Using shared genetic controls in studies of gene-environment interactions

Raymond J. Carroll
Department of Statistics
Center for Statistical Bioinformatics
Institute for Applied Mathematics and Computational Science

Texas A&M University
http://stat.tamu.edu/~carroll
Acknowledgment

- This work is joint with Nilanjan Chatterjee (NCI) and Yi-Hau Chen (Academia Sinica)
Basic Problem Formalized

- **Gene** and **Environment**

- **Question**: For women who carry the **BRCA1/2 mutation**, does **oral contraceptive use** provide any protection against ovarian cancer?
Basic Problem Formalized

• Gene and Environment

• Question: For people carrying a particular haplotype in the VDR pathway, does higher levels of serum Vitamin D protect against prostate cancer?
Basic Problem Formalized

• **Gene** and **Environment**

• **Question**: If you are a **current smoker**, are you protected against bladder cancer if you carry a particular **SNP**?
Prospective and Retrospective Studies

- \( D = \) disease status (binary)
- \( X = \) environmental variables
- \( G = \) gene status
- \( S = \) population stratification variables (later)
Prospective and Retrospective Studies

- **Prospective**: Classic random sampling of a population
  - You measure gene and environment on a cohort
  - You then follow up people for disease occurrence
Prospective and Retrospective Studies

- **Prospective Studies:**
  - *Expensive:* disease states are rare, so large sample sizes needed
  - *Time-consuming:* you have to wait for disease to develop
  - *They Exist:* Framingham Heart Study, NIH-AARP Diet and Health Study, Women’s Health Initiative, etc.
Prospective and Retrospective Studies

- **Prospective Studies:**

- **Daunting Task:** Only very large, very expensive prospective studies can find gene-environment interactions

- **Data Access:** Access to the Framingham Heart Study requires a university commitment to security
Prospective and Retrospective Studies

- **Retrospective Studies**: Usually called case-control studies

- Find a population of **cases**, i.e., people with a disease, and sample from it.

- Find a population of **controls**, i.e., people without the disease, and sample from it.
Prospective and Retrospective Studies

- **Retrospective Studies**: Because the gene $G$ and the environment $X$ are sampled after disease status is ascertained.

- **Microarray studies on humans**: most are case-control studies.

- **Genome Wide Association Studies (GWAS)**: most are case-control studies.
Prospective and Retrospective Studies

Case-control Studies:

- **Fast**: no need to wait for disease to develop
- **Cheapish**: sample sizes are much smaller
- **Subtle**: The controls need to be representative of the population of people without the disease.
The Need for Shared Controls

- **Case-control Studies**: Many case-control studies in genetic epidemiology are becoming enormous.

- For example, a 2011 study of bladder cancer pooled a number of studies and had 8,381 cases and 48,275 controls.

- There are a number of such studies out there.
The Need for Shared Controls

• **Gene-environment:**

  • In a standard gene-environment study, you have to get the environmental variables, $X$, as well as the genetic variables, $G$.

  • As we will see, for rare diseases any interaction between gene and environment can be found by using cases only, **under assumptions**
The Need for Shared Controls

- **Gene-environment:**

- So, it is tempting for the same budget to **sample cases only** (and hence more of them)
The Need for Shared Controls

• **Gene-environment:**

• In modern genomics studies, you may have controls, but to save money you may wish not to genotype the controls
We want to explore how one might reasonably do use such a sampling scheme

Alarm bells should go off

Assumptions must be made (not mistakes, mind you)
Logistic Regression

• **Logistic Function:**

\[ H(x) = \frac{\exp(x)}{1 + \exp(x)} \quad \text{and} \quad 1 - H(x) = \frac{1}{1 + \exp(x)} \]

• For a rare disease, \( 1 - H(x) \approx 1 \) hence

\[ H(x) \approx \exp(x) \]
Logistic Model

- As is usual, we will model disease by a logistic regression model, with an interaction.

\[
\text{pr}(D = 1| \ G, X) = H(\beta_0 + \beta_g G + \beta_x X + \beta_{xg} G \ast X)
\]

- All the parameters are of interest. For a rare disease,

\[
\text{pr}(D = 1| \ G, X) \approx \exp(\beta_0 + \beta_g G + \beta_x X + \beta_{xg} G \ast X)
\]
Gene-Environment Independence

- In many cases, it is reasonable to posit that your genetic status and your environmental exposure are *independent in the population*, after perhaps accounting for strata.

- A useful assumption with lots of literature and attempts at robustness to the assumption.
The Case Only Design

• In 1994, Piegorsch, et al discovered that in rare diseases, with gene-environment independence, you can estimate the interaction using cases only.

• Their calculation is retrospective:

\[
\frac{\text{pr}(G=1,X=1|D=1) \text{pr}(G=0,X=0|D=1)}{\text{pr}(G=0,X=1|D=1) \text{pr}(G=1,X=0|D=1)} \approx \exp(\beta_{xg})
\]
The Case Only Design

• The case-only design means that for the same budget, you can make your number of cases two times larger

• This means enormous gains in power.

