Correcting for selection bias in adaptive two-stage designs

David Robertson[1] Toby Prevost[2] Jack Bowden[1,3]

[1]MRC Biostatistics Unit, University of Cambridge

[2]Imperial College London

[3]MRC Integrative Epidemiology Unit, University of Bristol

MCP 2017

Outline

1. Introduction
2. General framework
3. GWAS data
4. Seamless phase II/III trials
5. Summary

Introduction

- The two-stage 'learn and confirm' strategy is widely implemented:
  - Seamless phase II/III trials
  - Biomarker research
  - Genome wide association study (GWAS)
- Ranking and selecting candidates can induce *bias* into estimates at study completion.

- An unbiased estimator can easily be found by just using the stage 2 data.
- However, this estimator suffers from lower precision.
- Instead we seek an efficient unbiased estimator that uses data from *both* stages.

- An unbiased estimator can easily be found by just using the stage 2 data.
- However, this estimator suffers from lower precision.
- Instead we seek an efficient unbiased estimator that uses data from *both* stages.

- Take the expectation of the stage 2 data, conditional on the stage 1 data and selection rules.

- The resulting estimator is the **UMVCUE**: uniformly minimum variance conditionally-unbiased estimator.
  - Lower variance than any other unbiased estimator.

- Unbiased estimation in two-stage framework introduced by Cohen and Sackrowitz (1989).
- Key assumption in the literature is that the stage 1 population parameter estimates are *independent* random variables.
- This may not be a reasonable assumption to make!

- In the GWAS setting, SNPs on the same genomic region may be in linkage disequilibrium.
- In biomarker trials, measurements of different biomarkers may be correlated within each person.
- In a multi-arm adaptive trial with common control group, the estimates of each treatment's benefit over the control are correlated by definition.

General framework

Stage 1

- K correlated continuous stage 1 parameter estimates $\mathbf{X} = (X_1, \ldots, X_K)$.
- $\mathbf{X} \sim N(\boldsymbol{\mu}, V)$ where $\boldsymbol{\mu}$ is vector of unknown means and $V = (V_{ij})$ is known covariance matrix.
- Ordered stage 1 estimates $X_{(i)}$, where $X_{(1)} \geq \cdots \geq X_{(K)}$.
- Let $\sigma_i^2 = V_{ii}$ for $i = 1, \ldots, K$.

Stage 2

- Let Y_j be stage 2 estimate of jth ranked candidate, where $Y_j \sim N(\mu_{(j)}, \tau_j^2)$.
- At the end of stage 2, the aim is to efficiently estimate $\mu_{(j)}$.

Estimation

- The MLE for $\mu_{(j)}$ is weighted average of the data:

$$\hat{\mu}_{(j)} = \frac{\tau_j^2 X_{(j)} + \sigma_{(j)}^2 Y_j}{\sigma_{(j)}^2 + \tau_j^2}.$$

- The MLE is biased because it does not take into account the selection rules or correlation.
- The stage 2 data Y_j is unbiased, but has lower precision.

Calculating the UMVCUE

- Let Q be the event $\{\mathbf{X} : X_1 \geq \cdots \geq X_K\}$.
- Without loss of generality, we condition on Q.

- The statistic $\mathbf{Z_j} = (Z_{1j}, \ldots, Z_{Kj})$ is sufficient and complete for $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$, where

$$Z_{ij} = X_i + \frac{V_{ij}}{\tau_j^2} Y_j$$

- We have a closed-form expression for the UMVCUE for μ_j, which we can write as

$$\text{UMVCUE} = \text{MLE} + \text{Bias}.$$

The UMVCUE for μ_j given Q is

$$\hat{U}_j = \frac{\tau_j^2 Z_{jj}}{\sigma_j^2 + \tau_j^2} - \frac{\tau_j^2}{\sqrt{\sigma_j^2 + \tau_j^2}} \frac{\phi(W_1) - \phi(W_2)}{\Phi(W_1) - \Phi(W_2)}$$

where

$$W_i = \frac{k_i \sqrt{\sigma_j^2 + \tau_j^2}}{\tau_j^2} - \frac{Z_{jj}}{\sqrt{\sigma_j^2 + \tau_j^2}} \qquad \text{for } i = 1, 2$$

$$k_1 = \min(A_1), \quad k_2 = \max(A_2),$$

$$A_1 = \left\{ \frac{\tau_j^2 (Z_{ij} - Z_{i+1,j})}{V_{ij} - V_{i+1,j}} : V_{ij} > V_{i+1,j} \; ; \; i = 1, \ldots, K-1 \right\},$$

$$A_2 = \left\{ \frac{\tau_j^2 (Z_{ij} - Z_{i+1,j})}{V_{ij} - V_{i+1,j}} : V_{ij} < V_{i+1,j} \; ; \; i = 1, \ldots, K-1 \right\}.$$

