Biometrika (2017), **xx**, x, *p.* 1
Printed in Great Britain

Supplementary Material to *Semiparametric Analysis of Complex Polygenic Gene-Environment Interactions in Case-Control Studies*

BY ODILE STALDER

Department of Clinical Research, and Institute of Social and Preventive Medicine, University of Bern, 3012 Bern, Switzerland
Odile.Stalder@gmail.com

ALEX ASHER, LIANG LIANG, RAYMOND J. CARROLL

Department of Statistics, Texas A&M University, 3143 TAMU, College Station, TX 77843-3143, U.S.A.
alexasher@stat.tamu.edu liang@stat.tamu.edu carroll@stat.tamu.edu

YANYUAN MA

Department of Statistics, Penn State University, University Park, PA 16802, U.S.A.
yanyuanma@gmail.com

AND NILANJAN CHATTERJEE

Department of Biostatistics, Bloomberg School of Public Health and Department of Oncology, School of Medicine, Johns Hopkins University, Baltimore, MD 21205, U.S.A.
nchatte2@jhu.edu

S·1. PROOF OF THEOREM 1

SKETCH OF TECHNICAL ARGUMENTS

Necessary U-Statistic theory

Consider the case of one sample. Let $Z_1,\ldots,Z_n$ be independent and identically distributed. Let $h_*(\cdot)$ be a function such that $E\{h_*(Z_1,Z_2)\}=0$. Define

$$U_{n*}=\textstyle\sum_{i=1}^{n}\sum_{j\neq i}^{n}h_*(Z_i,Z_j)/\{n(n-1)\}=\sum_{i=1}^{n}\sum_{j<i}h_*(Z_i,Z_j)/\{n(n-1)/2\}.$$

If $h_*(z_1,z_2)\neq h_*(z_2,z_1)$, we make it symmetric in its arguments by noticing that if

$$h(z_1,z_2)=\{h_*(z_1,z_2)+h_*(z_2,z_1)\}/2,$$

then

$$U_{n*}=U_n=\textstyle\sum_{i=1}^{n}\sum_{j\neq i}^{n}h(Z_i,Z_j)/\{n(n-1)\}.$$

We recognize U_n as a U-statistic of order 2 with a symmetric kernel $h(\cdot)$. Define

$$h_1(z)=2E\{h(z,Z_2)\}. \tag{1}$$

Then, as in Theorem 12.3 of Van der Vaart (1998),

$$n^{1/2}U_n=n^{-1/2}\textstyle\sum_{i=1}^{n}h_1(Z_i)+o_p(1). \tag{2}$$

Next we consider a special case of two samples, namely the n_0 controls and n_1 cases, denoted as $(U_1,\ldots,U_{n_0})$ and $(V_1,\ldots,V_{n_1})$, respectively, with $n=n_0+n_1$. The U-statistic of interest is

$$U_n=(n_0n_1)^{-1}\textstyle\sum_{i=1}^{n_0}\sum_{j=1}^{n_1}I(D_i=0)I(D_j=1)h(U_i,V_j), \tag{3}$$

where $0=E\{h(U_i,V_j)\mid D_i=0,D_j=1\}$. Let $n_0/n\to\lambda$ and $n_1/n\to 1-\lambda$, with $0<\lambda<1$. Define

$$\begin{aligned}h_{1,0}(u)&=E\{h(u,V)\mid D=1\};\\ h_{0,1}(v)&=E\{h(U,v)\mid D=0\}.\end{aligned}$$

Then, from Chapter 12.2 of Van der Vaart (1998),

$$n^{1/2}U_n=n^{1/2}n_0^{-1}\textstyle\sum_{i=1}^{n_0}I(D_i=0)h_{1,0}(U_i)+n^{1/2}n_1^{-1}\sum_{j=1}^{n_1}I(D_j=1)h_{0,1}(V_j)+o_p(1). \tag{4}$$

Preliminary Lemma

Let the data be $Z_i=(D_i,G_i,X_i)$ for $i=1,\ldots,n$, ordered so that the first n_0 observations are the controls, and the last n_1 observations are the cases.

Define $n_d=c_dn$.

In the proofs, for generic functions $T(\cdot)$ and $P(\cdot)$, we need to deal with terms

$$\begin{aligned}\mathcal{D}_n(P,T)&=\textstyle\sum_{d=0}^{1}(\pi_d/n_d)n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{r=0}^{1}I(D_j=d)\\ &\qquad\times P(X_i)\left\{T(r,G_j,X_i)-T_E(r,D_j,X_i)\right\}\\ &=\textstyle\sum_{t=0}^{1}\sum_{d=0}^{1}(\pi_d/n_d)n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{r=0}^{1}I(D_i=t,D_j=d)\\ &\qquad\times P(X_i)\left\{T(r,G_j,X_i)-T_E(r,d,X_i)\right\},\end{aligned}$$

where

$$T_E(r,d,x)=E\{T(r,G,x)\mid D=d\}.$$

We will use repeatedly the fact that for any constant x,

$$0=E\left[P(x)\left\{T(r,G,x)-T_E(r,d,x)\right\}\mid D=d\right]. \tag{5}$$

We will make the following notational convention. We define

$$E\left[P(X)\left\{T(r,g_i,X)-T_E(r,d,X)\right\}\mid D=t\right] \tag{6}$$

to mean

$$E\left[P(X)\left\{T(r,g,X)-T_E(r,d,X)\right\}\mid D=t\right]_{g=G_i}.$$

Similarly, $E\left[P(x_i)\left\{T(r,G,x_i)-T_E(r,d,x_i)\right\}\mid D=t\right]$ is

$$E\left[P(x)\left\{T(r,G,x)-T_E(r,d,x)\right\}\mid D=t\right]_{x=X_i}.$$

Below, we will prove the following Lemma, which relies of U-statistics of order 2 for one sample and U-statistics of order 1 for independent samples, namely the cases and the controls. We use the notation defined at (6).

LEMMA 1. *Define* $Z_i = (D_i, G_i, X_i)$. *As* $n \to \infty$ *in such a way that* $n_d = c_d n$ *for* $0 < c_0, c_1 < 1$,

$$
\begin{aligned}
&n^{1/2}\mathcal{D}_n(P,T) \\
&\quad = n^{-1/2}\textstyle\sum_{i=1}^{n_0}\sum_{d=0}^{1}\sum_{r=0}^{1}\{c_d\pi_{d_i}/c_{d_i}\}E\{P(X)T(r,g_i,X) \mid D=d\} \\
&\qquad -n^{-1/2}n_0 E\left[P(X)\left\{\pi_0\textstyle\sum_{r=0}^{1}T_E(r,0,X)+\pi_1\sum_{r=0}^{1}T_E(r,1,X)\right\} \mid D=0\right] \\
&\qquad -n^{-1/2}n_1 E\left[P(X)\left\{\pi_0\textstyle\sum_{r=0}^{1}T_E(r,0,X)+\pi_1\sum_{r=0}^{1}T_E(r,1,X)\right\} \mid D=1\right] + o_p(1).
\end{aligned}
$$

Proof of Lemma 1

Now, since there are only n terms with $i = j$, whereas the leading terms before the summations are $O(n^{-2})$, and because $(n-1)^{-1} - n^{-1} = O(n^{-2})$, and because the first n_0 observations are controls, to order $n^{1/2}$, analyzing D_n is equivalent to analyzing

$$\mathcal{D}_n(P,T) = \textstyle\sum_{t=0}^{1}\sum_{d=0}^{1}\mathcal{D}_n(P,T,t,d) + o_p(n^{-1}),$$

where

$$
\begin{aligned}
\mathcal{D}_n(P,T,0,0) &= (\pi_0/n_0)n^{-1}\textstyle\sum_{i=1}^{n_0}\sum_{j=1,j\neq i}^{n_0}I(D_i=0,D_j=0) \\
&\qquad \times P(X_i)\textstyle\sum_{r=0}^{1}\{T(r,G_j,X_i)-T_E(r,0,X_i)\} \\
&= \{n_0(n_0-1)\}^{-1}\textstyle\sum_{i=1}^{n_0}\sum_{j=1,j\neq i}^{n_0}I(D_i=0,D_j=0) \\
&\qquad \times(\pi_0 c_0)P(X_i)\textstyle\sum_{r=0}^{1}\{T(r,G_j,X_i)-T_E(r,0,X_i)\} + O_p(n^{-1});
\end{aligned}
$$

$$
\begin{aligned}
\mathcal{D}_n(P,T,0,1) &= (\pi_1/n_1)n^{-1}\textstyle\sum_{i=1}^{n_0}\sum_{j=n_0+1}^{n}I(D_i=0,D_j=1) \\
&\qquad \times P(X_i)\textstyle\sum_{r=0}^{1}\{T(r,G_j,X_i)-T_E(r,1,X_i)\} \\
&= (n_0 n_1)^{-1}\textstyle\sum_{i=1}^{n_0}\sum_{j=n_0+1}^{n}I(D_i=0,D_j=1) \\
&\qquad \times(\pi_1 c_0)P(X_i)\textstyle\sum_{r=0}^{1}\{T(r,G_j,X_i)-T_E(r,1,X_i)\};
\end{aligned}
$$

$$
\begin{aligned}
\mathcal{D}_n(P,T,1,0) &= (\pi_0/n_0)n^{-1}\textstyle\sum_{i=n_0+1}^{n}\sum_{j=1}^{n_0}I(D_i=1,D_j=0) \\
&\qquad \times P(X_i)\textstyle\sum_{r=0}^{1}\{T(r,G_j,X_i)-T_E(r,0,X_i)\} \\
&= (n_0 n_1)^{-1}\textstyle\sum_{i=1}^{n_0}\sum_{j=n_0+1}^{n}I(D_i=0,D_j=1) \\
&\qquad \times(\pi_0 c_1)P(X_j)\textstyle\sum_{r=0}^{1}\{T(r,G_i,X_j)-T_E(r,0,X_j)\};
\end{aligned}
$$

$$
\begin{aligned}
\mathcal{D}_n(P,T,1,1) &= (\pi_1/n_1)n^{-1}\textstyle\sum_{i=n_0+1}^{n}\sum_{j=n_0+1,j\neq i}^{n}I(D_i=1,D_j=1) \\
&\qquad \times P(X_i)\textstyle\sum_{r=0}^{1}\{T(r,G_j,X_i)-T_E(r,1,X_i)\} \\
&= \{n_1(n_1-1)\}^{-1}\textstyle\sum_{i=n_0+1}^{n}\sum_{j=n_0+1,j\neq i}^{n}I(D_i=1,D_j=1) \\
&\qquad \times(\pi_1 c_1)P(X_i)\textstyle\sum_{r=0}^{1}\{T(r,G_j,X_i)-T_E(r,1,X_i)\} + O_p(n^{-1}).
\end{aligned}
$$

Now, $\mathcal{D}_n(P,T,1,0)$ and $\mathcal{D}_n(P,T,0,1)$ are U-statistics of order 1 for 2 independent samples, while $\mathcal{D}_n(P,T,0,0)$ and $\mathcal{D}_n(P,T,1,1)$ are U-statistics of order 2 for a single sample, all with asymmetric kernels.

