Design and Analysis of a Two-Phase Study for Multivariate Longitudinal Outcomes

Chiara Di Gravio, Ran Tao, Jonathan Schildcrout

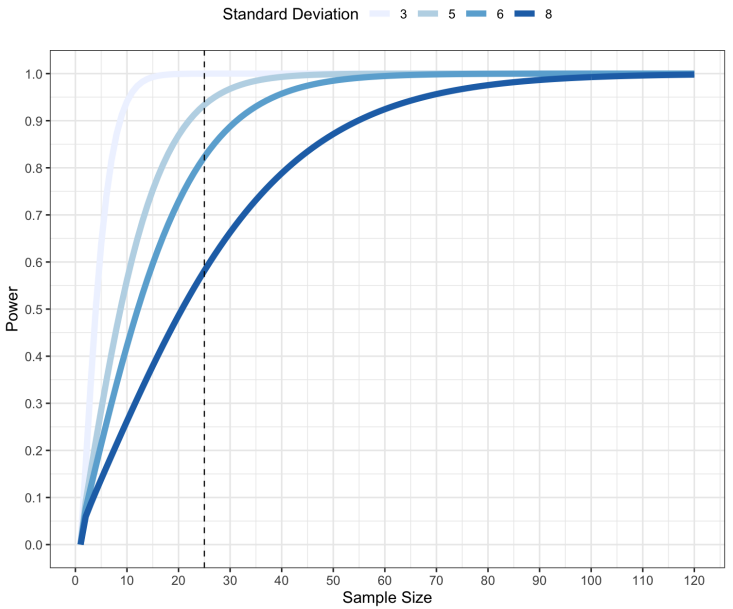
Vanderbilt University

September 29, 2021

- Motivation
- Lung Health Study
- Two-Phase Outcome Dependent Sampling
- Results
  - Simulation Studies
  - Lung Health Study

Motivation

- Electronic health records and existing cohort studies provide easily accessible data on phenotype
- Researchers might be interested in an exposure that is unavailable and expensive to collect
- We want to use the available data to identify the most informative subjects for whom the expensive exposure will be collected.



- By sampling informative individuals instead of sampling at random, we could decrease the standard deviation associated with an estimate. For a fixed sample size we would be able to achieve higher power

The Lung Health Study (LHS)

- LHS was a multi-center randomized clinical trial recruiting adults aged 35 to 60 with moderate lung function impairment. The intervention aimed at getting people to quit smoking.
- Hansel et al. (2013) conducted a GWAS and identified two novel SNPs associated with accelerated lung function decline.
- Even though Hansel et al. had complete DNA data, this is often not the case. In what follows we assume that genetic data had not yet been collected

- We are going to consider 2,563 smokers with at least two observations and examine the association between one of the SNPs identified by Hansel et al. and the rate of lung function decline over time
- We are going to consider a scenario where data on outcome and confounders are available on everyone, but data on SNP can only be collected on 800 subjects
- The expensive exposure is the presence/absence of at least one copy of the T-allele at SNP rs177852
- Two correlated longitudinal outcomes are: forced expiratory volume (FEV) in the first second of an exhalation following bronchodilators use, and forced vital capacity (FVC)

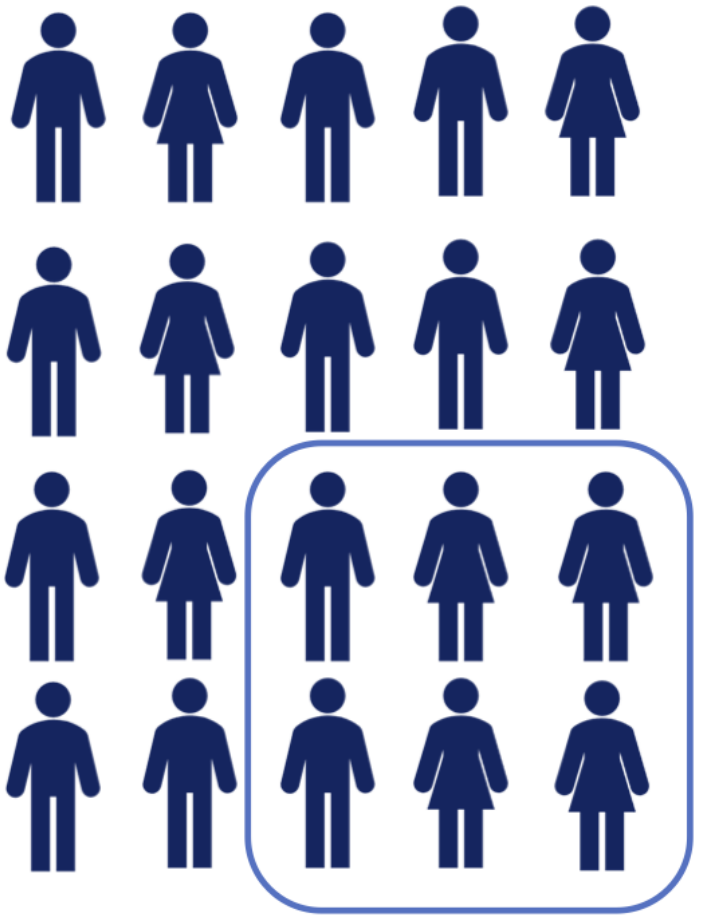
The model of interest is:

$$Y_{1ij} = \beta_{10} + \beta_{1s} snp_i + \beta_{1t} t_{ij} + \boldsymbol{\beta_{1c} c_i} + \beta_{1st} snp_i t_{ij} + b_{10i} + b_{11i} t_{ij} + \epsilon_{1ij}$$

$$Y_{2ij} = \beta_{20} + \beta_{2s} snp_i + \beta_{2t} t_{ij} + \boldsymbol{\beta_{2c} c_i} + \beta_{2st} snp_i t_{ij} + b_{20i} + b_{21i} t_{ij} + \epsilon_{2ij}$$