• It means many more discoveries at early stages that can be followed up later
The Case Only Design

- The case-only design however can only estimate the interaction

- It cannot estimate the genetic main effect

- It cannot estimate the environmental main effect

- *If we have gene status on controls*, we discovered a while ago how to get power gains and main effects
External Controls

• If we have no controls, what if we grab them from another study?

• The buzz word is **external controls**

• Sadly, nothing doing. The main effects are not identified unless the external controls have the same \((G,X)\) distribution as internal controls.

• And pigs fly
External Controls

• Blindly using external controls is obviously a bad idea.

• The distribution of (G,X) in your study surely differs from that in an external study.

• All sorts of biases then arise.
External and Internal Controls

• We need to link internal and external controls.

• One possibility: use external controls, but get X for internal controls, and do not go to the expense of getting G

• Prices keep falling, but G is not free
Stratification Variables

• Population stratification is a major issue in case-control studies, i.e., the cases and controls might not come from the same underlying study base.

• This is not always possible, so that differences of \((G,X)\) between cases and controls might be due to confounding.
Stratification Variables

• In modern studies, such as SNP arrays, a wealth of genetic information is gathered.

• It is common to take a large number of ancestry informative SNPs in the cases and controls.

• Then often PCA is done to define a few ancestry informative variables, $S$, and these are included in the risk model.
Stratification Variable Model

• This makes the risk model

\[ \text{pr}(D = 1| G, X) = H\{\beta_0 + m(G, X, \beta) + r(S, \delta)\} \]

• For example,

\[ m(G, X, \beta) = \beta_g G + \beta_x X + \beta_{xg} G \ast X \]
Data Design

- In our main study, we observe \((D=1,G,X,S)\) on all cases

- We observe on \((D=0,X)\) among our controls

- In the external study, we observe \((D=0,G,X,S)\) among their controls
Assumptions

- We assume **rare disease**

- **G and X are independent given S.** Thus, gene-environment independence within each sub-population (but not in the general population). Combined with the rare disease assumption

\[
\Pr(G|X,S) \approx \Pr(G|X,S,D=0) \\
\approx \Pr(G|S,D=0)
\]
Assumptions

- **The stratification works**: the distribution of $G$ given $S$ is the same in the main and the external study.

- Using the rare disease assumption, we model this distribution as

$$\text{pr}(G = g | S) \approx \text{pr}(G = g | S, D = 0) = Q(S, \theta)$$
A Simple Conditional Likelihood

- The simplest method uses the internal cases and the external controls only

- It throws away the X data among the internal controls

- Thus, it carefully makes no assumptions about the distribution of X given S
A Simple Conditional Likelihood

- We likelihood consists of two products. The first is that of the distribution of G among the internal cases

\[ \prod_{i=1}^{n_1} \text{pr}_{\text{Int}}(G_i \mid X_i, S_i, D_i = 1, \text{Internal}) \]

- The second is the distribution of G among the external controls (we toss away their X)

\[ \prod_{i=1}^{n_0} \text{pr}_{\text{Ext}}(G_i \mid S_i, D_i = 0, \text{External}) \]
A Simple Conditional Likelihood

- With the rare disease approximation, the resulting likelihood is independent of the main effect for X, and is also independent of $r(S, \delta)$ in the risk model

$$\text{pr}(D = 1| \ G, X) = H\{\beta_0 + m(G, X, \beta) + r(S, \delta)\}$$

- If the main effect of X is of interest, too bad, you cannot get it with this approach
- The calculation is easy, and the likelihood is explicit and easy to work with
A Simple Conditional Likelihood

- The main appeal of the conditional likelihood method is that you do not need to assume anything about how X and S are related.

- It does still require that G and X are independent given S.
A Simple Conditional Likelihood

• How can we get at the main effect of X?

• You have to use the **internal controls**!

• Because the risk model and the distribution of G depend on S, you have to make some sort of assumption about the relationship of X and S
A Simple Conditional Likelihood

• However, you do not want to make any distributional assumptions about X.

• Because X is multivariate and typically contains a combination of continuous and discrete random variables

• Modeling such things is not really practical

• Need a semiparametric approach
Additional Assumption

• The **distribution of X does not depend on S**.
  • Because we do not see S among the controls in our population
  • If the assumption fails → a nasty missing data problem
  • There is some sensitivity to major violations of this assumption
  • Testable in the external controls

\[
\text{pr}(X \mid S) \approx \text{pr}(X \mid S, D = 0)
\]

• Forgetting genetics, this is an implicit assumption in all environmental exposure studies
Profile Likelihood

• In many gene-environment interaction contexts, something is assumed known about the relationship between G and X
  • Independence
  • A model for G given X
• In these cases, nonparametric profile likelihood approaches are often used, that avoid assumptions about the distribution of X
  • Useful since X is usually multivariate with mixtures of discrete and continuous variables
Profile Likelihood