We next analyze $\mathcal{D}_n(P,T,0,1)$. The term $\mathcal{D}_n(P,T,0,1)$ has kernel

$$h(Z_i,Z_j,0,1) = (\pi_1 c_0)P(X_i)\textstyle\sum_{r=0}^{1}\{T(r,G_j,X_i)-T_E(r,1,X_i)\}.$$

Then

$$\begin{aligned} h_{1,0}(u,0,1) &= E\{h(u,Z_j,0,1) \mid D_j = 1\} = 0, \text{ by (5);} \\ h_{0,1}(v,0,1) &= E\{h(Z_i,v) \mid D_i = 0\} \\ &= (\pi_1 c_0) E\left[P(X){\textstyle\sum_{r=0}^{1}} \{T(r,v,X) - T_E(r,1,X)\} \mid D = 0\right]. \end{aligned}$$

Thus, from (4),

$$\begin{aligned} n^{1/2}\mathcal{D}_n(P,T,0,1) &= (n^{1/2}/n_1){\textstyle\sum_{j=n_0+1}^{n}} h_{0,1}(Z_j,0,1) + o_p(1) \\ &= n^{-1/2}{\textstyle\sum_{j=n_0+1}^{n}} c_1^{-1} h_{0,1}(Z_j,0,1) + o_p(1). \qquad (7) \end{aligned}$$

In the notation defined at (6),

$$\begin{aligned} n^{1/2}\mathcal{D}_n(P,T,0,1) &= n^{-1/2}{\textstyle\sum_{j=n_0+1}^{n}} (\pi_1 c_0/c_1) I(D_j = 1) \\ &\qquad \times{\textstyle\sum_{r=0}^{1}} E\left[P(X)\left\{T(r,g_j,X) - T_E(r,1,X)\right\} \mid D = 0\right] + o_p(1) \\ &= n^{-1/2}{\textstyle\sum_{i=n_0+1}^{n}} I(D_i = 1) \qquad (8) \\ &\qquad \times(\pi_1 c_0/c_1){\textstyle\sum_{r=0}^{1}} E\left[P(X)\left\{T(r,g_i,X) - T_E(r,1,X)\right\} \mid D = 0\right] + o_p(1). \end{aligned}$$

Now consider he term $\mathcal{D}_n(P,T,1,0)$, which has kernel

$$h(Z_i,Z_j,1,0) = (\pi_0 c_1) P(X_j){\textstyle\sum_{r=0}^{1}} \{T(r,G_i,X_j) - T_E(r,0,X_j)\}.$$

Then

$$\begin{aligned} h_{1,0}(u,1,0) &= E\{h(u,Z_j,1,0) \mid D_j = 1\} \\ &= (\pi_0 c_1) E\left[P(X_j){\textstyle\sum_{r=0}^{1}} \{T(r,u,X_j) - T_E(r,0,X_j)\} \mid D_j = 1\right]; \\ h_{0,1}(v,1,0) &= E\{h(Z_i,v,1,0) \mid D_i = 0\} = 0, \text{ by (5).} \end{aligned}$$

Thus, from (4),

$$\begin{aligned} n^{1/2}\mathcal{D}_n(P,T,1,0) &= (n^{1/2}/n_0){\textstyle\sum_{i=1}^{n_0}} h_{1,0}(Z_i,1,0) + o_p(1) \\ &= n^{-1/2}{\textstyle\sum_{i=1}^{n_0}} c_0^{-1} h_{1,0}(Z_j,1,0) + o_p(1). \qquad (9) \end{aligned}$$

In the notation defined at (6),

$$\begin{aligned} n^{1/2}\mathcal{D}_n(P,T,1,0) &= n^{-1/2}{\textstyle\sum_{i=1}^{n_0}} I(D_i = 0)(\pi_0 c_1/c_0) \qquad (10) \\ &\qquad \times{\textstyle\sum_{r=0}^{1}} E\left[P(X)\left\{T(r,g_i,X) - T_E(r,0,X)\right\} \mid D = 1\right] + o_p(1). \end{aligned}$$

We next analyze $\mathcal{D}_n(P,T,0,0)$, which is a U-statistic of order 2 but with an asymmetric kernel

$$h_*(Z_i,Z_j,0,0) = I(D_i = D_j = 0)(\pi_0 c_0) P(X_i){\textstyle\sum_{r=0}^{1}} \{T(r,G_j,X_i) - T_E(r,0,X_i)\}.$$

To make this a symmetric kernel, we define

$$\begin{aligned} h(Z_i,Z_j,0,0) = (1/2) I(D_i = D_j = 0)(\pi_0 c_0)\big[&P(X_i){\textstyle\sum_{r=0}^{1}} \{T(r,G_j,X_i) - T_E(r,0,X_i)\} \\ &+ P(X_j){\textstyle\sum_{r=0}^{1}} \big\{T(r,G_i,X_j) - T_E(r,0,X_j)\big\}\big]. \end{aligned}$$

We now apply (1), so that

$$\begin{aligned} h_1(z,0,0) &= (\pi_0 c_0)E\left[P(x){\textstyle\sum_{r=0}^{1}}\left\{T(r,G,x)-T_E(r,0,x)\right\} \mid D=0\right] \\ &\quad +(\pi_0 c_0)E\left[P(X){\textstyle\sum_{r=0}^{1}}\left\{T(r,g,X)-T_E(r,0,X)\right\} \mid D=0\right] \\ &= (\pi_0 c_0)E\left[P(X){\textstyle\sum_{r=0}^{1}}\left\{T(r,g,X)-T_E(r,0,X)\right\} \mid D=0\right], \text{ by (5).} \end{aligned}$$

From (2), this means that

$$\begin{aligned} n^{1/2}\mathcal{D}_n(P,T,0,0) &= (n^{1/2}/n_0){\textstyle\sum_{i=1}^{n_0}} h_1(Z_i,0,0)+o_p(1) \\ &= (n^{-1/2}/c_0){\textstyle\sum_{i=1}^{n_0}} h_1(Z_i,0,0)+o_p(1). \end{aligned}$$

Thus, in the notation defined at (6),

$$\begin{aligned} n^{1/2}\mathcal{D}_n(P,T,0,0) = n^{-1/2}&{\textstyle\sum_{i=1}^{n_0}} I(D_i=0)\pi_0 \\ &\times{\textstyle\sum_{r=0}^{1}} E\left[P(X)\left\{T(r,g_i,X)-T_E(r,0,X)\right\} \mid D=0\right]+o_p(1). \end{aligned} \tag{11}$$

We next analyze $\mathcal{D}_n(P,T,1,1)$, which is a U-statistic of order 2 but with an asymmetric kernel

$$h_*(Z_i,Z_j,1,1)=I(D_i=D_j=1)(\pi_1 c_1)P(X_i){\textstyle\sum_{r=0}^{1}}\left\{T(r,G_j,X_i)-T_E(r,1,X_i)\right\}.$$

To make this a symmetric kernel, we define

$$\begin{aligned} h(Z_i,Z_j,1,1)=(1/2)I(D_i=D_j=1)(\pi_1 c_1)\big[&P(X_i){\textstyle\sum_{r=0}^{1}}\left\{T(r,G_j,X_i)-T_E(r,1,X_i)\right\} \\ &+P(X_j){\textstyle\sum_{r=0}^{1}}\left\{T(r,G_i,X_j)-T_E(r,1,X_j)\right\}\big]. \end{aligned}$$

We now apply (1), so that

$$\begin{aligned} h_1(z,1,1) &= (\pi_1 c_1)E\left[P(x){\textstyle\sum_{r=0}^{1}}\left\{T(r,G,x)-T_E(r,1,x)\right\} \mid D=1\right] \\ &\quad +(\pi_1 c_1)E\left[P(X){\textstyle\sum_{r=0}^{1}}\left\{T(r,g,X)-T_E(r,1,X)\right\} \mid D=1\right] \\ &= (\pi_1 c_1)E\left[P(X){\textstyle\sum_{r=0}^{1}}\left\{T(r,g,X)-T_E(r,1,X)\right\} \mid D=1\right], \text{ by (5).} \end{aligned}$$

From (2),

$$\begin{aligned} n^{1/2}\mathcal{D}_n(P,T,1,1) &= (n^{1/2}/n_1){\textstyle\sum_{i=n_0+1}^{n}} h_1(Z_i,1,1)+o_p(1) \\ &= (n^{-1/2}/c_1){\textstyle\sum_{i=n_0+1}^{n}} h_1(Z_i,1,1)+o_p(1). \end{aligned}$$

Thus, in the notation at (6),

$$\begin{aligned} n^{1/2}\mathcal{D}_n(P,T,1,1) = n^{-1/2}&{\textstyle\sum_{i=n_0+1}^{n}} I(D_i=1) \\ &\times\pi_1 E\left[P(X){\textstyle\sum_{r=0}^{1}}\left\{T(r,g_i,X)-T_E(r,1,X)\right\} \mid D=1\right]+o_p(1). \end{aligned} \tag{12}$$

Collecting the terms (8), (10), (11) and (12), we get that

$$
\begin{aligned}
&n^{1/2}\mathcal{D}_n(P,T)\\
&= n^{-1/2}\textstyle\sum_{i=1}^{n_0} I(D_i=0)\pi_0\sum_{r=0}^{1} E\left[P(X)\left\{T(r,g_i,X)-T_E(r,0,X)\right\} \mid D=0\right]\\
&\quad + n^{-1/2}\textstyle\sum_{i=1}^{n_0} I(D_i=0)(\pi_0 c_1/c_0)\sum_{r=0}^{1} E\left[P(X)\left\{T(r,g_i,X)-T_E(r,0,X)\right\} \mid D=1\right]\\
&\quad + n^{-1/2}\textstyle\sum_{i=n_0+1}^{n} I(D_i=1)\pi_1 E\left[P(X)\sum_{r=0}^{1}\left\{T(r,g_i,X)-T_E(r,1,X)\right\} \mid D=1\right]\\
&\quad + n^{-1/2}\textstyle\sum_{i=n_0+1}^{n} I(D_i=1)(\pi_1 c_0/c_1)\sum_{r=0}^{1} E\left[P(X)\left\{T(r,g_i,X)-T_E(r,1,X)\right\} \mid D=0\right]\\
&\quad + o_p(1).
\end{aligned}
$$

This in turn is seen to be

$$
n^{1/2}\mathcal{D}_n(P,T)=\mathcal{G}_1-\mathcal{G}_2+o_p(1),
$$

where

$$
\begin{aligned}
\mathcal{G}_1(P,T) &= n^{-1/2}\textstyle\sum_{i=1}^{n}\sum_{d=0}^{1}\sum_{r=0}^{1}(c_d\pi_{d_i}/c_{d_i})E\{P(X)T(r,g_i,X)\mid D=d\};\\
\mathcal{G}_2(P,T) &= n^{-1/2}\textstyle\sum_{i=n_0+1}^{n}\sum_{r=0}^{1} I(D_i=0)\pi_0 E\left\{P(X)T_E(r,0,X)\mid D=0\right\}\\
&\quad + n^{-1/2}\textstyle\sum_{i=1}^{n_0}\sum_{r=0}^{1} I(D_i=0)(\pi_0 c_1/c_0)E\left\{P(X)T_E(r,0,X)\mid D=1\right\}\\
&\quad + n^{-1/2}\textstyle\sum_{i=n_0+1}^{n}\sum_{r=0}^{1} I(D_i=1)\pi_1 E\left\{P(X)T_E(r,1,X)\mid D=1\right\}\\
&\quad + n^{-1/2}\textstyle\sum_{i=n_0+1}^{n}\sum_{r=0}^{1} I(D_i=1)(\pi_1 c_0/c_1)\sum_{r=0}^{1} E\left\{P(X)T_E(r,1,X)\mid D=0\right\}.
\end{aligned}
$$