- Y_{1ij} is the FEV for subject i at visit j
- Y_{2ij} is the FVC for subject i at visit j
- snp_i is an indicator for the presence of at least one copy of the allele at rs177852
- $(b_{10i}, b_{11i}, b_{20i}, b_{21i}) \sim N(\mathbf{0}, \boldsymbol{D})$ are the random intercept and slope for subject i
- $\boldsymbol{c_i}$ is a set of covariates
- $(\epsilon_{1ij}, \epsilon_{2ij}) \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ are the error terms independent of the random effects

Two-Phase Outcome Dependent Sampling



Two-Phase Design

Outcome and inexpensive covariates are observed on all subjects in phase one. A subset of the subjects is chosen for a phase two where the expensive exposure is measured

Outcome Dependent Sampling

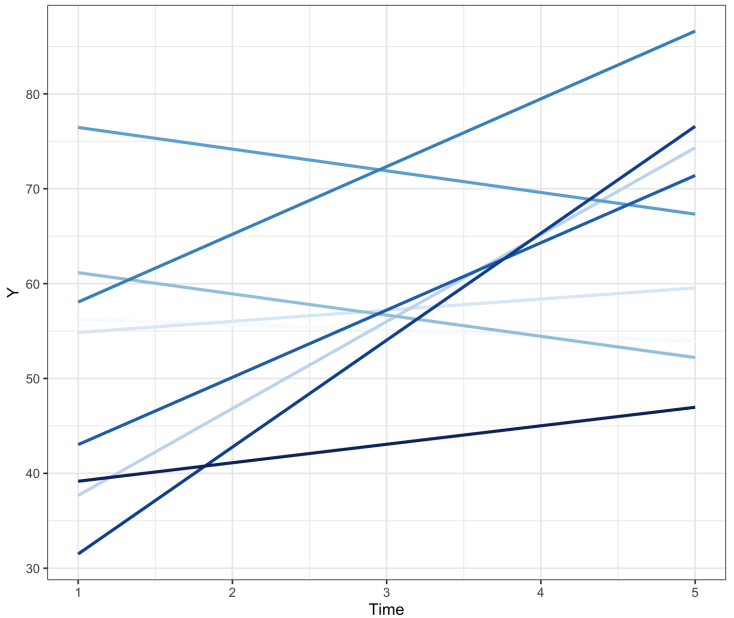
Subjects for whom the exposure is collected are selected based on their observed outcome

Who Are the Most Informative Subjects?

Outcome dependent sampling aims to maximize observed response variability

- Cross-sectional binary outcome: the case-control study with an equal number of cases and controls is the most cost-efficient design
- Cross-sectional continuous outcome: informative individuals are those with high or low values of the outcome

Longitudinal Continuous Outcome



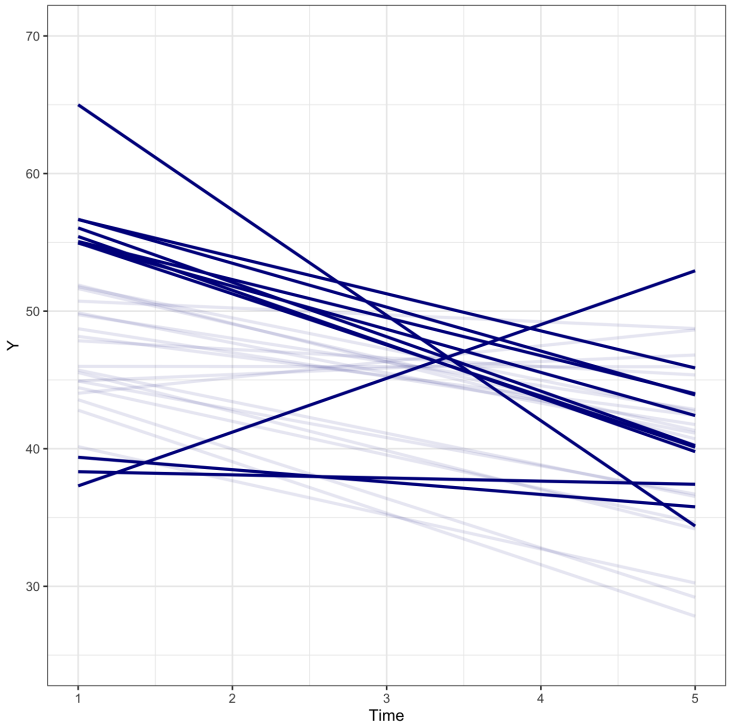
- The outcome is repeatedly collected over time. For a subject i the outcome is a vector $\boldsymbol{Y}_i = (Y_1, \ldots, Y_{n_i})$
- The multiple measures of the outcome allow us to separate changes over time within individuals from changes between individuals
- Who the most informative individuals are will depend on whether our interest lies in a time-varying exposure or a time-fixed exposure

