A fast, valid and powerful new test for $E \times GWAS$ interaction, with measurement error in $E$

JSM, 2014, Boston

Donna Spiegelman
Hughes Aschard
Molin Wang
Pete Kraft

Departments of Epidemiology and Biostatistics
Harvard School of Public Health
Boston, MA
Background

Gene-environment (GxE) interactions

A hot topic in genetic association studies that has received a lot of attention over the past years
Gene-environment (GxE) interactions in Human

However, despite numerous studies investigating GxE interactions in human traits and diseases, GWASs have been persistently ineffective in identifying replicable interactions.

Major issues in identifying interactions in genome-wide data:

- **Assessment of complex exposures** *(e.g. air pollution)*
- **Small effect size** *(smaller than marginal SNP effect, e.g. OR≤1.01)*
- **Unavailability of exposure** *(e.g. time dependent event)*
- **Measurement error in E?**

[Aschard et al. 2012]
[Bookman, et al. 2011]
[Duncan, et al. 2010]
Extending previous work on the linear discriminant function approach (Lyles et al. 2014; Armstrong et al. 1989, Efron 1975;…) to develop a SNP by exposure interaction test with similar objectives:

- Reducing or eliminating the bias in estimation of the interaction effect in the presence of exposure measurement error
- Increasing statistical power to detect the interaction effect in the presence of exposure measurement error
Principle

Logistic regression including a SNP x Exposure interaction

$$\text{logit}(pr(D = 1)) = \alpha_0 + \alpha_1 \times E + \alpha_2 \times G + \alpha_3 \times E \times G$$

New idea: Testing $\beta_3$ the interaction between $D$ and $G$, instead of $\alpha_3$ the interaction between $E$ and $G$.

Linear regression where the exposure is treated as the outcome

$$\text{Expectation}(E|D, G) = \beta_0 + \beta_1 \times D + \beta_2 \times G + \beta_3 \times D \times G$$

*Notation:*  
- $G$ is a genetic variant  
- $E$ is a continuous exposure  
- $D$ is a binary outcome
Two models need to be aligned:

The standard logistic model

\[ pr(D = 1|E, G) = \frac{e^{\alpha_0 + \alpha_1 E + \alpha_2 G + \alpha_3 E \times G}}{1 + e^{\alpha_0 + \alpha_1 E + \alpha_2 G + \alpha_3 E \times G}} \]

An exposure based model

\[ E|D, G = \mathcal{N}(\beta_0 + \beta_1 D + \beta_2 G + \beta_3 D \times G, \sigma^2) \]

**Notation:**
- \(G\) is a genetic variant
- \(E\) is a continuous exposure
- \(D\) is a binary outcome
Two models need to be aligned:

It can be shown that the parameters of these two models can be linked as follows:

\[ \beta_1 = \alpha_1 \sigma^2 \]

\[ \beta_3 = \alpha_3 \sigma^2 \]

\[ \beta_0 = \frac{2 \sigma^2 \times \left( \log \left( \frac{p_{10}}{1 - p_{10}} \right) - \alpha_0 \right) - \beta_1^2}{2 \times \beta_1} \]

\[ \beta_2 = \frac{2 \sigma^2 \times \left( \log \left( \frac{p_{11} \times (1 - p_{10})}{p_{10} \times (1 - p_{11})} \right) - \alpha_2 \right) - \beta_3^2 - 2 \beta_1 \beta_3 - \beta_0 \beta_3}{2 \times (\beta_1 + \beta_3)} \]

Notation:

\[ p_{10} = pr(D|G = 0) \]
\[ p_{11} = pr(D|G = 1) \]
\[ p_1 = pr(D) \]
\[ p_G = pr(G = 1) \]
Two models need to be aligned:

System to solve to write $\beta$ in terms of $\alpha$:

$$p_{11} \times p_G + p_{10} \times (1 - p_G) = pr(D) = p_1$$

$$p_{10} = \int \frac{e^{\alpha_0 + \alpha_1 E}}{1 + e^{\alpha_0 + \alpha_1 E}} \{p_{10} \times \mathcal{N}(\beta_0 + \beta_1, \sigma^2) + (1 - p_{10}) \times \mathcal{N}(\beta_0, \sigma^2)\} dE$$

$$p_{11} = \int \frac{e^{\alpha_0 + \alpha_1 E + \alpha_2 + \alpha_3 G \times E}}{1 + e^{\alpha_0 + \alpha_1 E + \alpha_2 + \alpha_3 G \times E}} \{p_{11} \times \mathcal{N}(\beta_0 + \beta_1 + \beta_2 + \beta_3, \sigma^2) + (1 - p_{11}) \times \mathcal{N}(\beta_0 + \beta_2, \sigma^2)\} dE$$