It is easily seen that

$$
\begin{aligned}
\mathcal{G}_2(P,T) &= n^{-1/2}n_0 E\left[P(X)\left\{\pi_0\textstyle\sum_{r=0}^{1}T_E(r,0,X)\right\}\mid D=0\right]\\
&\quad + n^{-1/2}n_1 E\left[P(X)\left\{\pi_0\textstyle\sum_{r=0}^{1}T_E(r,0,X)\right\}\mid D=1\right]\\
&\quad + n^{-1/2}n_1 E\left[P(X)\left\{\pi_1\textstyle\sum_{r=0}^{1}T_E(r,1,X)\right\}\mid D=1\right]\\
&\quad + n^{-1/2}n_0 E\left[P(X)\left\{\pi_1\textstyle\sum_{r=0}^{1}T_E(r,1,X)\right\}\mid D=0\right]\\
&= n^{-1/2}n_0 E\left[P(X)\left\{\pi_0\textstyle\sum_{r=0}^{1}T_E(r,0,X)+\pi_1\sum_{r=0}^{1}T_E(r,1,X)\right\}\mid D=0\right]\\
&\quad + n^{-1/2}n_1 E\left[P(X)\left\{\pi_0\textstyle\sum_{r=0}^{1}T_E(r,0,X)+\pi_1\sum_{r=0}^{1}T_E(r,1,X)\right\}\mid D=1\right].
\end{aligned}
$$

This completes the proof of Lemma 1.

Proof of Theorem 1

With a first-order Taylor series expansion, it is readily seen that

$$
n^{-1/2}\textstyle\sum_{i=1}^{n}\left\{\dfrac{S_\Omega(D_i,G_i,X_i,\widehat{\Omega})}{S(D_i,G_i,X_i,\widehat{\Omega})}-\dfrac{S_\Omega(D_i,G_i,X_i,\Omega)}{S(D_i,G_i,X_i,\Omega)}\right\}=\Gamma_1 n^{1/2}(\widehat{\Omega}-\Omega)+o_p(1).
$$

Similarly,

$$
n^{-1/2}\textstyle\sum_{i=1}^{n}\left\{\dfrac{\widehat{R}_\Omega(X_i,\widehat{\Omega})}{\widehat{R}(X_i,\widehat{\Omega})}-\dfrac{\widehat{R}_\Omega(X_i,\Omega)}{\widehat{R}(X_i,\Omega)}\right\}=\Gamma_2 n^{1/2}(\widehat{\Omega}-\Omega)+o_p(1).
$$

In a manner similar to that of Wei et al. (2013), we have that

$$\begin{aligned} 0 = \widehat{\mathcal{S}}_n(\widehat{\Omega}) &= \widehat{\mathcal{S}}_n(\Omega) + n^{-1/2}\frac{\partial \widehat{\mathcal{S}}_n(\Omega)}{\partial \Omega^{\mathrm{T}}} n^{1/2}(\widehat{\Omega}-\Omega) + o_p(1)\\ &= \widehat{\mathcal{S}}_n(\Omega) + (\Gamma_1-\Gamma_2)n^{1/2}(\widehat{\Omega}-\Omega)+o_p(1)\\ &= \mathcal{S}_n(\Omega) - n^{-1/2}{\textstyle\sum_{i=1}^n} \left\{ \frac{\widehat{R}_\Omega(X_i,\Omega)}{\widehat{R}(X_i,\Omega)} - \frac{R_\Omega(X_i,\Omega)}{R(X_i,\Omega)}\right\}\\ &\qquad +(\Gamma_1-\Gamma_2)n^{1/2}(\widehat{\Omega}-\Omega)+o_p(1). \end{aligned} \tag{13}$$

We now analyze the second term in (13), which equals

$$\begin{aligned} & n^{-1/2}{\textstyle\sum_{i=1}^n}\left[\frac{\widehat{R}_\Omega(X_i,\Omega)-R_\Omega(X_i,\Omega)}{R(X_i,\Omega)} - \frac{R_\Omega(X_i,\Omega)\{\widehat{R}(X_i,\Omega)-R(X_i,\Omega)\}}{R^2(X_i,\Omega)}\right] + o_p(1)\\ = \quad & n^{-1/2}{\textstyle\sum_{i=1}^n} P_1(X_i,\Omega)\{\widehat{R}_\Omega(X_i,\Omega)-R_\Omega(X_i,\Omega)\}\\ & -n^{-1/2}{\textstyle\sum_{i=1}^n} P_2(X_i,\Omega)\{\widehat{R}(X_i,\Omega)-R(X_i,\Omega)\}+o_p(1). \end{aligned}$$

Thus,

$$\begin{aligned} \mathcal{C}_n &= n^{-1/2}{\textstyle\sum_{i=1}^n}\{\frac{\widehat{R}_\Omega(X_i,\Omega)}{\widehat{R}(X_i,\Omega)} - \frac{R_\Omega(X_i,\Omega)}{R(X_i,\Omega)}\}\\ &= n^{-1/2}{\textstyle\sum_{i=1}^n}\frac{\widehat{R}_\Omega(X_i,\Omega)-R_\Omega(X_i,\Omega)}{R(X_i,\Omega)}\\ &\qquad -n^{-1/2}{\textstyle\sum_{i=1}^n}\frac{R_\Omega(X_i,\Omega)}{R^2(X_i,\Omega)}\{\widehat{R}(X_i,\Omega)-R(X_i,\Omega)\}+o_p(1)\\ &= \mathcal{C}_{n1}-\mathcal{C}_{n2}+o_p(1). \end{aligned}$$

First, we calculate that

$$\begin{aligned} \widehat{R}(x,\Omega)-R(x,\Omega) &= {\textstyle\sum_{j=1}^n\sum_{r=0}^1\sum_{d=0}^1}(\pi_d/n_d)I(D_j=d)S(r,G_j,x,\Omega)\\ &\qquad -{\textstyle\sum_{r=0}^1\sum_{d=0}^1}\pi_d S_E(r,d,x,\Omega)\\ &= {\textstyle\sum_{j=1}^n}\{{\textstyle\sum_{r=0}^1\sum_{d=0}^1}(\pi_d/n_d)I(D_j=d)S(r,G_j,x,\Omega)\\ &\qquad -{\textstyle\sum_{r=0}^1\sum_{d=0}^1}(\pi_d/n_d)I(D_j=d)S_E(r,d,x,\Omega)\}\\ &= {\textstyle\sum_{d=0}^1} n_d^{-1}{\textstyle\sum_{j=1}^n\sum_{r=0}^1} I(D_j=d)\pi_d\\ &\qquad \times\{S(r,G_j,x,\Omega)-S_E(r,d,x,\Omega)\}. \end{aligned} \tag{14}$$

Similarly,

$$\begin{aligned} \widehat{R}_\Omega(x,\Omega)-R_\Omega(x,\Omega) &= {\textstyle\sum_{d=0}^1} n_d^{-1}{\textstyle\sum_{j=1}^n\sum_{r=0}^1} I(D_j=d)\pi_d\\ &\qquad \times\{S_\Omega(r,G_j,x,\Omega)-S_{E,\Omega}(r,d,x,\Omega)\}. \end{aligned} \tag{15}$$

Then, from (14) and (15),

$$\begin{aligned}
\mathcal{C}_{n1} &= n^{-1/2}{\textstyle\sum_{i=1}^{n}} P_1(X_i,\Omega)\{\widehat{R}_\Omega(X_i,\Omega) - R_\Omega(X_i,\Omega)\} \\
&= {\textstyle\sum_{d=0}^{1}}(\pi_d/n_d) n^{-1/2}{\textstyle\sum_{i=1}^{n}} P_1(X_i,\Omega) \\
&\qquad\qquad \times{\textstyle\sum_{j=1}^{n}\sum_{r=0}^{1}} I(D_j=d)\left\{S_\Omega(r,G_j,X_i,\Omega) - S_{E,\Omega}(r,D_j,X_i,\Omega)\right\}; \\
\mathcal{C}_{n2} &= n^{-1/2}{\textstyle\sum_{i=1}^{n}} P_2(X_i,\Omega)\{\widehat{R}(X_i,\Omega) - R(X_i,\Omega)\} \\
&= {\textstyle\sum_{d=0}^{1}}(\pi_d/n_d) n^{-1/2}{\textstyle\sum_{i=1}^{n}} P_2(X_i,\Omega) \\
&\qquad\qquad \times{\textstyle\sum_{j=1}^{n}\sum_{r=0}^{1}} I(D_j=d)\left\{S(r,G_j,X_i,\Omega) - S_E(r,D_j,X_i,\Omega)\right\}.
\end{aligned}$$

In the notation defined at (6),

$$\begin{aligned}
\mathcal{C}_{n1} &= n^{1/2}\mathcal{D}_n(P_1,S_\Omega) + o_p(1); \\
\mathcal{C}_{n2} &= n^{1/2}\mathcal{D}_n(P_2,S) + o_p(1).
\end{aligned}$$

Thus, with Lemma 1,

$$\begin{aligned}
\mathcal{C}_n &= \mathcal{C}_{n1} - \mathcal{C}_{n2} + o_p(1) \\
&= n^{-1/2}{\textstyle\sum_{i=1}^{n}\sum_{d=0}^{1}\sum_{r=0}^{1}}\frac{c_d\pi_{d_i}}{c_{d_i}} E\left[\{P_1(X,\Omega)S_\Omega(r,g_i,X) - P_2(X,\Omega)S(r,g_i,X)\} \mid D=d\right] \\
&\quad -n^{-1/2}n_0 E\left[P_1(X,\Omega)\left\{\pi_0{\textstyle\sum_{r=0}^{1}}S_{E,\Omega}(r,0,X) + \pi_1{\textstyle\sum_{r=0}^{1}}S_{E,\Omega}(r,1,X)\right\} \mid D=0\right] \\
&\quad -n^{-1/2}n_1 E\left[P_1(X,\Omega)\left\{\pi_0{\textstyle\sum_{r=0}^{1}}S_{E,\Omega}(r,0,X) + \pi_1{\textstyle\sum_{r=0}^{1}}S_{E,\Omega}(r,1,X)\right\} \mid D=1\right] \\
&\quad +n^{-1/2}n_0 E\left[P_2(X,\Omega)\left\{\pi_0{\textstyle\sum_{r=0}^{1}}S_{E}(r,0,X) + \pi_1{\textstyle\sum_{r=0}^{1}}S_{E}(r,1,X)\right\} \mid D=0\right] \\
&\quad +n^{-1/2}n_1 E\left[P_2(X,\Omega)\left\{\pi_0{\textstyle\sum_{r=0}^{1}}S_{E}(r,0,X) + \pi_1{\textstyle\sum_{r=0}^{1}}S_{E}(r,1,X)\right\} \mid D=1\right] + o_p(1) \\
&= n^{-1/2}{\textstyle\sum_{i=1}^{n}\sum_{d=0}^{1}\sum_{r=0}^{1}}\frac{c_d\pi_{d_i}}{c_{d_i}} E\left[\{P_1(X,\Omega)S_\Omega(r,g_i,X) - P_2(X,\Omega)S(r,g_i,X)\} \mid D=d\right] \\
&\quad -n^{-1/2}n_0 E\left\{P_1(X,\Omega)R_\Omega(X,\Omega) \mid D=0\right\} \\
&\quad -n^{-1/2}n_1 E\left\{P_1(X,\Omega)R_\Omega(X,\Omega) \mid D=1\right\} \\
&\quad +n^{-1/2}n_0 E\left\{P_2(X,\Omega)R(X,\Omega) \mid D=0\right\} \\
&\quad +n^{-1/2}n_1 E\left\{P_2(X,\Omega)R(X,\Omega) \mid D=1\right\} + o_p(1).
\end{aligned}$$

However,

$$\begin{aligned}
P_1(X,\Omega)R_\Omega(X,\Omega) &= \{R(X,\Omega)\}^{-1}R_\Omega(X,\Omega); \\
P_2(X,\Omega)R(X,\Omega) &= \{R(X,\Omega)\}^{-1}R_\Omega(X,\Omega),
\end{aligned}$$

so the last 4 terms above cancel, completing the proof of Theorem 1.