Design 1: ODS Design (Schildcrout et al, 2013)

- Sample informative individuals based on their estimated subject-specific intercept and/or slope.
- For each subject fit $E(Y_{ij}|t_{ij}) = q_{0i} + q_{1i}t_{ij}$
    - q_{0i} is the subject-specific mean outcome at baseline, q_{1i} is the subject-specific rate of change

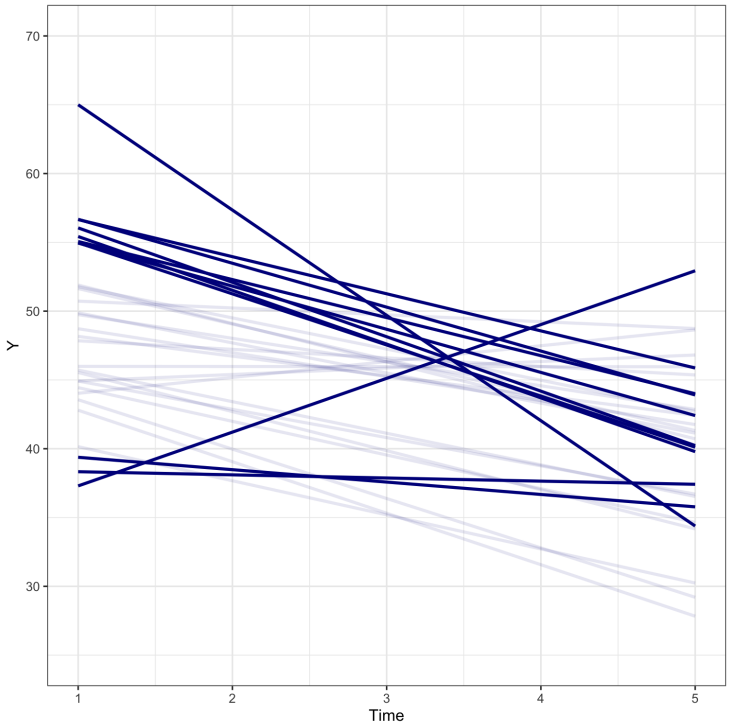
ODS Intercept

- Assign higher probability of being sampled to subjects with extreme q_{0i}



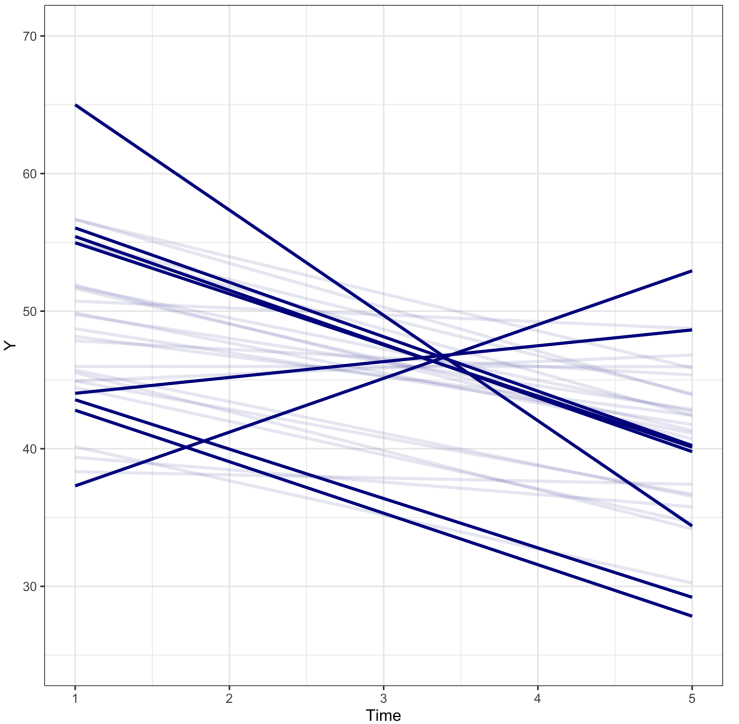
ODS Intercept

- Assign higher probability of being sampled to subjects with extreme q_{0i}

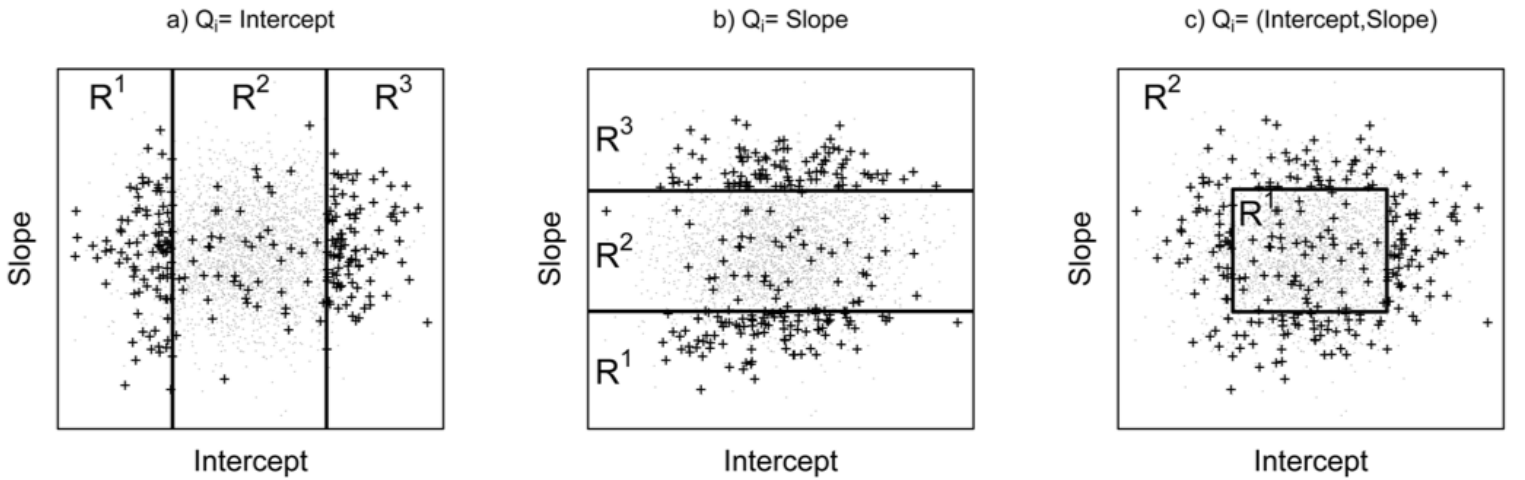


ODS Slope

- Assign higher probability of being sampled to subjects with extreme q_{1i}



- Sort values of q_{0i} and/or q_{1i}
- Introduce cutpoints that define sampling strata from which we sample with different probabilities



Picture taken directly from Schildcrout et al (2013)

Design 2: BDS Design (Sun et al, 2017)

- Sample informative individuals based on the best linear unbiased predictor (BLUP) estimates of random intercept and slope
- Fit $Y_{ij} = \alpha_0 + \alpha_t t_{ij} + \boldsymbol{\alpha_c c_i} + a_{0i} + a_{1i} t_{ij} + \epsilon_{ij}$
- Estimate a_{0i} and a_{1i}
- Sort values of a_{0i} and/or a_{1i} and introduce cutpoints that define sampling strata from which we sample with different probabilities

Design 2: BDS Design (Sun et al, 2017)

- Sample informative individuals based on the best linear unbiased predictor (BLUP) estimates of random intercept and slope
- Fit $Y_{ij} = \alpha_0 + \alpha_t t_{ij} + \boldsymbol{\alpha_c c_i} + a_{0i} + a_{1i} t_{ij} + \epsilon_{ij}$