$$\sigma^2 = \int \{p_{11} \times p_G \times \mathcal{N}(\beta_0 + \beta_1 + \beta_2 + \beta_3, \sigma^2) + p_{10} \times (1 - p_G) \times \mathcal{N}(\beta_0 + \beta_1, \sigma^2) + (1 - p_{11}) \times p_G \times \mathcal{N}(\beta_0 + \beta_2, \sigma^2) + (1 - p_{10}) \times (1 - p_G) \times \mathcal{N}(\beta_0, \sigma^2)\} dE$$

Notations:
- $p_{10} = pr(D | G = 0)$
- $p_{11} = pr(D | G = 1)$
- $p_1 = pr(D)$
- $p_G = pr(G = 1)$
Two models need to aligned:

System to solve to obtain all parameters

\[ p_{11} \times p_G + p_{10} \times (1 - p_G) = pr(D) = p_1 \]

\[ p_{10} = \int \frac{e^{\alpha_0 + \alpha_1 E}}{1 + e^{\alpha_0 + \alpha_1 E}} \{p_{10} \times \mathcal{N}(\beta_0 + \beta_1, \sigma^2) + (1 - p_{10}) \times \mathcal{N}(\beta_0 + \beta_2, \sigma^2)\} dE \]

\[ p_{11} = \int \frac{e^{\alpha_0 + \alpha_1 E + \alpha_2 + \alpha_3 G \times E}}{1 + e^{\alpha_0 + \alpha_1 E + \alpha_2 + \alpha_3 G \times E}} \{p_{11} \times \mathcal{N}(\beta_0 + \beta_1 + \beta_2 + \beta_3, \sigma^2)\} dE \]

\[ \sigma^2 = \int E \times \{p_{11} \times p_G \times \mathcal{N}(\beta_0 + \beta_1 + \beta_2 + \beta_3, \sigma^2) + p_{10} \times \mathcal{N}(\beta_0 + \beta_1, \sigma^2) + (1 - p_{11}) \times p_G \times \mathcal{N}(\beta_0 + \beta_2, \sigma^2) \times (1 - p_G) \times \mathcal{N}(\beta_0, \sigma^2)\} dE \]

No simple solution to the system of equations
Simulation framework

Simulation scheme:

- Analysis of case/control study design
- A binary genetic variant $G$
- A normally distributed exposure $E$
- 10,000 replicate simulated per scenario considered

Algorithm:

1) set $p_{10}, p_{11}, \beta_0, \beta_1, \beta_2, \beta_3, \sigma$ and $p_G$ (for simplicity we fixed $\beta_0 = 0$, $\sigma = 1$ and $p_G = 0.3$ in all simulations)

2) for each replicate, generate $G$, $E$ and $D$ as follows:
   a. Sample a binary $G$ from Bernoulli ($p_G$)
   b. Sample $D$ from Bernoulli ($p_{11}$) if $G=1$ and from Bernoulli ($p_{10}$) if $G=0$
   c. Sample $E$ from $\mathcal{N}(\beta_0 + \beta_1 D + \beta_2 G + \beta_3 D \times G, \sigma^2)$
   d. Sample cases/controls from the full cohort
Null model

The test shows correct type I error rate across all parameters we considered

\[ \text{logit}(pr(D = 1)) = \alpha_0 + \alpha_1 \times E + \alpha_2 \times G + \alpha_3 \times E \times G \]

\[ E|D, G = \mathcal{N}(\beta_0 + \beta_1 D + \beta_2 G + \beta_3 D \times G, \sigma^2) \]