S·2. Alternative Proof Based on a Hypothetical Population

Here we give an alternative argument using the hypothetical population framework of Ma (2010). Define $\mathcal{K}_1(D,G,X,\Omega) = S_\Omega(D,G,X,\Omega)/S(D,G,X,\Omega)$ and $\mathcal{K}_2(X,\Omega) = R_\Omega(X,\Omega)/R(X,\Omega)$. Solving (7) in the main paper leads to the expansion

$$
\begin{aligned}
\mathbf{0} &= n^{-1/2}\sum_{i=1}^{n}\left\{\mathcal{K}_1(D_i,G_i,X_i,\widehat{\Omega}) - \frac{\widehat{R}_\Omega(X_i,\widehat{\Omega})}{\widehat{R}(X_i,\widehat{\Omega})}\right\}\\
&= n^{-1/2}\sum_{i=1}^{n}\left\{\mathcal{K}_1(D_i,G_i,X_i,\Omega) - \frac{\widehat{R}_\Omega(X_i,\Omega)}{\widehat{R}(X_i,\Omega)}\right\}\\
&\quad + \left[n^{-1}\textstyle\sum_{i=1}^{n}\partial\left\{\mathcal{K}_1(D_i,G_i,X_i,\Omega) - \dfrac{\widehat{R}_\Omega(X_i,\Omega)}{\widehat{R}(X_i,\Omega)}\right\}/\partial\Omega^{\mathrm{T}} + o_p(1)\right]\sqrt{n}(\widehat{\Omega}-\Omega)\\
&= n^{-1/2}\sum_{i=1}^{n}\left[\mathcal{K}_1(D_i,G_i,X_i,\Omega) - \mathcal{K}_2(X_i,\Omega) - \frac{\widehat{R}_\Omega(X_i,\Omega) - R_\Omega(X_i,\Omega)}{R(X_i,\Omega)}\right.\\
&\quad \left. + \frac{R_\Omega(X_i,\Omega)}{R^2(X_i,\Omega)}\left\{\widehat{R}(X_i,\Omega) - R(X_i,\Omega)\right\}\right] + (\Gamma_1-\Gamma_2)\sqrt{n}(\widehat{\Omega}-\Omega) + o_p(1)\\
&= n^{-1/2}\sum_{i=1}^{n}\left\{\mathcal{K}_1(D_i,G_i,X_i,\Omega) - \mathcal{K}_2(X_i,\Omega) - P_1(X_i,\Omega)\widehat{R}_\Omega(X_i,\Omega)\right.\\
&\quad \left. + P_2(X_i,\Omega)\widehat{R}(X_i,\Omega)\right\} + (\Gamma_1-\Gamma_2)\sqrt{n}(\widehat{\Omega}-\Omega) + o_p(1).
\end{aligned}
$$

Now using U-statistics properties,

$$
\begin{aligned}
&n^{-1/2}\textstyle\sum_{i=1}^{n}\left\{P_1(X_i,\Omega)\widehat{R}_\Omega(X_i,\Omega) - P_2(X_i,\Omega)\widehat{R}(X_i,\Omega)\right\}\\
&= \textstyle\sum_{r=0}^{1}\sum_{d=0}^{1} n^{-3/2}\sum_{i=1}^{n}\sum_{j=1}^{n}\dfrac{\pi_d}{c_d}I(D_j=d)\left\{P_1(X_i,\Omega)S_\Omega(r,G_j,X_i,\Omega)\right.\\
&\quad \left. - P_2(X_i,\Omega)S(r,G_j,X_i,\Omega)\right\}\\
&= \textstyle\sum_{r=0}^{1}\sum_{d=0}^{1} n^{-1/2}\sum_{i=1}^{n} E\left[\dfrac{\pi_d}{c_d}I(D=d)\left\{P_1(x_i,\Omega)S_\Omega(r,G,x_i,\Omega) - P_2(x_i,\Omega)S(r,G,x_i,\Omega)\right\}\right]\\
&\quad + \textstyle\sum_{r=0}^{1}\sum_{d=0}^{1} n^{-1/2}\sum_{j=1}^{n} E\left[\dfrac{\pi_d}{c_d}I(d_j=d)\left\{P_1(X,\Omega)S_\Omega(r,g_j,X,\Omega) - P_2(X,\Omega)S(r,g_j,X,\Omega)\right\}\right]\\
&\quad - \textstyle\sum_{r=0}^{1}\sum_{d=0}^{1} n^{-1/2}\sum_{j=1}^{n} E\left[\dfrac{\pi_d}{c_d}I(D_j=d)\left\{P_1(X,\Omega)S_\Omega(r,G_j,X,\Omega) - P_2(X,\Omega)S(r,G_j,X,\Omega)\right\}\right]\\
&\quad + o_p(1).
\end{aligned}
$$

Further, we thus have that

$$
\begin{aligned}
&n^{-1/2}\textstyle\sum_{i=1}^{n}\left\{P_1(X_i,\Omega)\widehat{R}_\Omega(X_i,\Omega) - P_2(X_i,\Omega)\widehat{R}(X_i,\Omega)\right\}\\
&= \textstyle\sum_{r=0}^{1}\sum_{d=0}^{1} n^{-1/2}\sum_{i=1}^{n}\pi_d\left\{P_1(X_i,\Omega)S_{E,\Omega}(r,d,X_i,\Omega) - P_2(X_i,\Omega)S_E(r,d,X_i,\Omega)\right\}\\
&\quad + \textstyle\sum_{t=0}^{1}\sum_{r=0}^{1}\sum_{d=0}^{1} n^{-1/2}\sum_{i=1}^{n}\dfrac{\pi_d c_t}{c_d}I(d_i=d)\\
&\quad \times E\left[\left\{P_1(X,\Omega)S_\Omega(r,g_i,X,\Omega) - P_2(X,\Omega)S(r,g_i,X,\Omega)\right\} \mid D=t\right]\\
&\quad - \textstyle\sum_{t=0}^{1}\sum_{r=0}^{1}\sum_{d=0}^{1} n^{1/2}\pi_d c_t E\left\{P_1(X,\Omega)S_{E,\Omega}(r,d,X,\Omega) - P_2(X,\Omega)S_E(r,d,X,\Omega) \mid D=t\right\}\\
&\quad + o_p(1).
\end{aligned}
$$

Thus,

$$
\begin{aligned}
& n^{-1/2}\textstyle\sum_{i=1}^{n} \left\{P_1(X_i,\Omega)\widehat{R}_\Omega(X_i,\Omega) - P_2(X_i,\Omega)\widehat{R}(X_i,\Omega)\right\} \\
= & \textstyle\sum_{r=0}^{1}\sum_{d=0}^{1} n^{-1/2}\sum_{i=1}^{n}\pi_d \left\{P_1(X_i,\Omega)S_{E,\Omega}(r,d,X_i,\Omega) - P_2(X_i,\Omega)S_E(r,d,X_i,\Omega)\right\} \\
& +\textstyle\sum_{d=0}^{1}\sum_{r=0}^{1} n^{-1/2}\sum_{i=1}^{n}\dfrac{\pi_{d_i}c_d}{c_{d_i}} E\left[\left\{P_1(X,\Omega)S_\Omega(r,g_i,X,\Omega) - P_2(X,\Omega)S(r,g_i,X,\Omega)\right\} \mid D=d\right] \\
& -\textstyle\sum_{t=0}^{1}\sum_{r=0}^{1}\sum_{d=0}^{1} n^{1/2}\pi_d c_t E\left\{P_1(X,\Omega)S_{E,\Omega}(r,d,X,\Omega) - P_2(X,\Omega)S_E(r,d,X,\Omega) \mid D=t\right\} \\
& +o_p(1) \\
= & \textstyle\sum_{d=0}^{1}\sum_{r=0}^{1} n^{-1/2}\sum_{i=1}^{n}\dfrac{\pi_{d_i}c_d}{c_{d_i}} E\left[\left\{P_1(X,\Omega)S_\Omega(r,g_i,X,\Omega) - P_2(X,\Omega)S(r,g_i,X,\Omega)\right\} \mid D=d\right] \\
& +o_p(1).
\end{aligned}
$$

Here the last step is because for any X,

$$
\begin{aligned}
& \textstyle\sum_{r=0}^{1}\sum_{d=0}^{1}\pi_d \left\{P_1(X,\Omega)S_{E,\Omega}(r,d,X,\Omega) - P_2(X,\Omega)S_E(r,d,X,\Omega)\right\} \\
= & \dfrac{1}{R(X,\Omega)}\textstyle\sum_{r=0}^{1}\sum_{d=0}^{1}\pi_d S_{E,\Omega}(r,d,X,\Omega) - \dfrac{R_\Omega(X,\Omega)}{R^2(X,\Omega)}\sum_{r=0}^{1}\sum_{d=0}^{1}\pi_d S_E(r,d,X,\Omega) \\
= & \dfrac{R_\Omega(X,\Omega)}{R(X,\Omega)} - \dfrac{R(X,\Omega)R_\Omega(X,\Omega)}{R^2(X,\Omega)} \\
= & \mathbf{0}.
\end{aligned}
$$

This leads to the result.

S·3. Score and Hessian: Rare Disease Case of §2.2 in the Main Paper

We consider models in which $\kappa + m(g,x,\boldsymbol{\beta}) = Q^{\mathrm{T}}(g,x)\Omega$, and $\Omega = (\kappa, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}}$. The point of this section is to show that both the log-pseudolikelihood score and its Hessian are very simply calculated, and that the Hessian is negative semidefinite.