- It is tempting to try profile likelihood in this case

**Question**: What is profile likelihood?

- The likelihood function has multiple components

- For the **main data cases**, the retrospective likelihood is

$$
\text{pr}_{\text{Int}} (G = g, X = x, S = s | D = 1)
$$
Profile Likelihood

- For the **main data controls**, since we see neither G nor S, the retrospective likelihood is

  \[
  \text{pr}_{\text{Int}}(X = x \mid D = 0)
  \]

- For the **external data controls**, since we do not care about the likelihood of (X,S) in the external study, the retrospective likelihood contribution is

  \[
  \text{pr}_{\text{Ex}}(G = g \mid X, S, D = 0) = \text{pr}_{\text{Ex}}(G = g \mid S, D = 0)
  \]
Profile Likelihood

- The density/mass of $X$ in the primary study population (not the case-control study) is
  \[ \xi(x) = \Pr_{\text{int}}(X = x) \]

- The density/mass of $S$ in the primary study population (not the case-control study) is
  \[ \alpha(s) = \Pr_{\text{int}}(S = s) \]
Profile Likelihood

- Skipping the details, in terms of the density $\xi(x)$ and the density $\alpha(s)$, the likelihood function is a product of the former across all main study $X$ and the latter among all main study $S$, then times linear functionals across all observations of the form

$$\int S(x, s) \xi(x) \alpha(s) dx \, ds$$

- Here, $S(x, s)$ also depends on the parameters
Profile Likelihood

- In profile likelihood, you treat both the density for X and the density for S as if they were discrete with mass at the observed data points.

- You then hold the parameters fixed and maximize over those discrete distributions.

- In principle, you then have a function only of the parameters.
Profile Likelihood

• In general, the trouble with such profiling is that maximizing over $\xi(x)$ and $\alpha(s)$ as if they were discrete has no explicit solution.

• In those cases, it is hopeless.

• There is also a Lagrange multiplier, since they have to sum to 1.0.
Profile Likelihood

• In our case, profiling has a **semi-explicit solution**!

• The profiled likelihood function depends on the model parameters and an **unconstrained, unknown** but smooth function $\psi(S)$

• Makes some sense because of our risk model

$$
\text{pr}(D = 1 \mid G, X) = H\{\beta_0 + m(G, X, \beta) + r(S, \delta)\}
$$
Profile Likelihood

- The likelihood depends on the internal cases \( \psi(S_1), \ldots, \psi(S_{n_1}) \)

- One possibility is to treat these as unknown parameters, and to maximize over them and the model parameters.

- Obviously, with 1,000 cases, that is a lot of things to maximize over!
Profile Likelihood

- The likelihood depends on the internal cases $\psi(S_1), \ldots, \psi(S_{n_1})$

- We have been able to do this for 1,000 cases, and to get the Fisher Information for the model parameters

- However, dimension reduction seems useful!
**Working Likelihood**

- It is known that $\psi(\bullet)$ is smooth

- Working likelihoods model it flexibly, e.g., with a low-dimensional spline, or a higher dimensional penalized spline

- The low dimensional spline allows easy Fisher information calculations

- $\# \text{ parameters} = \# \text{ model parameters} + \# \text{ knots}$
Working Likelihood

• It is known that \( \psi(\bullet) \) is smooth

• It is even easier to make \( \psi(\bullet) \) linear in \( S \! \)

• Obvious model robustness concerns, etc.
Pseudo-Likelihood

• More is known about $\psi(\bullet)$

• It is a function of the model parameters, $S$, and

$$E\left[ \exp\{m(g, X, \beta)\} \right]$$

• Making the rare disease approximation

$$E\left[ \exp\{m(g, X, \beta)\} \right] \approx E\left[ \exp\{m(g, X, \beta)\} \mid D = 0 \right]$$

$$\approx n_0^{-1} \sum_{i=1}^{n_0} \exp\{m(g, X_i, \beta)\}$$
Pseudo-Likelihood

- The pseudo-likelihood only depends on the model parameters

- Fisher information is not applicable, but the theory is easy and standard errors result
Simulations

- Extensive simulations show what is expected.

- When $X$ is independent of $S$, profile and pseudo-likelihood are more efficient than conditional likelihood (which cannot estimate the main effect of $X$).

- Working likelihood is also fine as long as the model for $\psi(\bullet)$ is correct, otherwise bias.
Simulations

• When $X$ strongly depends on $S$, profile and pseudo-likelihood have biases in their main effects

• Although not in the interaction term

• Something has to give if model assumptions are wrong!
Bladder Cancer

- We did an illustrative analysis of a regular case-control study (PLCO) for bladder cancer

- Roughly 700 cases and 700 controls

- We contrasted a standard logistic analysis with our analysis when we “ignored” the genetic information among the controls.
Bladder Cancer

• The external data set had 1,800 controls

• We used 3 eigenvectors to adjust for population stratification (S)

• Analysis of the actual case control study showed no strong link between smoking, age, sex and S
Bladder Cancer

• Since we had the (G,S) among the internal controls, we could contrast our analysis with regular logistic regression.

• All the methods gave similar results

• While smoking status was a significant main effect, no evidence of a gene-environment interaction
Conclusion

• For studies which use external controls for genetic information, we have listed assumptions and developed methods of estimation