Application to GWAS data

- We apply our methodology to data from a GWAS for Crohn's disease by the Welcome Trust Case Control Consortium.
- Identified 12 SNPs associated with disease status at genome-wide significance.
- A replication study was then reported by Parkes et al. (2007) in a follow-up cohort.
- This is a two-stage design with a genome-wide association study (stage 1) followed by a replication study (stage 2).

- The table below shows the estimated odds ratios (ORs) for stages 1 and 2, as well as the overall MLE.
- The UMVCUEs are calculated assuming that log ORs for the SNPs are uncorrelated.

Chr	SNP	Stage 1	Stage 2	MLE	UMVCUE
5p13	**rs17234657**	**1.55**	**1.16**	**1.39**	**1.16**
5p13	**rs9292777**	**1.38**	**1.34**	**1.37**	**1.39**
10q24	rs10883365	1.27	1.18	1.24	1.16
18p11	rs2542151	1.35	1.15	1.27	1.15
5q33	**rs13361189**	**1.51**	**1.38**	**1.46**	**1.40**
3p21	rs9858542	1.26	1.17	1.22	1.17
5q33	**rs4958847**	**1.35**	**1.36**	**1.36**	**1.35**
5q23	rs10077785	1.29	1.19	1.25	1.19
1q24	rs12035082	1.22	1.14	1.19	1.15
21q22	rs2836754	1.22	1.15	1.19	1.16
1q31	rs10801047	1.38	1.47	1.42	1.44

- The 1st and 2nd ranked SNPs are on 5p13.
- The 5th and 7th ranked SNPs are on 5q33.
- A natural question to ask is how the OR estimates are affected by linkage disequilibrium.
- Only those SNPs that meet a selection criteria in stage 1 continue to stage 2.
- We extend our framework to account for ranking by p-value:

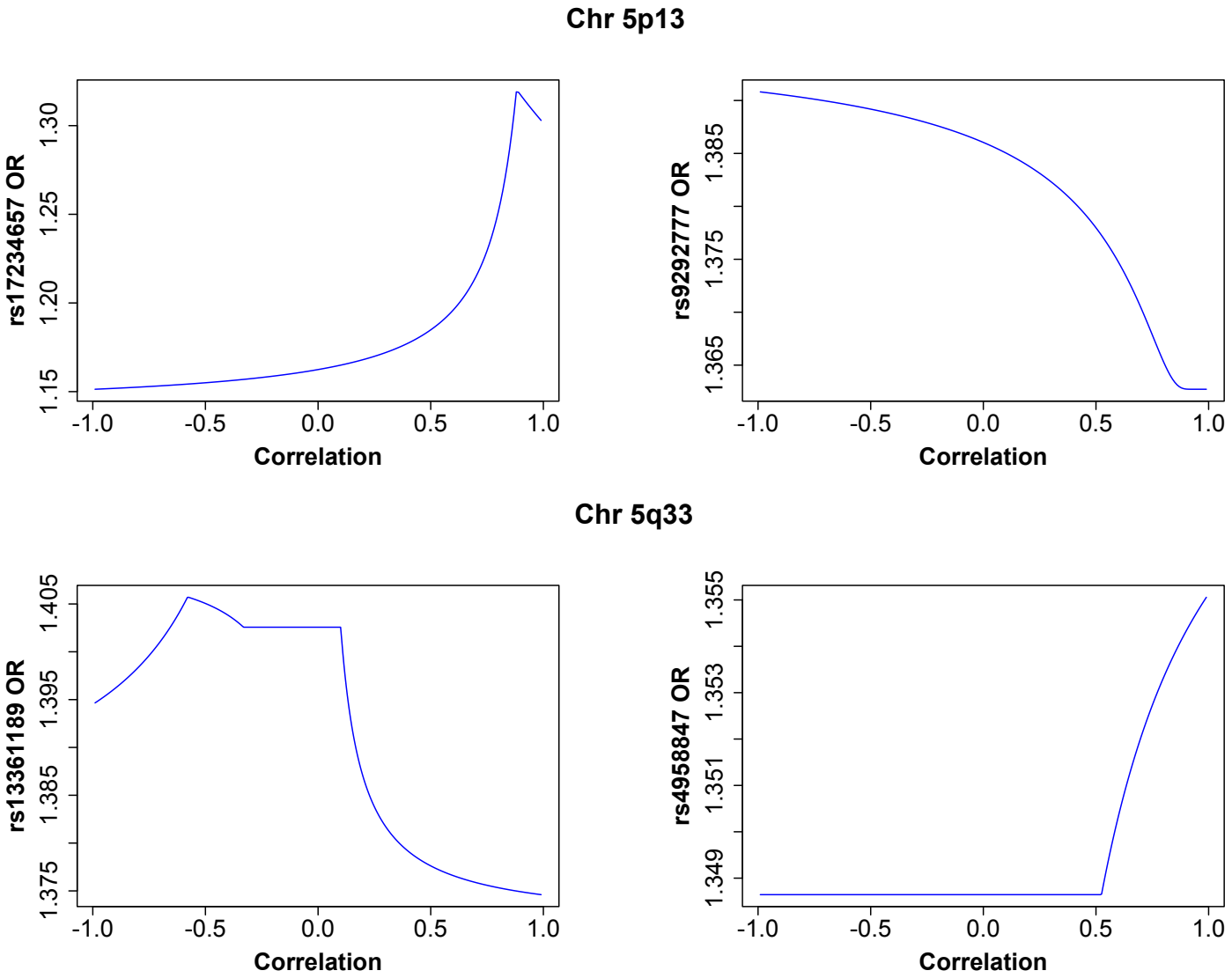
$$Q_2 = \left\{ \mathbf{X} : \frac{|X_1|}{\sigma_1} \geq \frac{|X_2|}{\sigma_2} \geq \cdots \geq \frac{|X_K|}{\sigma_K} \geq \Phi^{-1}(1 - p_{\text{crit}}/2) \right\}.$$

Results

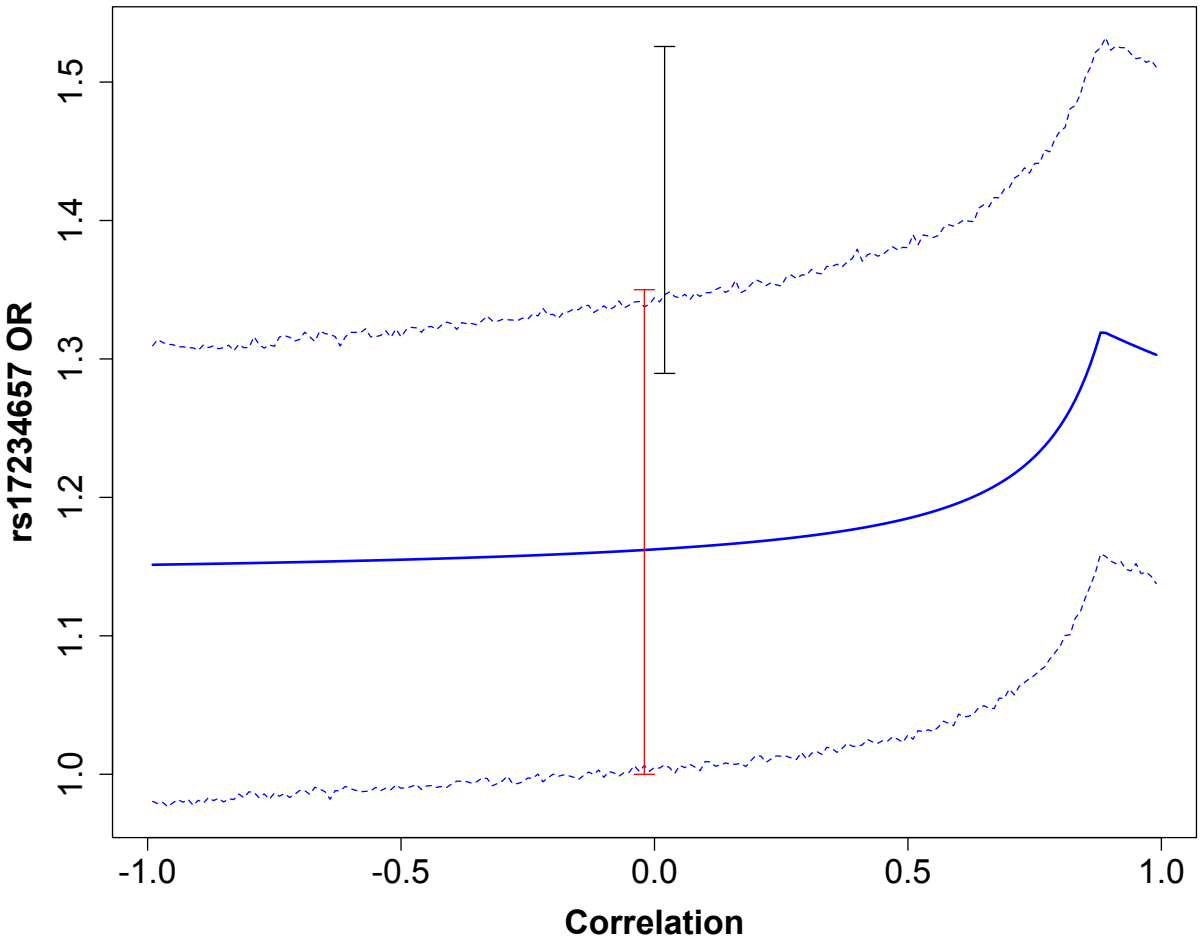
- We take each pair of SNPs on the same chromosomal region and calculate the UMVCUE as the correlation between the log ORs change.