In the rare disease case,

$$S(d,g,x,\Omega) = \exp\{dQ^{\mathrm{T}}(g,x)\Omega\}, \tag{16}$$

and thus

$$\log\{S(d,g,x,\Omega)\} = dQ^{\mathrm{T}}(g,x)\Omega.$$

This means that

$$\partial \log\{S(d,g,x,\Omega)\}/\partial\Omega = dQ(g,x),$$

and also that

$$\partial^2 \log\{S(d,g,x,\Omega)\}\partial\Omega\partial\Omega^{\mathrm{T}} = 0.$$

Similarly, in the rare disease case,

$$\widehat{R}(X,\Omega) = n_0^{-1}\textstyle\sum_{j=1}^{n}\sum_{r=0}^{1} I(D_j = 0)S(r, G_j, X, \Omega).$$

From (16),

$$\begin{aligned}\widehat{R}_\Omega(X,\Omega) = \partial\widehat{R}(X,\Omega)/\partial\Omega &= n_0^{-1}\textstyle\sum_{j=1}^{n}\sum_{r=0}^{1} I(D_j = 0)S(r, G_j, X, \Omega)rQ(G_j,X)\\ &= n_0^{-1}\textstyle\sum_{j=1}^{n} I(D_j = 0)S(1, G_j, X, \Omega)Q(G_j,X). \qquad (17)\end{aligned}$$

Thus,

$$\begin{aligned}\widehat{R}_{\Omega\Omega}(X,\Omega) &= \partial^2\widehat{R}(X,\Omega)/\partial\Omega\partial\Omega^{\mathrm{T}}\\ &= n_0^{-1}\textstyle\sum_{j=1}^{n} I(D_j = 0)S(1, G_j, X, \Omega)Q(G_j,X)Q^{\mathrm{T}}(G_j,X). \qquad (18)\end{aligned}$$

This means that the Hessian for the log-pseudolikelihood in equation (6) of the main paper is

$$\begin{aligned}-\frac{\partial\{\widehat{R}_\Omega(X,\Omega)/\widehat{R}(X,\Omega)\}}{\partial\Omega^{\mathrm{T}}} &= -\frac{\widehat{R}_{\Omega\Omega}(X,\Omega)}{\widehat{R}(X,\Omega)} + \frac{\widehat{R}_\Omega(X,\Omega)\widehat{R}_\Omega^{\mathrm{T}}(X,\Omega)}{\widehat{R}^2(X,\Omega)}\\ &= \{\widehat{R}(X,\Omega)\}^{-2}\left\{-\widehat{R}_{\Omega\Omega}(X,\Omega)\widehat{R}(X,\Omega) + \widehat{R}_\Omega(X,\Omega)\widehat{R}_\Omega^{\mathrm{T}}(X,\Omega)\right\}.\end{aligned}$$

Write $V_j = I(D_j = 0)S(1, G_j, X, \Omega)$. For matrices, define $A \leq B$ to be that $B - A$ is positive semidefinite. By Hölder's inequality

$$\begin{aligned}&\widehat{R}_\Omega(X,\Omega)\widehat{R}_\Omega^{\mathrm{T}}(X,\Omega)\\ &\quad= n_0^{-1}\textstyle\sum_{j=1}^{n} V_jQ(G_j,X) \times n_0^{-1}\sum_{j=1}^{n} V_jQ^{\mathrm{T}}(G_j,X)\\ &\quad\leq n_0^{-1}\textstyle\sum_{j=1}^{n} V_jQ(G_j,X)Q^{\mathrm{T}}(G_j,X) \times n_0^{-1}\sum_{j=1}^{n} V_j\\ &\quad= \widehat{R}_{\Omega\Omega}(X,\Omega)\widehat{R}(X,\Omega).\end{aligned}$$

Hence, the Hessian is negative semidefinite as claimed.

S·4. Stratification and the Independence Assumption

The assumption of gene-environment independence may not hold when there may exist underlying strata in the population, e.g. defined by ethnicity, across which the distribution of both genetic and environmental factors vary. In this case, as discussed in Section 3.1 of Chatterjee & Carroll (2005), we extend our framework to account for the scenario where the genetic and environmental factors can be assumed to be independent conditional on a discrete stratification $\mathcal{A}$ with $a = 1, ..., A$ levels.

To apply the method in Section 2.1 in the main paper to this case, for stratum a, we replace π_d by π_{da}, the probability that $D = d$ in the a^{th} stratum of the source population, and we replace n, n_0 and n_1 by n_a, n_{0a} and n_{1a}, the number of subjects, controls, and cases in stratum a, respectively. We modify (1) to $\text{pr}(D = 1|G, X, \mathcal{A} = a) = H\{\alpha_{0a} + m(G, X, \beta)\}$: more complex models with possible interactions between (G, X) and the strata can also be considered. We then set $\kappa_a = \alpha_{0a} + \log(n_{1a}/n_{0a}) - \log(\pi_{1a}/\pi_{0a})$. The parameters to be estimated are then $\Omega = (\kappa_1, ..., \kappa_A, \beta^{\mathrm{T}})^{\mathrm{T}}$. We also replace $S(d, g, x, \Omega)$ by

$$S_a(d, g, x, \Omega) = \frac{\exp[d\{\kappa_a + m(g, x, \beta)\}]}{1 + \exp\{\kappa_a + \log(\pi_{1a}/\pi_{0a}) - \log(n_{1a}/n_{0a}) + m(g, x, \beta)\}}.$$

Next, set $n = \sum_{a=1}^{A} n_a$, and replace (5) by

$$\widehat{R}_a(x, \Omega) = \textstyle\sum_{j=1}^{n}\sum_{r=0}^{1}\sum_{d=0}^{1}(\pi_{da}/n_{da})I(D_j = d, \mathcal{A}_j = a)S_a(r, G_j, x, \Omega),$$

and the estimated loglikelihood (6) becomes

$$\mathcal{L}(\Omega) = \textstyle\sum_{a=1}^{A}I(\mathcal{A}_i = a)[\sum_{i=1}^{n}\log\{S_a(D_i, G_i, X_i, \Omega)\} - \sum_{i=1}^{n}\log\{\widehat{R}_a(X_i, \Omega)\}],$$

which is then maximized to obtain the estimate $\widehat{\Omega}$. Now replace the score function (7) by

$$\widehat{\mathcal{S}}_n(\Omega) = n^{-1/2}\textstyle\sum_{a=1}^{A}\sum_{i=1}^{n}I(\mathcal{A}_i = a)\left\{\dfrac{S_{\Omega,a}(D_i, G_i, X_i, \Omega)}{S_a(D_i, G_i, X_i, \Omega)} - \dfrac{\widehat{R}_{\Omega,a}(X_i, \Omega)}{\widehat{R}_a(X_i, \Omega)}\right\},$$

using the obvious definitions of $S_{\Omega,a}(\cdot)$, $\widehat{R}_{\Omega,a}(\cdot)$, $P_{1a}(X, \Omega)$, $P_{2a}(X, \Omega)$ and with $Z_i = (D_i, G_i, X_i, \mathcal{A}_i)$.

In terms of the asymptotic theory of Section 2.3 of the main paper, we replace (Γ_1, Γ_2) by

$$\Gamma_1 = \textstyle\sum_{a=1}^{A}\sum_{d=0}^{1}(n_{da}/n)E\left\{\dfrac{\partial S_{\Omega,a}(D, G, X, \Omega)/S_a(D, G, X, \Omega)}{\partial\Omega^{\mathrm{T}}}\middle|\mathcal{A} = a, D = d\right\};$$

$$\Gamma_2 = \textstyle\sum_{a=1}^{A}\sum_{d=0}^{1}(n_{da}/n)E\left\{\dfrac{\partial R_{\Omega,a}(X, \Omega)/R_a(X, \Omega)}{\partial\Omega^{\mathrm{T}}}\middle|\mathcal{A} = a, D = d\right\}.$$

Then define

$$\begin{aligned}\zeta_a(Z_i, \Omega) = I(\mathcal{A}_i = a)&\frac{S_{\Omega,a}(Z_i, \Omega)}{S_a(Z_i, \Omega)} - \frac{R_{\Omega,a}(X_i, \Omega)}{R_a(X_i, \Omega)}\\ &-\sum_{d=0}^{1}\sum_{r=0}^{1}\frac{c_{d,a}\pi_{d_i,a}}{c_{d_i,a}}\\ &\qquad\times E\left[\{P_{1a}(X, \Omega)S_{\Omega,a}(r, g_i, X) - P_{2a}(X, \Omega)S_a(r, g_i, X)\} \mid \mathcal{A} = a, D = d\right],\end{aligned}$$

and now Σ becomes

$$\Sigma = \textstyle\sum_{a=1}^{A}\sum_{d=0}^{1}(n_{da}/n)\text{cov}\{\zeta_a(D, X, G, \Omega)|D = d, \mathcal{A} = a\}.$$

S·5. Additional Simulations

S·5·1. *Comparison with the Method of Chatterjee & Carroll (2005)*

Table 2 of this Supplementary Material gives results in the same simulation setting as in Section 3 in the main paper, except that to compare with Chatterjee & Carroll (2005), we only use the first SNP for our method and for the Chatterjee-Carroll method. The latter method uses the R package CGEN in Bioconductor, and is based on that package's function *snp.logistic*, which allows for SNP levels 0, 1, 2 and X values 0,1, as in our simulation. The results of that analysis and our method are very similar, indicating that our method is, in this case, almost efficient.

S·5·2. *Misspecification of Population Disease Rate*

Table 3 of this Supplementary Material reports the results of a simulation to evaluate the robustness of our method to misspecification of the population disease rate, using a sample of 1000 cases and 1000 controls. We considered actual disease rates of $\pi_1 = 0.03$, 0.05, 0.085 and 0.12, and compared the results for the rare disease approximation and when the assumed disease rate was $\pi_1 = 0.03$. For the method using the a rare disease approximation, it was only when the rate was $\pi_1 = 0.12$ that there was a deterioration in the coverage probabilities, but even then the lowest coverage rate was 91.8%. When the disease rate was assumed to be $\pi_1 = 0.03$, nominal coverage was seen except when the exact disease rate was $\pi_1 = 0.12$, and even at the worst case the lowest coverage rate was 93.1%, almost nominal. This indicates a surprising robustness to disease rate misspecification.

S·5·3. *Violations of the Gene-Environment Independence Assumption*

Tables S.4, S.5 and S.6 of this Supplementary Material contain simulations to examine the robustness of our method to violations of the gene-environment independence assumption. In these simulations, the genetic variables are generated as described in Section 3 of the main paper, but the environmental variable is normally distributed with mean αG_1, αG_2, or αG_3. We let $\alpha = 0.032$ to introduce a dependence between X and G with $R^2 = 0.001$. Here $\beta_G = \{\log(1.2), \log(1.2), 0, \log(1.2), 0\}$ as in Section 3 of the main paper, but $\beta_X = \log(1.73)$ and $\beta_{GX} = \{\log(1.42), 0, 0, \log(1.42), 0\}$. In each simulation, the logistic intercept was chosen to give a 3% population disease prevalence. In Table S.4 X is correlated with G_1, which has a nonzero main effect and a nonzero interaction; in Table S.5 X is correlated with G_2, which has a nonzero main effect but no interaction effect; in Table S.6 X is correlated with G_3, which has neither main nor interaction effects.

Similarly to Chatterjee & Carroll (2005), we find that violating the G-E independence assumption induces a bias in the parameter estimates. In Section S·4 of this Supplementary Material we describe how to remove this bias when G and E are independent conditional on a discrete stratification variable $\mathcal{A}$. Mukherjee & Chatterjee (2008) and Chen et al. (2009) show how to use empirical-Bayes methods as well to provide additional robustness against violations of the gene-environment independence assumption.