- Estimate a_{0i} and a_{1i}
- Sort values of a_{0i} and/or a_{1i} and introduce cutpoints that define sampling strata from which we sample with different probabilities
- **BDS Intercept**. If we are interested in a time-fixed exposure we can assign higher probability of being sampled to subjects with extreme a_{0i}.
- **BDS Slope**. If we are interested in a time-varying exposure we can assign higher probability of being sampled to subjects with extreme a_{1i}.

- ODS designs require each subject to be observed at least two times, while BDS designs can include subjects with one observation
- BDS designs can be more cost-efficient than ODS designs when we have unbalanced data or when there is a strong association between the outcome and the confounders

A Note

- Sampling at the extremes can have practical difficulties:
  - Requires a large underlying population available to have enough individuals with extreme values of the outcome
  - Results in sampling subjects with conditions that are not of interest

Analysis

- Failure to account for the design will generally result in biased estimates
- Analysis procedures that properly account for a two-phase ODS design can be divided in two groups:
  - **Conditional Likelihood:** only include subjects sampled in phase two
  - **Full Likelihood:** combine complete information available on subjects sampled in phase two with partial information (outcome and confounders) available on subjects not sampled in phase two

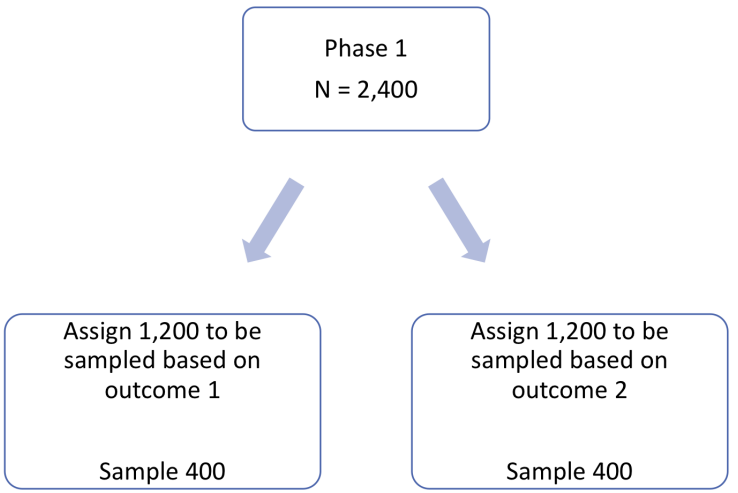
Two-Phase ODS with Multivariate Outcomes

Why Multivariate Outcomes?