<table>
<thead>
<tr>
<th>Simulation parameters</th>
<th>Estimated ORs</th>
<th>rejection rate of H0 for alpha=5%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Logit.</td>
</tr>
<tr>
<td>( p_{10} )</td>
<td>( p_{11} )</td>
<td>( \exp^\beta_1(D) )</td>
</tr>
<tr>
<td>0.01</td>
<td>0.015</td>
<td>1.0</td>
</tr>
<tr>
<td>0.01</td>
<td>0.015</td>
<td>1.5</td>
</tr>
<tr>
<td>0.01</td>
<td>0.015</td>
<td>2.0</td>
</tr>
<tr>
<td>0.01</td>
<td>0.020</td>
<td>1.0</td>
</tr>
<tr>
<td>0.01</td>
<td>0.020</td>
<td>1.5</td>
</tr>
<tr>
<td>0.01</td>
<td>0.020</td>
<td>2.0</td>
</tr>
<tr>
<td>0.05</td>
<td>0.073</td>
<td>1.0</td>
</tr>
<tr>
<td>0.05</td>
<td>0.073</td>
<td>1.5</td>
</tr>
<tr>
<td>0.05</td>
<td>0.07</td>
<td>2.0</td>
</tr>
<tr>
<td>0.05</td>
<td>0.095</td>
<td>1.0</td>
</tr>
<tr>
<td>0.05</td>
<td>0.10</td>
<td>1.5</td>
</tr>
<tr>
<td>0.05</td>
<td>0.10</td>
<td>2.0</td>
</tr>
<tr>
<td>0.30</td>
<td>0.39</td>
<td>1.0</td>
</tr>
<tr>
<td>0.30</td>
<td>0.39</td>
<td>1.5</td>
</tr>
<tr>
<td>0.30</td>
<td>0.39</td>
<td>2.0</td>
</tr>
<tr>
<td>0.30</td>
<td>0.46</td>
<td>1.0</td>
</tr>
<tr>
<td>0.30</td>
<td>0.46</td>
<td>1.5</td>
</tr>
<tr>
<td>0.30</td>
<td>0.46</td>
<td>2.0</td>
</tr>
</tbody>
</table>

\textbf{Notation:}

\( p_{10} = pr(D|G = 0) \)

\( p_{11} = pr(D|G = 1) \)

\( p_1 = pr(D) \)

\( p_G = pr(G = 1) \)
Improving in power

Power increases with main exposure effect and effects but is almost independent of the main genetic effect

Relative power of the approaches was measured by deriving the ratio of median chi-squared

Model simulated

\( OR_E = \{1; 3\} \)
\( OR_{GE} = \{1, \ldots, 2\} \)
\( \Pr(G) = 0.3 \)
\( \Pr(D = 1) = 0.1 \)

\( OR_G = 1.2 \)

\( OR_G = 2 \)
Impact of measurement error on interaction effect estimate

\[
\text{pr}(D = 1|E, G) = \frac{e^{\alpha_0 + \alpha_1 E + \alpha_2 G + \alpha_3 E \times G}}{1 + e^{\alpha_0 + \alpha_1 E + \alpha_2 G + \alpha_3 E \times G}}
\]

\[
E|D, G = N(\beta_0 + \beta_1 E + \beta_2 G + \beta_3 E \times G, \sigma^2)
\]

Under measurement error: \(E_{\text{measure}} = E + N(0, \sigma^2_{\text{noise}})\)

\[
\text{pr}(D = 1|E_{\text{measure}}, G) = \frac{e^{\alpha_0^* + \alpha_1^* E + \alpha_2^* G + \alpha_3^* E \times G}}{1 + e^{\alpha_0^* + \alpha_1^* E + \alpha_2^* G + \alpha_3^* E \times G}}
\]

\[
E_{\text{measure}}|D, G = N(\beta_0^* + \beta_1^* E + \beta_2^* G + \beta_3^* E \times G, \sigma^2_{\text{noise}})
\]

Let \(\rho = \frac{\sigma^2}{\sigma^2 + \sigma^2_{\text{noise}}}\),

We have \(\alpha_3 = \frac{\alpha_3^*}{\rho}, \beta_3 = \beta_3^*\)
Improving GxE effect estimation

Impact of measurement error on effect estimates

Interaction effect decreases toward 0 for logistic regression

Noise is generated so that:

\[ E_{measured} = E + \mathcal{N}(0, \sigma_{noise}) \]

Model simulated

\[ \text{OR}_E = 2 \]
\[ \text{OR}_G = 1.5 \]
\[ \text{OR}_{GE} = 1.5 \]
\[ \text{Pr}(G) = 0.3 \]
\[ \text{Pr}(D = 1) = 0.1 \]
Improving GxE effect estimation

Impact of measurement error on effect estimates

Interaction effect decreases toward 0 for logistic regression while it remains unbiased for the reverse test.