S·6. Properties of $\widehat{R}(x,\Omega)$ in equation (5) of the Main Paper

Equation (5) of the main paper is

$$\widehat{R}(x,\Omega)=\textstyle\sum_{j=1}^{n}\sum_{r=0}^{1}\sum_{d=0}^{1}(\pi_d/n_d)I(D_j=d)S(r,G_j,x,\Omega).$$

Computing its expectation is facilitated by seeing that

$$E\{I(D_j=d)S(r,G_j,x,\Omega)\}=E\{S(r,G_j,x,\Omega)|D_j=d\}=E\{S(r,G,x,\Omega)|D=d\}$$

Hence, recognizing that there are n_d subjects with $D=d$,

$$\begin{aligned}E\{\widehat{R}(x,\Omega)\}&=\textstyle\sum_{j=1}^{n}\sum_{r=0}^{1}\sum_{d=0}^{1}(\pi_d/n_d)E\{S(r,G,x,\Omega)|D=d\}\\&=\textstyle\sum_{r=0}^{1}\sum_{d=0}^{1}\pi_d/n_d)E\{S(r,G,x,\Omega)|D=d\}\\&=R(x,\Omega).\end{aligned}$$

Hence, (5) of the main paper is unbiased for $R(x,\Omega)$. Further, we see that

$$\begin{aligned}&\widehat{R}(x,\Omega)-R(x,\Omega)\\&=\textstyle\sum_{r=0}^{1}\sum_{d=0}^{1}(\pi_d/n_d)\sum_{j=1}^{n}\\&\qquad\qquad\times\left[I(D_j=d)S(r,G_j,x,\Omega)-E\{I(D_j=d)S(r,G_j,x,\Omega)\}\right],\end{aligned}$$

so that $\widehat{R}(x,\Omega)$ is $n^{1/2}$-consistent for $R(x,\Omega)$, and with proper normalization is asymptotically normally distributed.

S·7. SNPs Involved in Creating the Polygenic Risk Score

Table S.1. *SNPs involved in creating the polygenic risk score, and their regression coefficients*

Actual RS	Variable	
Number	Name	Coefficient
rs11249433	gene1	-0.02813492
rs1045485	gene2	-0.09307971
rs13387042	gene3	-0.26203658
rs4973768	gene4	0.08013260
rs10069690	gene5	0.06459363
rs10941679	gene6	0.09185539
rs889312	gene7	-0.00565121
rs17530068	gene8	0.09668742
rs2046210	gene9	0.09851217
rs1562430	gene10	-0.14871719
rs1011970	gene11	0.05329783
rs865686	gene12	-0.02913340
rs2380205	gene13	-0.01821032
rs10995190	gene14	-0.04275836
rs2981582	gene15	0.14008397
rs909116	gene16	0.04955235
rs614367	gene17	0.06438418
rs3803662	gene18	0.27080105
rs6504950	gene19	-0.17586244
rs8170	gene20	0.08570773
rs999737_as	gene21	-0.13737833

S·8. Comparison with the Method of Chatterjee & Carroll (2005) in a Special Case

Table S.2. *Results of 1000 simulations with 3% disease prevalence as described in Section 3 of the main paper, except that to compare with Chatterjee & Carroll (2005), we only use the first SNP. We compare our semiparametric pseudolikelihood estimator to the method of Chatterjee & Carroll (2005) and to ordinary logistic regression. The simulations were performed with 500 cases and 500 controls*

	500 cases & 500 controls			1000 cases & 1000 controls		
	β_{G1}	β_X	β_{G1X}	β_{G1}	β_X	β_{G1X}
True	0.182	0.405	0.262	0.182	0.405	0.262
			Logistic			
Bias	-0.011	0.001	0.015	0.009	0.003	-0.001
CI (%)	93.9	94.1	93.7	95.2	94.2	95.6
			Chatterjee Carroll			
Bias	-0.008	0.005	-0.004	0.013	0.006	-0.016
CI (%)	95.1	94.1	93.6	96.0	94.6	94.4
MSE Eff	1.405	1.108	2.227	1.321	1.118	2.183
			SPMLE, Rare			
Bias	-0.007	0.004	-0.001	0.013	0.006	-0.015
CI (%)	95.1	94.1	94.1	95.8	94.5	94.8
MSE Eff	1.381	1.104	2.166	1.290	1.113	2.141
			SPMLE, π_1 known			
Bias	-0.014	0.001	0.014	0.006	0.003	0.000
CI (%)	95.1	94.2	94.8	95.9	94.7	94.4
MSE Eff	1.359	1.100	2.016	1.292	1.113	2.021

Logistic is ordinary logistic regression; *Chatterjee Carroll* is the method of Chatterjee & Carroll (2005); *SPMLE, Rare* is our estimator using the rare disease approximation with unknown π_1 (Section 2.2 of the main paper); *SPMLE, π_1 known* is our estimator when π_1 is known in the source population (Section 2.1 of the main paper); *Bias* is the mean bias; *CI (%)* is the coverage in percent of a nominal 95% confidence interval (calculated using the asymptotic standard error); *MSE Eff* is the mean squared error efficiency of the method compared to logistic regression.

S·9. SIMULATION WHEN THE DISEASE RATE IS MISSPECIFIED

Table S.3. *Results of 1000 simulations as described in §3 of the main paper, except that the logistic intercept has been modified to give population disease rates* $(0.03, 0.05, 0.085, 0.12)$. *We compare ordinary logistic regression, our method using the rare disease approximation, and our method with "known"* $\pi_1 = 0.03$, *which is misspecified when* $\pi_1 > 0.03$. *The simulations were performed with 1000 cases and 1000 controls*

	β_{G1}	β_{G2}	β_{G3}	β_{G4}	β_{G5}	β_X	β_{G1X}	β_{G2X}	β_{G3X}	β_{G4X}	β_{G5X}
True	0.18	0.18	0.00	0.18	0.00	0.41	0.26	0.00	0.00	0.26	0.00
						Logistic					
Disease Rate $= 0.03$											
Bias	0.00	0.01	0.00	0.01	-0.01	0.01	0.01	-0.01	0.00	0.00	0.01
CI (%)	94.3	95.2	95.7	95.1	94.7	94.6	94.9	94.2	94.5	96.0	94.2
Disease Rate $= 0.05$											
Bias	0.00	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	0.00	0.00	0.01
CI (%)	95.8	95.2	95.9	94.7	94.4	95.6	95.7	95.5	95.3	94.8	95.3
Disease Rate $= 0.085$											
Bias	-0.01	0.01	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.01
CI (%)	94.2	94.8	95.6	94.4	93.7	94.4	94.9	94.3	94.9	95.9	94.2
Disease Rate $= 0.12$											
Bias	0.00	0.01	-0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
CI (%)	94.8	95.5	94.9	95.2	93.8	95.7	94.4	95.9	94.9	95.3	95.0
						SPMLE, Rare					
Disease Rate $= 0.03$											
Bias	0.01	0.00	0.00	0.02	-0.01	0.02	-0.02	-0.01	0.01	-0.02	0.01
CI (%)	95.2	95.4	96.4	95.8	95.3	95.1	95.4	94.8	96.1	95.5	94.9
MSE Eff			All G: 1.28			X: 1.26			All $G*X$: 2.18		
Disease Rate $= 0.05$											
Bias	0.02	0.00	0.00	0.02	-0.01	0.03	-0.04	0.00	0.00	-0.03	0.00
CI (%)	94.4	95.4	96.8	94.4	95.0	95.1	93.8	94.6	96.3	94.5	94.4
MSE Eff			All G: 1.25			X: 1.23			All $G*X$: 1.99		
Disease Rate $= 0.085$											
Bias	0.02	0.01	0.00	0.02	0.00	0.05	-0.05	-0.01	0.00	-0.05	0.00
CI (%)	95.0	94.5	96.1	94.1	93.9	93.5	93.9	94.8	95.8	94.5	95.6
MSE Eff			All G: 1.25			X: 1.14			All $G*X$: 2.02		
Disease Rate $= 0.12$											
Bias	0.03	0.01	-0.01	0.03	0.00	0.06	-0.08	-0.01	0.00	-0.06	0.00
CI (%)	94.2	95.5	94.6	93.3	93.9	93.4	92.0	96.1	94.5	91.8	94.4
MSE Eff			All G: 1.21			X: 1.02			All $G*X$: 1.88		
						SPMLE, $\pi_1 = 0.03$					
Disease Rate $= 0.03$											
Bias	0.00	0.00	0.00	0.01	-0.01	0.01	0.00	-0.01	0.01	-0.01	0.01
CI (%)	95.1	95.5	96.4	95.8	95.0	95.5	95.6	94.6	95.9	95.2	94.5
MSE Eff			All G: 1.28			X: 1.28			All $G*X$: 2.07		
Disease Rate $= 0.05$											
Bias	0.01	0.00	0.00	0.01	-0.01	0.01	-0.01	0.00	0.00	-0.01	0.01
CI (%)	94.6	95.4	96.4	94.7	94.7	95.8	94.3	94.6	96.0	94.5	94.1
MSE Eff			All G: 1.25			X: 1.27			All $G*X$: 1.90		
Disease Rate $= 0.085$											
Bias	0.01	0.01	0.00	0.01	0.00	0.03	-0.03	-0.01	0.00	-0.03	0.00
CI (%)	95.1	94.8	96.4	94.4	93.9	94.7	94.9	95.1	95.8	94.9	95.2
MSE Eff			All G: 1.25			X: 1.21			All $G*X$: 1.95		
Disease Rate $= 0.12$											
Bias	0.02	0.01	-0.01	0.03	0.00	0.05	-0.06	-0.01	0.00	-0.05	0.01
CI (%)	94.3	95.6	94.9	93.6	93.8	94.4	93.5	96.3	94.6	93.1	94.6
MSE Eff			All G: 1.22			X: 1.10			All $G*X$: 1.84		

Logistic is ordinary logistic regression; *SPMLE, Rare* is our estimator using the rare disease approximation with unknown π_1 (Section 2.2 of the main paper); *SPMLE,* $\pi_1 = 0.03$ is our estimator calculated as if the disease rate in the source population were known to be 0.03 (Section 2.1 of the main paper); *Bias* is the mean bias; *CI (%)* is the coverage in percent of a nominal 95% confidence interval (calculated using the asymptotic standard error); *MSE Eff* is the mean squared error efficiency of our method compared to logistic regression, averaged over G, over X and over the $G * X$ interactions.