- Many biomedical applications aim to study multiple (potentially correlated) outcomes
- Permit efficiency gains for multiple targets
- For example, the National Heart, Lung and Blood Institute Exome Sequencing Project sought to boost statistical power to detect genetic associations by oversampling subjects with extreme values of LDL cholesterol, blood pressure or BMI

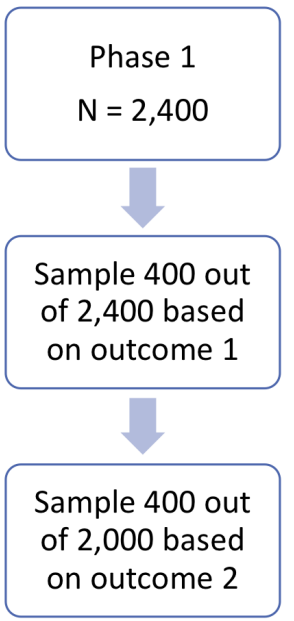
Independent Sampling

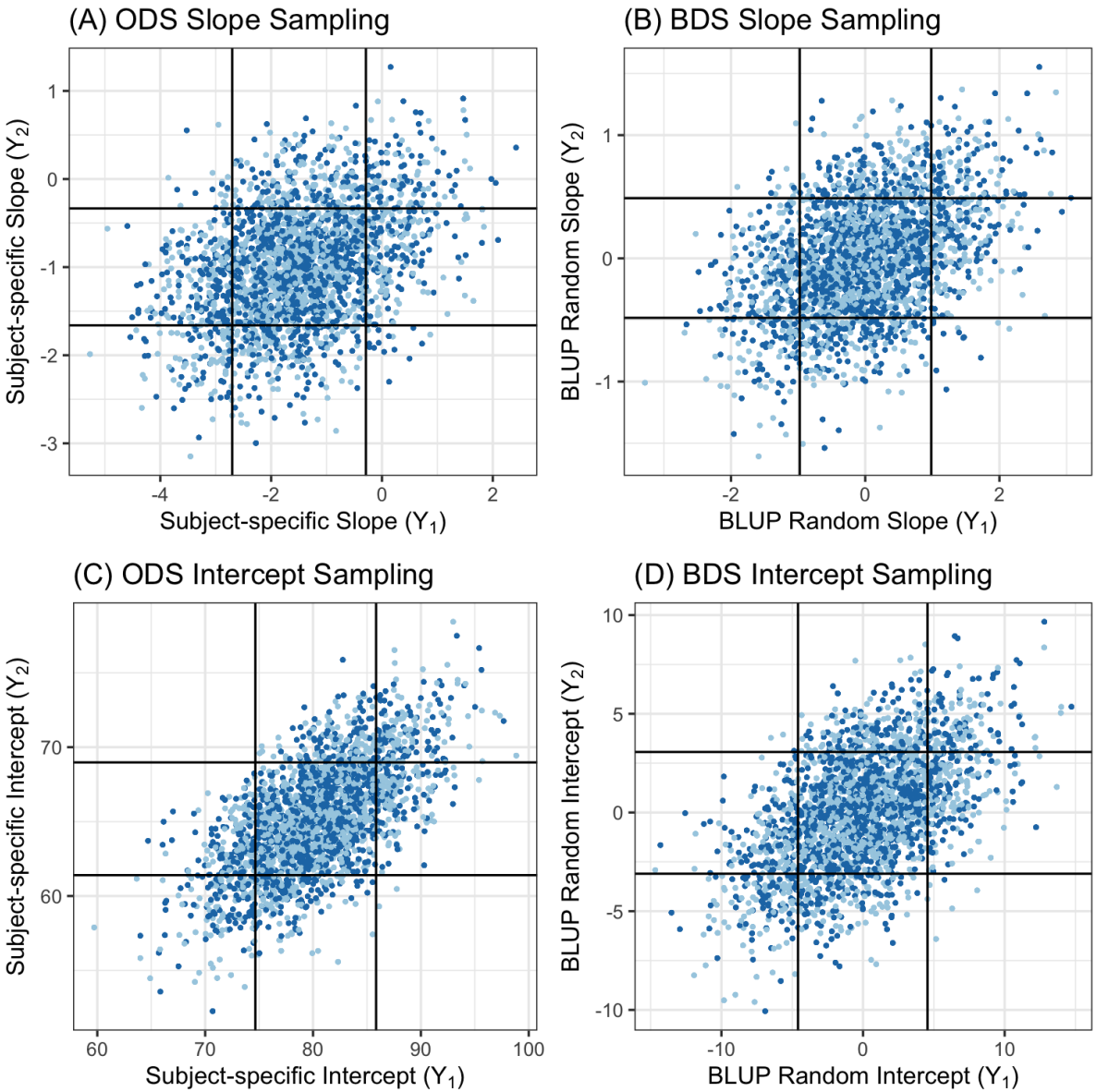
- Assign a fraction of individuals to be sampled based on outcome 1 and the remaining based on outcome 2
- Use the univariate ODS or BDS to select the most informative subjects for each outcome



Sequential Sampling

- Sample individuals based on outcome 1 first using univariate ODS or BDS
- Sample the remaining individuals based on outcome 2 using univariate ODS or BDS




(A) ODS Slope Sampling
Subject-specific Slope (Y_2)
Subject-specific Slope (Y_1)
(B) BDS Slope Sampling
BLUP Random Slope (Y_2)
BLUP Random Slope (Y_1)
(C) ODS Intercept Sampling
Subject-specific Intercept (Y_2)
Subject-specific Intercept (Y_1)
(D) BDS Intercept Sampling
BLUP Random Intercept (Y_2)
BLUP Random Intercept (Y_1)