- We assume the log ORs X_{j_1} and X_{j_2} follow a bivariate normal distribution, with correlation coefficient ρ

$$\begin{pmatrix} X_{j_1} \\ X_{j_2} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_{j_1} \\ \mu_{j_2} \end{pmatrix}, \begin{pmatrix} \sigma^2_{j_1} & \rho\sigma_{j_1}\sigma_{j_2} \\ \rho\sigma_{j_1}\sigma_{j_2} & \sigma^2_{j_2} \end{pmatrix} \right).$$


Chr 5p13
rs1723465 7 OR
1.15
1.20
1.25
1.30
-1.0
-0.5
0.0
0.5
1.0
Correlation
rs9292777 OR
1.365
1.375
1.385
-1.0
-0.5
0.0
0.5
1.0
Correlation
Chr 5q33
rs13361189 OR
1.375
1.385
1.395
1.405
-1.0
-0.5
0.0
0.5
1.0
Correlation
rs4958847 OR
1.349
1.351
1.353
1.355
-1.0
-0.5
0.0
0.5
1.0
Correlation

Can also construct confidence intervals using parametric bootstrap:



Seamless phase II/III trials

- Starting point is the adaptive seamless design (ASD) used in Kimani et al. (2013).
- Consider an ASD where stage 1 is used to select the most promising treatment and stage 2 is for confirmatory analysis.
- Assume stage 1 sample means $X_i \sim N(\mu_i, \sigma^2_{1i})$.
- Let n_{1i} denote the number of subjects allocated to treatment i in stage 1, where $i = 0$ corresponds to the control treatment.

- At the end of stage 1, rank the treatments according to their standardised treatment difference:

$$\frac{X_i - X_0}{\sqrt{\text{Var}(X_i - X_0)}} \geq \frac{X_j - X_0}{\sqrt{\text{Var}(X_j - X_0)}}$$

$$\Longrightarrow \quad \frac{X_i - X_0}{\sqrt{\sigma_{1i}^2 + \sigma_{10}^2}} \geq \frac{X_j - X_0}{\sqrt{\sigma_{1j}^2 + \sigma_{10}^2}}$$

- Let the treatment with the highest ranking be denoted by S.
- Early stopping of the trial for futility: the trial continues to stage 2 if $\frac{X_S - X_0}{\sqrt{\sigma_{1S}^2 + \sigma_{10}^2}} \geq b$.

