pi}_i^*)^2}{\hat{\pi}_i^* \times (1-\hat{\pi}_i^*)} I(\hat{\pi}_i^* \in R_c)$ and normalize $T_{n,c}$ such that mean is zero and variance is one.

3. Let $\tilde{T}_{n,c}$ be the normalized $T_{n,c}$ and set the observed statistic to be

$$T_n = \max_c |\tilde{T}_{n,c}|.$$

4. Generate disease-status variable D$^* = (D_1^*, ..., D_n^*)$ from Bernoulli distribution with success probabilities as the fitted values $\hat{\boldsymbol{\pi}}^* = (\hat{\pi}_1^*, ..., \hat{\pi}_n^*)$.

5. Form a test statistic

$$T_{n,c}^* = \sum_{i=1}^{n} \frac{(D_i^* - \hat{\pi}_i^*)^2}{\hat{\pi}_i^* \times (1 - \hat{\pi}_i^*)} I(\hat{\pi}_i^* \in R_c)$$

as above.

6. Normalize $T_{n,c}^*$ such that $\tilde{T}_{n,c}^* = \frac{n^{-\frac{1}{2}}(T_{n,c}^* - n_c)}{\sigma^{*2}_{n,c}}$ where $n_c = \sum_{i=1}^{n} I(\hat{\pi}_i^* \in R_c)$ and $\sigma^{*2}_{n,c} = Var(Z_{n,c,i,1})$ where $Z_{n,c,i,1} = \frac{(D_i^* - \pi_{i0}^*)}{\pi_{i0}^*(1-\pi_{i0}^*)}(1 - 2\pi_{i0}^*) I(\pi_{i0}^* \in R_c)$. Note that we do not need to take into account the uncertainty of the estimation of $\boldsymbol{\beta}$ and $\mathbf{q}$ since they are estimated using D.

7. Form a statistic

$$T_n^* = \max_c |\tilde{T}_{n,c}^*|.$$

8. Repeat steps 4-7 a total of B times to obtain null statistics T_n^{*b} for $b = 1, ..., B$.

9. Compute the p-value as

$$p - value = \frac{\#\{T_n^{*b} > T_n; b = 1, ..., B\}}{B}.$$

References

Billingsley, P. (1968) *Convergence of Probability Measures.* Wiley, New York.

Moore, D. S., and Spruill, M. C. (1975) Unified large-sample theory of general chi-squared statistics for tests of fit. *The Annals of Statistics* **3**, 599-616.

Table 1. *Distribution of cases and controls by cohorts in BPC3 study*

	Cases					Controls				
Cohort	CPS2	EPIC	MEC	NHS	All	CPS2	EPIC	MEC	NHS	All
Complete data	245	1690	422	1811	4168	257	1570	431	2672	4930
Complete and missing data	786	4155	553	2561	10525	870	5238	574	3843	8035