Re-Use Data

- Using multivariate outcomes allow us to re-use data from a previous two-phase ODS study where sampling was conducted based on outcome 1, but interest shifts to the association between the expensive exposure and outcome 2
- Re-using data can be seen as a special case of independent sampling where we assign each subject to be sampled based on outcome 1

Conditional Likelihood Approach

- We consider $w = 2$ outcomes and assume n individuals are sampled for phase two
- The conditional likelihood approach explicitly conditions on a subject being sampled for phase two. Analyses are based on the ascertainment corrected log-likelihood:

$$\sum_{i=1}^{n} \left[log f(\boldsymbol{Y}_{1i}, \boldsymbol{Y}_{2i} | snp_i, \boldsymbol{t}_i, \boldsymbol{c}_i) - \underbrace{\sum_{w=1}^{2} \tau_{wi} log \left\{ \sum_k \pi(R_{wk}) \int_{R_{wk}} f(\boldsymbol{q}_{wi} | snp_i, \boldsymbol{t}_i, \boldsymbol{c}_i) d\boldsymbol{q}_{wi} \right\}}_{\text{ASCERTAINMENT CORRECTION}} \right]$$

Conditional Likelihood Approach

- We consider $w = 2$ outcomes and assume n individuals are sampled for phase two
- The conditional likelihood approach explicitly conditions on a subject being sampled for phase two. Analyses are based on the ascertainment corrected log-likelihood:

$$\sum_{i=1}^{n} \left[log f(\boldsymbol{Y}_{1i}, \boldsymbol{Y}_{2i} | snp_i, \boldsymbol{t}_i, \boldsymbol{c}_i) - \underbrace{\sum_{w=1}^{2} \tau_{wi} log \left\{ \sum_k \pi(R_{wk}) \int_{R_{wk}} f(\boldsymbol{q}_{wi} | snp_i, \boldsymbol{t}_i, \boldsymbol{c}_i) d\boldsymbol{q}_{wi} \right\}}_{\text{ASCERTAINMENT CORRECTION}} \right]$$

- $\pi(R_{wk})$ is the probability of being sampled associated to stratum R_{wk}
- q_{wi} is the subject i observed value of the sampling variable for outcome w
- τ_{wi} indicates whether subject i has been assigned to be sampled based on outcome w

Multiple Imputation

- We fill-in the missing data on snp in individuals not sampled $(S_i = 0)$ by drawing from $[snp_i|\boldsymbol{Y}_{1i}, \boldsymbol{Y}_{2i}, \boldsymbol{t}_i, \boldsymbol{c}_i, S_i = 0]$
- Because sampling depends only on $(\boldsymbol{Y}_{1i}, \boldsymbol{Y}_{2i}, \boldsymbol{t}_i, \boldsymbol{c}_i)$:

$$pr(snp_i|\boldsymbol{Y}_{1i}, \boldsymbol{Y}_{2i}, \boldsymbol{t}_i, \boldsymbol{c}_i, S_i = 0) = pr(snp_i|\boldsymbol{Y}_{1i}, \boldsymbol{Y}_{2i}, \boldsymbol{t}_i, \boldsymbol{c}_i) = pr(snp_i|\boldsymbol{Y}_{1i}, \boldsymbol{Y}_{2i}, \boldsymbol{t}_i, \boldsymbol{c}_i, S_i = 1)$$

- We build the imputation model using available data on all subjects. By Bayes' theorem:

$$\frac{pr(snp_i = 1|\boldsymbol{Y}_{1i}, \boldsymbol{Y}_{2i}, \boldsymbol{t}_i, \boldsymbol{c}_i, S_i = 0)}{pr(snp_i = 0|\boldsymbol{Y}_{1i}, \boldsymbol{Y}_{2i}, \boldsymbol{t}_i, \boldsymbol{c}_i, S_i = 0)} = \frac{f(\boldsymbol{Y}_{1i}, \boldsymbol{Y}_{2i}|snp_i = 1, \boldsymbol{t}_i, \boldsymbol{c}_i)}{f(\boldsymbol{Y}_{1i}, \boldsymbol{Y}_{2i}|snp_i = 0, \boldsymbol{t}_i, \boldsymbol{c}_i)} \times \frac{pr(snp_i = 1|\boldsymbol{t}_i, \boldsymbol{c}_i)}{pr(snp_i = 0|\boldsymbol{t}_i, \boldsymbol{c}_i)}$$

- We assume the Gaussian linear mixed effects model and we let $\boldsymbol{Y}_i = (\boldsymbol{Y}_{1i}, \boldsymbol{Y}_{2i})$, $\boldsymbol{\mu}_{x,i} = E(\boldsymbol{Y}_i | snp_i = x, \boldsymbol{t}_i, \boldsymbol{c}_i)$ and $\boldsymbol{V}_i = Var(\boldsymbol{Y}_i | snp_i, \boldsymbol{t}_i, \boldsymbol{c}_i)$

- After log-transforming both sides of the equation, the imputation model becomes:

$$\boldsymbol{Y}_i^T \boldsymbol{V}_i^{-1} (\boldsymbol{\mu}_{1,i} - \boldsymbol{\mu}_{0,i}) - \frac{1}{2}\left(\boldsymbol{\mu}_{1,i}^T \boldsymbol{V}_i^{-1} \boldsymbol{\mu}_{1,i} - \boldsymbol{\mu}_{0,i}^T \boldsymbol{V}_i^{-1} \boldsymbol{\mu}_{0,i}\right) + log\left[\frac{pr(snp_i = 1|\boldsymbol{c}_i)}{pr(snp_i = 0|\boldsymbol{c}_i)}\right]$$