S·10. Simulations When the Gene-Environment Independence Assumption is Violated

Table S.4. *Results of 1000 simulations with G as described in Section 3 of the main paper, but $X \sim \mathbf{N}(0.032G_1, 1)$. We compare our semiparametric pseudolikelihood estimator to ordinary logistic regression. Three simulations were performed with sample sizes of (1000, 2000, 3000) cases and controls each*

	β_{G1}	β_{G2}	β_{G3}	β_{G4}	β_{G5}	β_X	β_{G1X}	β_{G2X}	β_{G3X}	β_{G4X}	β_{G5X}
True	0.18	0.18	0.00	0.18	0.00	0.55	0.35	0.00	0.00	0.35	0.00
Logistic: 1000 cases											
Bias	-0.01	0.00	0.01	0.00	-0.01	0.01	0.01	0.01	-0.01	0.01	0.01
CI (%)	94.5	96.2	95.8	94.8	93.7	94.0	95.4	95.7	95.6	95.5	95.3
Logistic: 2000 cases											
Bias	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CI (%)	95.4	94.7	94.8	95.2	95.0	94.5	95.6	96.1	94.0	94.7	95.9
Logistic: 3000 cases											
Bias	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CI (%)	94.1	94.1	95.1	95.7	94.6	94.2	94.2	95.4	94.8	95.0	94.7
SPMLE, π_1 known: 1000 cases											
Bias	-0.01	0.00	0.01	0.00	-0.01	-0.03	0.10	0.00	0.00	0.01	0.00
CI (%)	94.2	95.9	95.0	95.2	93.8	93.3	80.4	94.9	94.9	95.0	94.8
MSE Eff			All G: 1.07			X: 1.31			All $G * X$: 1.75		
SPMLE, π_1 known: 2000 cases											
Bias	0.00	0.00	0.00	0.01	0.00	-0.03	0.10	0.00	0.00	0.00	0.00
CI (%)	94.2	94.8	95.1	95.5	95.6	90.9	71.4	95.5	94.1	95.0	95.6
MSE Eff			All G: 1.07			X: 1.08			All $G * X$: 1.53		
SPMLE, π_1 known: 3000 cases											
Bias	-0.01	0.00	0.00	0.00	0.00	-0.03	0.10	0.00	0.00	0.00	0.00
CI (%)	94.7	95.3	95.7	95.2	94.2	88.0	54.8	94.2	95.7	95.0	93.9
MSE Eff			All G: 1.06			X: 0.95			All $G * X$: 1.27		

Logistic is ordinary logistic regression; *SPMLE, π_1 known* is our estimator when π_1 is known in the source population (Section 2.1 of the main paper); *Bias* is the mean bias; *CI (%)* is the coverage in percent of a nominal 95% confidence interval (calculated using the asymptotic standard error); *MSE Eff* is the mean squared error efficiency of our method compared to logistic regression, averaged over G, over X and over the $G * X$ interactions.

Table S.5. *Results of 1000 simulations with G as described in Section 3 of the main paper, but $X \sim \mathbf{N}(0.032G_2, 1)$. We compare our semiparametric pseudolikelihood estimator to ordinary logistic regression. Three simulations were performed with sample sizes of (1000, 2000, 3000) cases and controls each*

	β_{G1}	β_{G2}	β_{G3}	β_{G4}	β_{G5}	β_X	β_{G1X}	β_{G2X}	β_{G3X}	β_{G4X}	β_{G5X}
True	0.18	0.18	0.00	0.18	0.00	0.55	0.35	0.00	0.00	0.35	0.00
Logistic: 1000 cases											
Bias	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.00
CI (%)	93.4	95.1	94.5	93.0	95.7	94.4	94.4	93.7	94.8	93.4	94.4
Logistic: 2000 cases											
Bias	0.00	-0.01	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.00
CI (%)	95.3	94.0	94.4	94.6	93.2	94.9	94.6	94.8	94.2	95.5	93.8
Logistic: 3000 cases											
Bias	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CI (%)	94.1	94.5	94.5	95.3	95.2	95.9	94.7	93.9	94.4	95.6	95.3
SPMLE, π_1 known: 1000 cases											
Bias	0.00	-0.01	0.00	0.01	-0.01	-0.04	0.01	0.06	0.00	0.00	0.00
CI (%)	93.7	95.3	95.4	94.0	95.1	89.4	93.8	86.0	95.0	94.6	94.9
MSE Eff			All G: 1.06			X: 1.12			All $G * X$: 2.19		
SPMLE, π_1 known: 2000 cases											
Bias	0.00	-0.01	0.00	0.00	0.00	-0.04	0.01	0.06	0.00	0.00	0.00
CI (%)	95.6	94.2	94.9	94.4	93.9	88.1	94.3	78.7	95.1	95.4	95.6
MSE Eff			All G: 1.08			X: 0.91			All $G * X$: 1.91		
SPMLE, π_1 known: 3000 cases											
Bias	0.00	0.00	0.00	0.00	0.00	-0.04	0.01	0.06	0.00	0.00	0.00
CI (%)	94.8	94.2	94.9	95.9	94.9	84.3	95.4	72.7	95.4	95.3	95.5
MSE Eff			All G: 1.08			X: 0.72			All $G * X$: 1.82		

Logistic is ordinary logistic regression; *SPMLE, π_1 known* is our estimator when π_1 is known in the source population (Section 2.1 of the main paper); *Bias* is the mean bias; *CI (%)* is the coverage in percent of a nominal 95% confidence interval (calculated using the asymptotic standard error); *MSE Eff* is the mean squared error efficiency of our method compared to logistic regression, averaged over G, over X and over the $G * X$ interactions.

Table S.6. *Results of 1000 simulations with G as described in Section 3 of the main paper, but $X \sim \mathbf{N}(0.032G_3, 1)$. We compare our semiparametric pseudolikelihood estimator to ordinary logistic regression. Three simulations were performed with sample sizes of (1000, 2000, 3000) cases and controls each*

	β_{G1}	β_{G2}	β_{G3}	β_{G4}	β_{G5}	β_X	β_{G1X}	β_{G2X}	β_{G3X}	β_{G4X}	β_{G5X}
True	0.18	0.18	0.00	0.18	0.00	0.55	0.35	0.00	0.00	0.35	0.00
						Logistic: 1000 cases					
Bias	-0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.01
CI (%)	95.5	94.4	95.2	96.2	95.3	94.7	94.9	94.0	94.9	95.5	94.9
						Logistic: 2000 cases					
Bias	0.00	0.00	-0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00
CI (%)	94.0	94.1	94.4	94.6	94.9	95.2	95.5	95.1	95.5	94.0	94.7
						Logistic: 3000 cases					
Bias	0.00	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	0.00	0.00	0.00
CI (%)	95.9	94.2	94.1	94.8	94.3	95.1	95.4	95.9	95.8	92.9	94.4
						SPMLE, π_1 known: 1000 cases					
Bias	0.00	0.00	0.00	0.00	0.00	-0.04	0.01	0.00	0.06	0.01	0.00
CI (%)	95.6	94.8	95.5	96.3	95.3	92.0	94.8	95.5	88.3	95.7	96.2
MSE Eff			All G: 1.07			X: 1.20			All $G * X$: 2.12		
						SPMLE, π_1 known: 2000 cases					
Bias	0.00	0.00	-0.01	0.00	0.00	-0.04	0.01	0.00	0.06	0.00	0.00
CI (%)	95.2	94.4	94.5	94.0	94.8	89.4	95.0	94.8	82.3	94.9	94.6
MSE Eff			All G: 1.06			X: 0.95			All $G * X$: 1.95		
						SPMLE, π_1 known: 3000 cases					
Bias	0.00	0.00	0.00	0.00	-0.01	-0.04	0.00	0.00	0.06	0.00	0.00
CI (%)	95.3	94.7	94.0	95.3	94.2	84.5	94.4	94.9	75.7	95.0	94.8
MSE Eff			All G: 1.06			X: 0.76			All $G * X$: 1.82		

Logistic is ordinary logistic regression; *SPMLE, π_1 known* is our estimator when π_1 is known in the source population (Section 2.1 of the main paper); *Bias* is the mean bias; *CI (%)* is the coverage in percent of a nominal 95% confidence interval (calculated using the asymptotic standard error); *MSE Eff* is the mean squared error efficiency of our method compared to logistic regression, averaged over G, over X and over the $G * X$ interactions.

S·11. The Simulation in Table 1 of the Main Paper With Componentwise Mean Squared Error Efficiencies

Table S.7. *Results of 1000 simulations as described in Section 3 of the main paper, with mean bias, coverage probabilities of a 95% nominal confidence interval, and mean squared error efficiency of our semiparametric pseudolikelihood estimator compared to ordinary logistic regression. The sample sizes were performed with 500 cases and 500 controls, and again with 1000 cases and 1000 controls*

	β_{G1}	β_{G2}	β_{G3}	β_{G4}	β_{G5}	β_X	β_{G1X}	β_{G2X}	β_{G3X}	β_{G4X}	β_{G5X}
True	0.18	0.18	0.00	0.18	0.00	0.41	0.26	0.00	0.00	0.26	0.00
						Logistic, 500 cases					
Bias	0.02	-0.02	0.02	-0.01	0.01	0.00	0.00	0.02	-0.02	0.02	-0.01
CI (%)	94.7	94.9	94.8	94.5	95.2	96.4	94.3	93.6	94.3	94.9	95.4
						Logistic, 1000 cases					
Bias	0.00	0.01	0.00	0.01	-0.01	0.01	0.01	-0.01	0.00	0.00	0.01
CI (%)	94.3	95.2	95.7	95.1	94.7	94.6	94.9	94.2	94.5	96.0	94.2
						SPMLE, Rare, 500 cases					
Bias	0.02	-0.01	0.02	0.00	0.00	0.01	-0.01	0.01	-0.02	-0.01	0.00
CI (%)	95.0	95.8	94.2	94.5	95.5	95.6	95.8	95.3	94.3	95.0	95.9
MSE Eff	1.37	1.34	1.23	1.27	1.27	1.29	2.44	2.13	1.87	1.91	2.22
						SPMLE, Rare, 1000 cases					
Bias	0.01	0.00	0.00	0.02	-0.01	0.02	-0.02	-0.01	0.01	-0.02	0.01
CI (%)	95.2	95.4	96.4	95.8	95.3	95.1	95.4	94.8	96.1	95.5	94.9
MSE Eff	1.35	1.25	1.29	1.25	1.24	1.26	2.36	2.00	2.19	2.02	2.21
						SPMLE, π_1 known: 500 cases					
Bias	0.01	-0.01	0.02	-0.01	0.00	0.00	0.01	0.01	-0.02	0.01	0.00
CI (%)	95.0	95.7	94.3	94.4	95.5	95.7	95.4	95.1	94.3	94.9	95.7
MSE Eff	1.39	1.34	1.22	1.26	1.28	1.28	2.31	2.01	1.78	1.81	2.09
						SPMLE, π_1 known: 1000 cases					
Bias	0.00	0.00	0.00	0.01	-0.01	0.01	0.00	-0.01	0.01	-0.01	0.01
CI (%)	95.1	95.5	96.4	95.8	95.0	95.5	95.6	94.6	95.9	95.2	94.5
MSE Eff	1.36	1.25	1.28	1.27	1.24	1.28	2.25	1.91	2.06	1.96	2.08

Logistic is ordinary logistic regression; *SPMLE, Rare* is our estimator using the rare disease approximation with unknown π_1, Section 2.2; *SPMLE, π_1 known* is our estimator when π_1 is known in the source population, Section 2.1; *CI (%)* is the coverage in percent of a nominal 95% confidence interval, calculated using the asymptotic standard error; *MSE Eff* is the mean squared error efficiency of our method compared to logistic regression.

S·12. Skewness, Kurtosis and qq-Plots for the Simulation in Table 1 of the Main Paper

Table S.8 gives skewness and kurtosis for the simulation in Table 1 of the main paper with 1000 cases and controls.

Figure S.1 presents q–q plots for the main effects for $(G_1, \ldots, G_5, X)$ in the same simulation.

Figure S.2 presents q–q plots for the interaction effects for X and $(G_1, \ldots, G_5)$ in the same simulation.