- Stage 2 sample means $Y_i \sim N(\mu_i, \sigma_{2i}^2)$.
- Let n_{2i} denote the number of subjects allocated to treatment i $(i = 0, S)$ in stage 2.
- If there is a common variance σ^2, then $\sigma_{2i}^2 = \sigma^2 / n_{2i}$.
- **Aim:** to estimate the treatment difference $\theta_S = \mu_S - \mu_0$.

- If we have unequal treatment effect variances, using the theory for the multivariate normal setting, we can derive the UMVCUE.
- Let $\Theta_i = X_i - X_0$ denote the stage 1 sample mean treatment difference for treatment i. Then $\Theta_i \sim N\left(\mu_i - \mu_0, \sigma_{1i}^2 + \sigma_{10}^2\right)$.
- $\boldsymbol{\Theta}$ follows a multivariate normal distribution with mean $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$ and covariance matrix Σ, where $\theta_i = \mu_i - \mu_0$ and $\Sigma_{ij} = \text{Cov}(\Theta_i, \Theta_j)$. Hence

$$\begin{aligned} \Sigma_{ii} &= \sigma_{1i}^2 + \sigma_{10}^2 & i &\in \{1, \ldots, K\} \\ \Sigma_{ij} &= \sigma_{10}^2 & i, j &\in \{1, \ldots, K\}\ ,\ i \neq j \end{aligned}$$

- Note how the treatment differences are *correlated* in this framework.

Multiple testing with the closure principle

- Now look at the context of formal hypothesis testing.
- For a single null hypothesis H with first stage p-value p_1, the trial is stopped early if $p_1 > \alpha_0$.
- Assume we are testing K directional null hypotheses $H_i : \mu_i \leq \mu_0$, comparing the K treatments with the control.
- Want to strongly control the familywise error rate (FWER) at a pre-specified level α.
- Control the FWER strongly using the *closure principle* (CP).

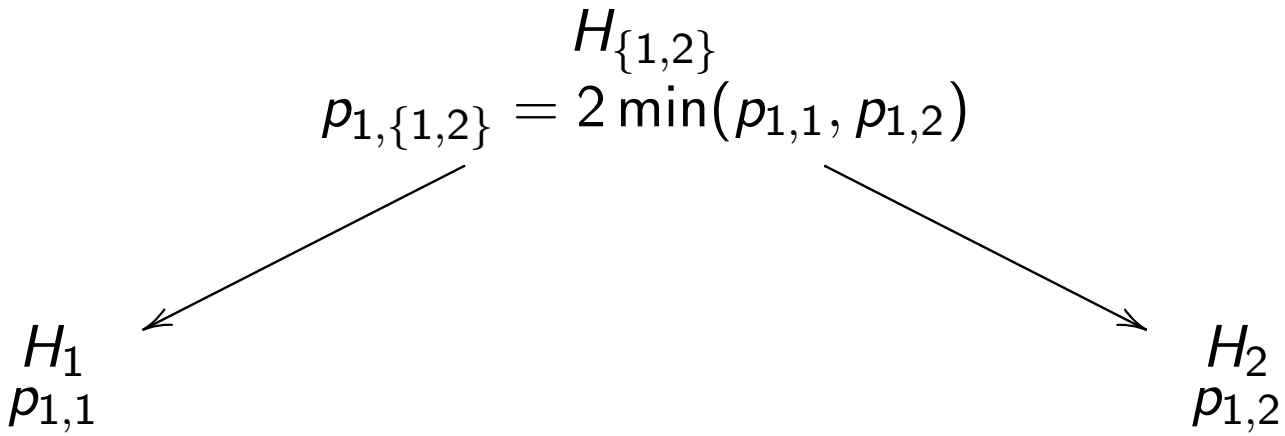
UMVCUE with Bonferonni correction

- Consider using the closed testing procedure for the stage 1 data with early stopping for futility, using the Bonferonni correction for multiplicity.
- Stage 1 (unadjusted) p-values $p_{1,i}$

$$p_{1,i} = 1 - \Phi\left(\frac{X_i - X_0}{\sqrt{\sigma_{1i}^2 + \sigma_{10}^2}}\right)$$

- Let $r(X_i) = \dfrac{X_i - X_0}{\sqrt{\sigma_{1i}^2 + \sigma_{10}^2}}$

- Comparing $K = 2$ treatments with control:



- The Bonferroni adjusted first stage p-value $p_{1,\{1,2\}}$ for the intersection hypothesis $H_{\{1,2\}}$ is

$$p_{1,\{1,2\}} = 2\min(p_{1,1}, p_{1,2}) = 2\left[1 - \Phi\left(\max_{i\in\{1,2\}} r(X_i)\right)\right].$$

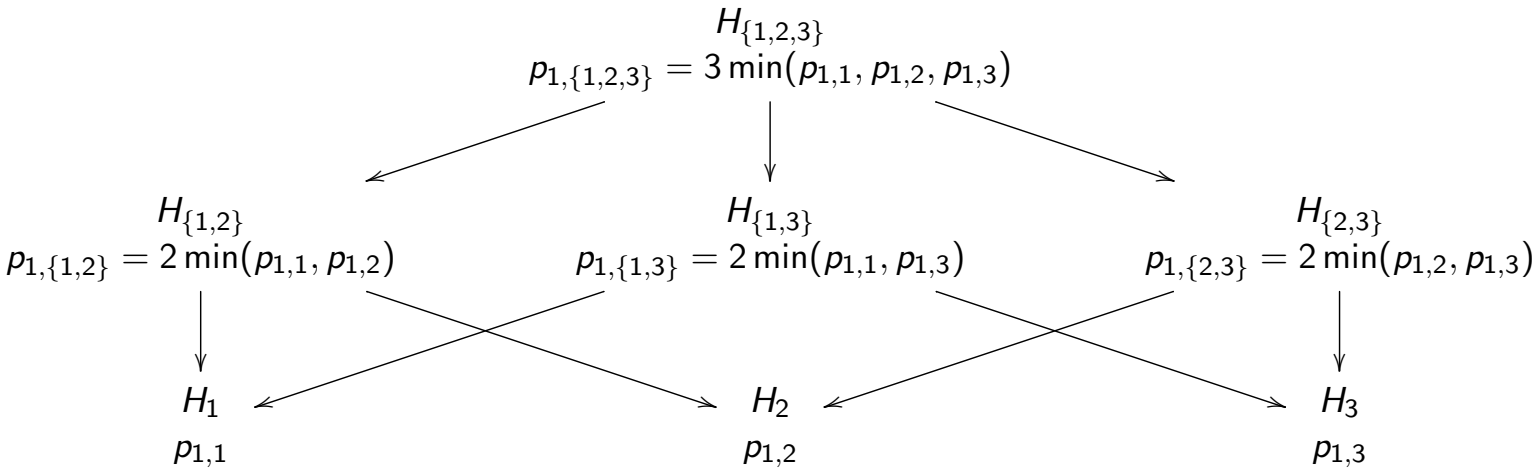
- By the CP, treatment $j \in \{1, 2\}$ continues to stage 2 if

$$p_{1,\{1,2\}} < \alpha_0 \implies \max_{i \in \{1,2\}} r(X_i) > \Phi^{-1}(1 - \alpha_0/2)$$

$$p_{1,j} < \alpha_0 \implies r(X_j) > \Phi^{-1}(1 - \alpha_0)$$

- Suppose treatment 1 is ranked above treatment 2, i.e. $r(X_1) > r(X_2)$.
- Then $\max_{i \in \{1,2\}} r(X_i) = r(X_1)$, and treatment 1 continues to stage 2 if $r(X_1) > \Phi^{-1}(1 - \alpha_0/2)$.

- So conditional on $Q = \{\mathbf{X} : r(X_1) > r(X_2)\}$, the UMVCUE for $\theta_1 = \mu_1 - \mu_0$ fits into the model framework, where $b = \Phi^{-1}(1 - \alpha_0/2)$.


$H_{\{1,2,3\}}$
$p_{1,\{1,2,3\}} = 3\min(p_{1,1}, p_{1,2}, p_{1,3})$
$H_{\{1,2\}}$
$p_{1,\{1,2\}} = 2\min(p_{1,1}, p_{1,2})$
$H_{\{1,3\}}$
$p_{1,\{1,3\}} = 2\min(p_{1,1}, p_{1,3})$
$H_{\{2,3\}}$
$p_{1,\{2,3\}} = 2\min(p_{1,2}, p_{1,3})$
H_1
$p_{1,1}$
H_2
$p_{1,2}$
H_3
$p_{1,3}$

Example

- Compare three experimental drugs with a placebo for the treatment of generalised anxiety disorder. Outcomes are normally distributed with common standard deviation $\sigma = 6$.
- Trial is planned with equal allocations to each treatment, with $n_1 = n_2 = 71$ subjects per group, but the randomisation procedure used leads to an unequal allocations.