Noise is generated so that:

\[ E_{\text{measured}} = E + \mathcal{N}(0, \sigma_{\text{noise}}) \]

**Model simulated**

\[ OR_E = 2 \]
\[ OR_G = 1.5 \]
\[ OR_{GE} = 1.5 \]
\[ \Pr(G) = 0.3 \]
\[ \Pr(D = 1) = 0.1 \]
Computation time in hours for the analysis of a simulated ExGWAS case-control study of one million SNPs (common setting in genome-wide association studies):

<table>
<thead>
<tr>
<th>Sample size (cases+controls)</th>
<th>Expected time in hours*</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Linear regression</td>
<td>Logistic regression</td>
</tr>
<tr>
<td>1,000</td>
<td>0.7</td>
<td>1.8</td>
</tr>
<tr>
<td>2,000</td>
<td>1.0</td>
<td>3.4</td>
</tr>
<tr>
<td>10,000</td>
<td>3.4</td>
<td>19.2</td>
</tr>
<tr>
<td>20,000</td>
<td>6.7</td>
<td>35.7</td>
</tr>
<tr>
<td>60,000</td>
<td>21.1</td>
<td>109.8</td>
</tr>
</tbody>
</table>

*Average over 1,000 simulated tests and scaled to a million tests
We evaluated the performance of the proposed approach for the analysis of breast cancer cases and controls (Cancer Genetic Markers of Susceptibility – CGEMS (Nature Genetics, Hunter et al 2007 (PMID: 17529973))

- 1133 breast cancer cases and 1133 age-matched controls.
- Genome-wide data were available for all individuals. It included ~500,000 genotyped SNPs

We considered two exposures

- BMI assessed at time of case diagnosis
- Alcohol use (gm/day) assessed at time of case diagnosis
Quality control and data filtering:

We removed all SNPs with:
- minor allele frequency < 0.01
- Hardy-Weinberg p-value < 0.0001

We removed subjects with:
- missing exposure
- pre-menopausal women at enrollment

⇒ After QC, 2,256 individuals and 459,999 SNPs were available for analysis
Real data analysis

Distribution of exposures:

**BMI**

- Mean BMI = 26.13
- Std BMI = 4.95

**Alcohol Consumption**

- Mean Alc. use = 65.33
- Std Alc. use = 80.4
Real data analysis: SNP x BMI interaction

SNP x BMI interaction effect on breast cancer

Logistic regression

\[
\logit(\Pr(D = 1)) \sim \alpha_1 \text{BMI} + \alpha_2 G + \alpha_3 \text{BMI} \times G + \alpha_4 C
\]

Test \( \alpha_3 = 0 \)

(Covariates \( C \) include PC of genotypes)
SNP x BMI interaction effect on breast cancer

Reverse test

\[ BMI \sim \beta_1 D + \beta_2 G + \beta_3 D \times G + \beta_4 C \]

Test \( \beta_3 = 0 \)

(Covariates \( C \) include PC of genotypes)
Real data analysis: **SNP x BMI interaction**

**SNP x BMI interaction effect on breast cancer**

We confirmed the validity of the test, i.e. both approaches show uniform $p$-value distribution (expected in GWAS, as most variant tested are assumed to be under the null)
Real data analysis: SNP x BMI interaction

Top signals do match between the two tests
Real data analysis: **SNP x BMI interaction**

But no analysis showed genome-wide significant signal (p-value < 5x10^{-8})
Real data analysis: **SNP x alcohol interaction**

**SNP x alcohol interaction effect on breast cancer**

Logistic regression

$$
\text{logit}(\Pr(D = 1)) \sim \alpha_1 ALC + \alpha_2 G + \alpha_3 ALC \times G + \alpha_4 C
$$

Test $\alpha_3 = 0$

*(Covariates $C$ include PC of genotypes)*
Real data analysis: SNP x alcohol interaction

**SNP x alcohol interaction effect on breast cancer**

Reverse test

\[ ALC \sim \beta_1 D + \beta_2 G + \beta_3 D \times G + \beta_4 C \]

Test \( \beta_3 = 0 \)

(Covariates \( C \) include PC of genotypes)
Real data analysis: SNP x alcohol interaction

Again, some signals matching but no genome-wide significant signal with either approach.
Real data analysis: Comparison with previous studies

**Previous studies:** no well-established gene-environment interaction published for breast cancer despite several studies

However, at least one candidate exists for alcohol use: SNP rs17468277 (gene *CASP8*) by alcohol use interaction on breast cancer risk [Nickels et al, 2013]

But no signal in the CGEMS data:

<table>
<thead>
<tr>
<th>Test</th>
<th>beta</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic regression</td>
<td>0.9906</td>
<td>0.3044</td>
</tr>
<tr>
<td>Reverse test</td>
<td>-0.9931</td>
<td>0.2394</td>
</tr>
</tbody>
</table>
The reverse test is valid, more powerful and much faster to calculate

Further work could include the following:

• Power loss due to measurement error/sample size calculations for $ExGWAS$ that consider this feature
• Extension to ordinal and continuous $E$
• Sensitivity of valid and power improvement of the reverse tests to deviations from conditional normality of $E|D, G$