Table S.8. *Skewness and kurtosis for the simulation in Table 1 of the main paper with* 1000 *cases and controls. Kurtosis = 0 for the normal distribution*

Skewness	Kurtosis
-0.02	-0.08
-0.05	0.12
-0.13	0.07
-0.02	-0.15
0.06	-0.04
-0.21	0.15
-0.01	-0.20
-0.03	-0.10
0.04	0.11
0.09	-0.13
0.01	-0.06
0.14	0.25

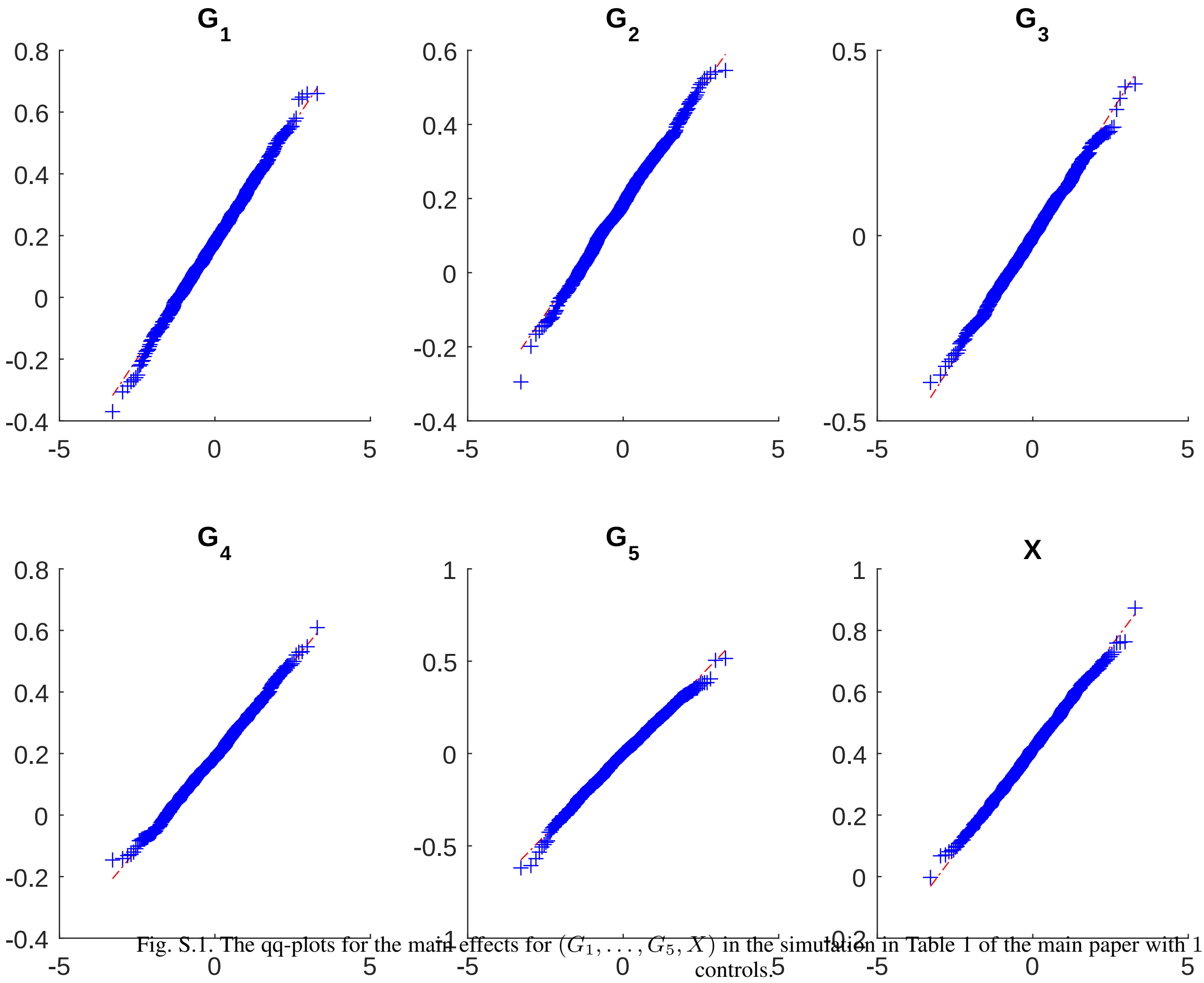

Fig. S.1. The qq-plots for the main effects for $(G_1, \ldots, G_5, X)$ in the simulation in Table 1 of the main paper with 1000 cases and controls.

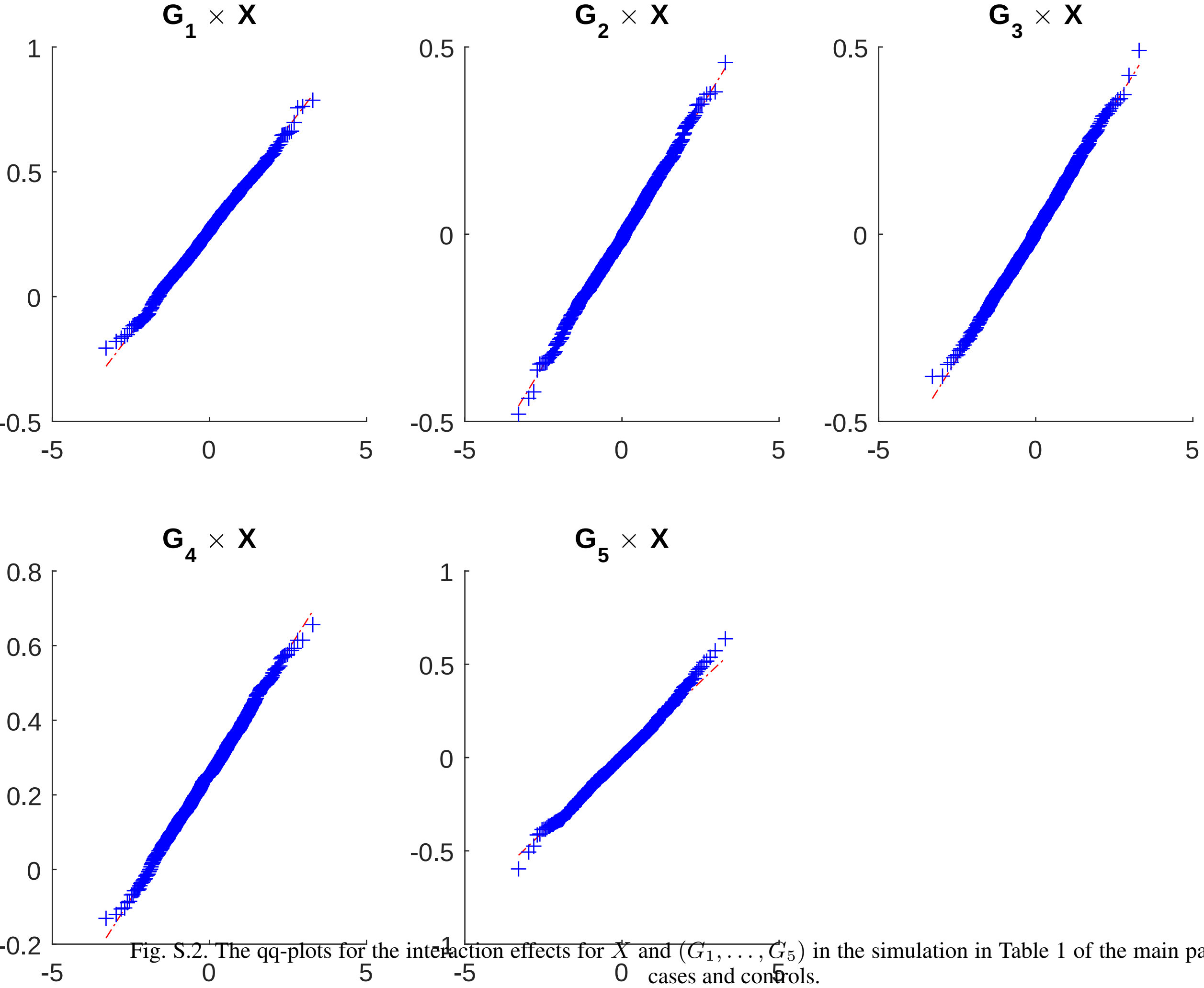

Fig. S.2. The qq-plots for the interaction effects for X and $(G_1, \dots, G_5)$ in the simulation in Table 1 of the main paper with 1000 cases and controls.

S·13. The Simulation in Table 1 of the Main Paper With 500 Cases and Controls

Table S.9. *Results of 1000 simulations as described in §3 of the main paper, with mean bias, coverage probabilities of a 95% nominal confidence interval, and mean squared error efficiency of our semiparametric pseudolikelihood estimator compared to ordinary logistic regression. The simulations were performed with 500 cases and 500 controls*

	β_{G1}	β_{G2}	β_{G3}	β_{G4}	β_{G5}	β_X	β_{G1X}	β_{G2X}	β_{G3X}	β_{G4X}	β_{G5X}
True	0.18	0.18	0.00	0.18	0.00	0.41	0.26	0.00	0.00	0.26	0.00
Logistic: 500 cases											
Bias	0.02	-0.02	0.02	-0.01	0.01	0.00	0.00	0.02	-0.02	0.02	-0.01
CI (%)	94.7	94.9	94.8	94.5	95.2	96.4	94.3	93.6	94.3	94.9	95.4
SPMLE, Rare: 500 cases											
Bias	0.02	-0.01	0.02	0.00	0.00	0.01	-0.01	0.01	-0.02	-0.01	0.00
CI (%)	95.0	95.8	94.2	94.5	95.5	95.6	95.8	95.3	94.3	95.0	95.9
Avg MSE Eff			All G: 1.30			X: 1.29			All $G * X$: 2.13		
SPMLE, π_1 known: 500 cases											
Bias	0.01	-0.01	0.02	-0.01	0.00	0.00	0.01	0.01	-0.02	0.01	0.00
CI (%)	95.0	95.7	94.3	94.4	95.5	95.7	95.4	95.1	94.3	94.9	95.7
Avg MSE Eff			All G: 1.30			X: 1.28			All $G * X$: 2.02		

Logistic is ordinary logistic regression; *SPMLE, Rare* is our estimator using the rare disease approximation with unknown π_1 (§2.2); *SPMLE, π_1 known* is our estimator when π_1 is known in the source population (§2.1); *CI (%)* is the coverage in percent of a nominal 95% confidence interval (calculated using the asymptotic standard error); *Avg MSE Eff* is the mean squared error efficiency of our method compared to logistic regression averaged over G, over X and over the $G * X$ interactions, respectively.

References

Chatterjee, N. & Carroll, R. J. (2005). Semiparametric maximum likelihood estimation in case-control studies of gene-environment interactions. *Biometrika* **92**, 399–418.

Chen, Y. H., Chatterjee, N. & Carroll, R. J. (2009). Shrinkage estimators for robust and efficient inference in haplotype-based case-control studies. *J. Am. Statist. Assoc.* **104**, 220–233.

Ma, Y. (2010). A semiparametric efficient estimator in case-control studies. *Bernoulli* **16**, 585–603.

Mukherjee, B. & Chatterjee, N. (2008). Exploiting gene-environment independence for analysis of case–control studies: An empirical bayes-type shrinkage estimator to trade-off between bias and efficiency. *Biometrics* **64**, 685–694.

Van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.

Wei, J., Carroll, R. J., Muller, U., Van Keilegom, I. & Chatterjee, N. (2013). Locally efficient estimation for homoscedastic regression in the secondary analysis of case-control data. *J. R. Statist. Soc.* B **75**, 185–206.