- The imputation model is an offsetted logistic regression

Algorithm

1) On sampled subjects, fit the linear mixed effects model of interest to obtain estimates $\hat{\boldsymbol{\beta}}^{(0)}$ and $\widehat{Cov}(\hat{\boldsymbol{\beta}}^{(0)})$

2) Draw $\beta^{(k)}$ from $N(\hat{\boldsymbol{\beta}}^{(k-1)}, \widehat{Cov}(\hat{\boldsymbol{\beta}}^{(k-1)}))$ and calculate the offset

3) Fit the logistic imputation model using the offset calculated in 2) and obtain the parameters $\hat{\boldsymbol{\gamma}}^{(k)}$ and $\widehat{Cov}(\hat{\boldsymbol{\gamma}}^{(k)})$

4) Draw $\boldsymbol{\gamma}^{(k)}$ from $N(\hat{\boldsymbol{\gamma}}^{(k)}, \widehat{Cov}(\hat{\boldsymbol{\gamma}}^{(k)}))$ and calculate $\hat{p}^{(k)} = P(snp_i = 1|\boldsymbol{c}_i; \boldsymbol{\gamma}^{(k)})$

5) For unsampled subjects impute snp_i using $\hat{p}^{(k)}$

6) Fit the linear mixed effect model of interest on everyone to obtain estimates $\hat{\boldsymbol{\beta}}^{(k)}$ and $\widehat{Cov}(\hat{\boldsymbol{\beta}}^{(k)})$

Algorithm

1) On sampled subjects, fit the linear mixed effects model of interest to obtain estimates $\hat{\boldsymbol{\beta}}^{(0)}$ and $\widehat{Cov}(\hat{\boldsymbol{\beta}}^{(0)})$

2) Draw $\beta^{(k)}$ from $N(\hat{\boldsymbol{\beta}}^{(k-1)}, \widehat{Cov}(\hat{\boldsymbol{\beta}}^{(k-1)}))$ and calculate the offset

3) Fit the logistic imputation model using the offset calculated in 2) and obtain the parameters $\hat{\boldsymbol{\gamma}}^{(k)}$ and $\widehat{Cov}(\hat{\boldsymbol{\gamma}}^{(k)})$

4) Draw $\boldsymbol{\gamma}^{(k)}$ from $N(\hat{\boldsymbol{\gamma}}^{(k)}, \widehat{Cov}(\hat{\boldsymbol{\gamma}}^{(k)}))$ and calculate $\hat{p}^{(k)} = P(snp_i = 1|\boldsymbol{c}_i; \boldsymbol{\gamma}^{(k)})$

5) For unsampled subjects impute snp_i using $\hat{p}^{(k)}$

6) Fit the linear mixed effect model of interest on everyone to obtain estimates $\hat{\boldsymbol{\beta}}^{(k)}$ and $\widehat{Cov}(\hat{\boldsymbol{\beta}}^{(k)})$

7) Repeat steps 2) to 6) m times to create a complete dataset, fit the linear mixed effects model of interest on the complete data, and store the results

8) Repeat the imputation + analysis M times and combine the results using Rubin's rule

Results

Simulation Study

- We generate data from the bivariate mixed effects model:

$$Y_{1ij} = \beta_{10} + \beta_{1s} snp_i + \beta_{1t} t_{ij} + \beta_{1st} snp_i t_{ij} + \beta_{1c} c_i + \beta_{1ct} c_i t_{ij} + b_{10i} + b_{11i} t_{ij} + \epsilon_{1ij}$$

$$Y_{2ij} = \beta_{20} + \beta_{2s} snp_i + \beta_{2t} t_{ij} + \beta_{2st} snp_i t_{ij} + \beta_{2c} c_i + \beta_{2ct} c_i t_{ij} + b_{20i} + b_{21i} t_{ij} + \epsilon_{2ij}$$

- $Pr(snp_i = 1) = 0.3$
- $c_i \sim N(-0.15 + \delta_s snp_i, 1)$ where δ_s controls the correlation between snp_i and c_i
- We assume that data on outcome and confounders are available on 2,400 subjects, but due to resource constraints we can only collect snp on 800 subjects

- We consider 3 designs:
  - simple random sampling (RS)
  - ODS slope sampling
  - BDS slope sampling
- For each outcome w, we define three regions and select (N_{w1}, N_{w2}, N_{w3}) subjects from the three regions
- We estimate the parameters of interest using the ascertainment corrected log-likelihood and multiple imputation

Scenario	(N_{w1}, N_{w2}, N_{w3})	δ_s
A	(150, 100, 150)	-0.05
B	(150, 100, 150)	-2
C	(180, 40, 180)	-0.05

We consider 6 design + inference procedures:

Consider the 800 people sampled in phase two

- RS with maximum likelihood
- ODS slope with ascertainment corrected maximum likelihood
- BDS slope with ascertainment corrected maximum likelihood