	Stage 1				**Stage 2**	
	n_{1i}	Observed	z-statistic	p_{1i}	n_{2i}	Observed
Placebo	70	0.4	—	—	68	−0.3
Treatment 1	72	2.2	1.787	0.0369	75	1.7
Treatment 2	68	2.4	1.958	0.0251	70	2.2
Treatment 3	74	3.2	2.799	0.0026	71	1.9

- Aim is to take forward as many treatments as possible that pass a first stage p-value futility threshold, set at $\alpha_0 = 0.1$.
- Stage 1 Bonferroni-adjusted p-values are:

$$p_{1,\{1,2,3\}} = 0.0077$$
$$p_{1,\{1,2\}} = 0.0503,\ p_{1,\{1,3\}} = 0.0051,\ p_{1,\{2,3\}} = 0.0051$$
$$p_{1,1} = 0.0369,\ p_{1,2} = 0.0251,\ p_{1,3} = 0.0026$$

- All of the adjusted p-values are less than $\alpha_0 \implies$ all of the treatments (and placebo) continue to stage 2.

Stage 1 Rank	Treatment	Stage 2	Naïve	UMVCUE
1	3	2.200	2.505	2.285
2	2	2.500	2.250	2.020
3	1	2.000	1.900	2.062

The Kimani estimator for treatment 3 is 2.197, using a pooled variance.

Summary

- We have a general framework for unbiased estimation in two-stage trials in the presence of selection and correlation.
- The UMVCUE can be decomposed into the MLE + Bias.
- It is important to correctly account for correlation – the UMVCUE that ignores correlation can be substantially biased.
- Our estimation strategy can be applied in practice to the GWAS and seamless phase II/III trial settings.
- Further work:
  - Construction of confidence intervals
  - Multi-stage trials

References

J. Bowden and F. Dudbridge
Unbiased estimation of odds ratios: combining genomewide association scans with replication studies.
Genetic epidemiology, 33(5):406–418, 2009.

J. Bowden and E. Glimm
Unbiased estimation of selected treatment means in two-stage trials.
Biometrical Journal, 50(4):515–527, 2008.

A. Cohen and H. B. Sackrowitz
Two stage conditionally unbiased estimators of the selected mean.
Statistics & Probability Letters, 8(3):273–278, 1989.

P. K. Kimani, S. Todd and N. Stallard
Conditionally unbiased estimation in phase II/III clinical trials with early stopping for futility.
Statistics in Medicine, 32(17):2893–2910, 2013.

M. Parkes et al.
Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's diseases susceptibility.
Nature Genetics, 39(7):830–832, 2007.

D. S. Robertson, A. T. Prevost and J. Bowden
Accounting for selection and correlation in the analysis of two-stage genome-wide association studies.
Biostatistics, 17(4):634–649, 2016.

D. S. Robertson, A. T. Prevost and J. Bowden
Unbiased estimation in seamless phase II/III trials with unequal treatment effect variances and hypothesis driven selection rules.
Statistics in Medicine, 35(22):3907–3922, 2016.

UMVCUE for two-sided test

The UMVCUE for μ_j given Q_2 is

$$\hat{U}_j = \frac{\tau_j^2 Z_{jj}}{\sigma_j^2 + \tau_j^2} - \frac{\tau_j^2}{\sqrt{\sigma_j^2 + \tau_j^2}} \frac{\sum_{i=1}^{M} \phi(W_{1i}) - \phi(W_{2i})}{\sum_{i=1}^{M} \Phi(W_{1i}) - \Phi(W_{2i})}$$

where

$$W_{1i} = \frac{b_i \sqrt{\sigma_j^2 + \tau_j^2}}{\tau_j^2} - \frac{Z_{jj}}{\sqrt{\sigma_j^2 + \tau_j^2}}, \qquad W_{2i} = \frac{a_i \sqrt{\sigma_j^2 + \tau_j^2}}{\tau_j^2} - \frac{Z_{jj}}{\sqrt{\sigma_j^2 + \tau_j^2}},$$

$$\bigcup_{i=1}^{M} [a_i, b_i] = \left(\bigcap_{i=1}^{K-1} (A_{1i} \cap A_{2i}) \cup (A_{3i} \cap A_{4i}) \right) \cap (A_5 \cup A_6)$$

$$A_{1i} = \left\{ Y : (\sigma_i V_{i+1,j} - \sigma_{i+1} V_{ij}) Y \geq \tau_j^2 (\sigma_i Z_{i+1,j} - \sigma_{i+1} Z_{ij}) \right\},$$

$$A_{2i} = \left\{ Y : (\sigma_i V_{i+1,j} + \sigma_{i+1} V_{ij}) Y \leq \tau_j^2 (\sigma_{i+1} Z_{ij} + \sigma_i Z_{i+1,j}) \right\},$$

$$A_{3i} = \left\{ Y : (\sigma_i V_{i+1,j} + \sigma_{i+1} V_{ij}) Y \geq \tau_j^2 (\sigma_{i+1} Z_{ij} + \sigma_i Z_{i+1,j}) \right\},$$

$$A_{4i} = \left\{ Y : (\sigma_i V_{i+1,j} - \sigma_{i+1} V_{ij}) Y \leq \tau_j^2 (\sigma_i Z_{i+1,j} - \sigma_{i+1} Z_{ij}) \right\},$$

$$A_5 = \left\{ Y : V_{Kj} Y \leq \tau_j^2 \left[Z_K - \sigma_K \Phi^{-1}(1 - p_{\text{crit}}/2) \right] \right\},$$

$$A_6 = \left\{ Y : V_{Kj} Y \geq \tau_j^2 \left[Z_K + \sigma_K \Phi^{-1}(1 - p_{\text{crit}}/2) \right] \right\}.$$

UMVCUE for treatment difference

The UMVCUE for $\theta_1 = \mu_1 - \mu_0$ given Q is

$$\hat{U} = \frac{\tau^2 Z_1}{\nu^2 + \tau^2} - \frac{\tau^2}{\sqrt{\nu^2 + \tau^2}} \frac{\phi(W_1) - \phi(W_2)}{\Phi(W_1) - \Phi(W_2)}$$

where

$$W_i = \frac{k_i \sqrt{\nu^2 + \tau^2}}{\tau^2} - \frac{Z_1}{\sqrt{\nu^2 + \tau^2}} \qquad \text{for } i = 1, 2$$

$$k_1 = \min(A_1, A_2, A_3), \quad k_2 = \max(A_4, A_5),$$

$$A_1 = \frac{\tau^2}{\nu^2}\left(Z_1 - \frac{b}{\lambda_1}\right)$$

$$A_2 = \left\{ \frac{\tau^2(\lambda_1 Z_1 - \lambda_2 Z_2)}{\sigma_{10}^2(\lambda_1 - \lambda_2) + \lambda_1 \sigma_{11}^2} : \lambda_1 \sigma_{11}^2 > (\lambda_2 - \lambda_1)\sigma_{10}^2 \right\},$$

$$A_3 = \left\{ \frac{\tau^2(\lambda_j Z_j - \lambda_{j+1} Z_{j+1})}{\sigma_{10}^2(\lambda_j - \lambda_{j+1})} : \sigma_{1,j+1}^2 > \sigma_{1j}^2 \, ; \, j = 2, \ldots, K-1 \right\},$$

$$A_4 = \left\{ \frac{\tau^2(\lambda_1 Z_1 - \lambda_2 Z_2)}{\sigma_{10}^2(\lambda_1 - \lambda_2) + \lambda_1 \sigma_{11}^2} : \lambda_1 \sigma_{11}^2 < (\lambda_2 - \lambda_1)\sigma_{10}^2 \right\},$$

$$A_5 = \left\{ \frac{\tau^2(\lambda_j Z_j - \lambda_{j+1} Z_{j+1})}{\sigma_{10}^2(\lambda_j - \lambda_{j+1})} : \sigma_{1,j+1}^2 < \sigma_{1j}^2 \, ; \, j = 2, \ldots, K-1 \right\}.$$