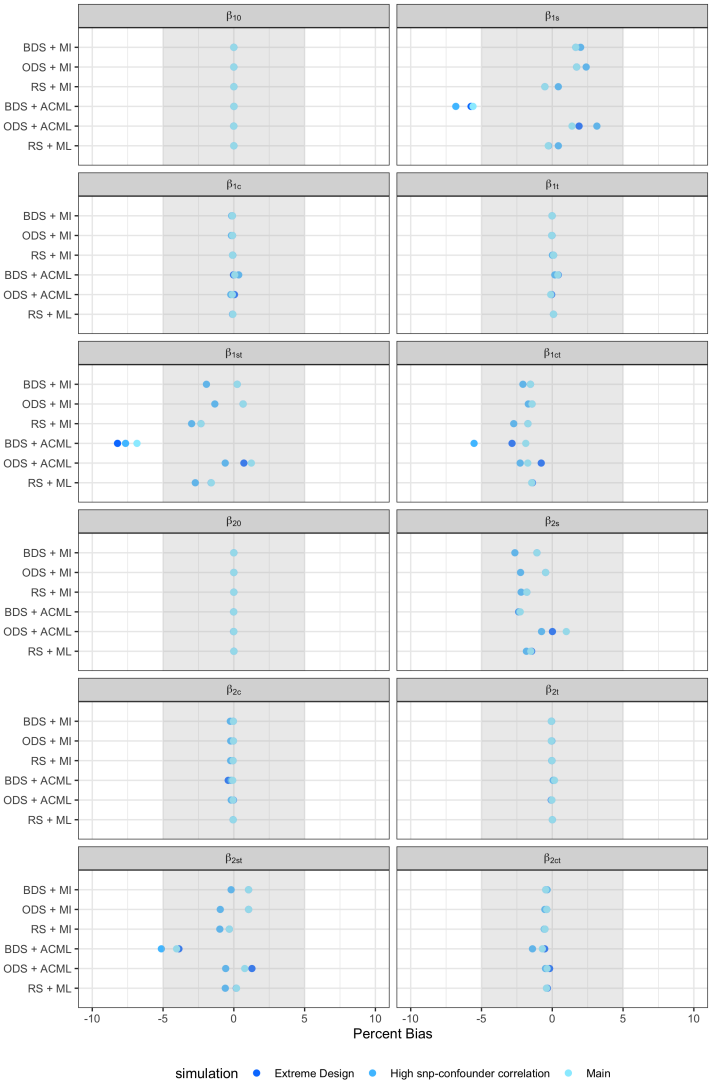
We consider 6 design + inference procedures:

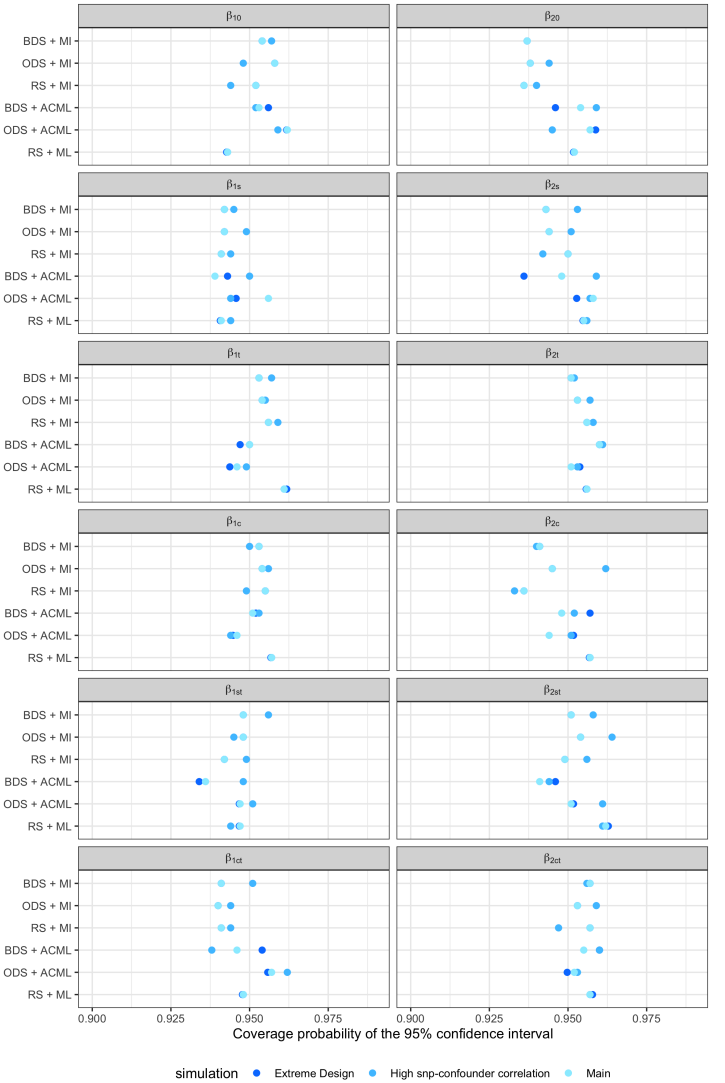
Consider the 800 people sampled in phase two

- RS with maximum likelihood
- ODS slope with ascertainment corrected maximum likelihood
- BDS slope with ascertainment corrected maximum likelihood

Consider everyone regardless of sampling status

- RS with multiple imputation
- ODS slope with multiple imputation
- BDS slope with multiple imputation

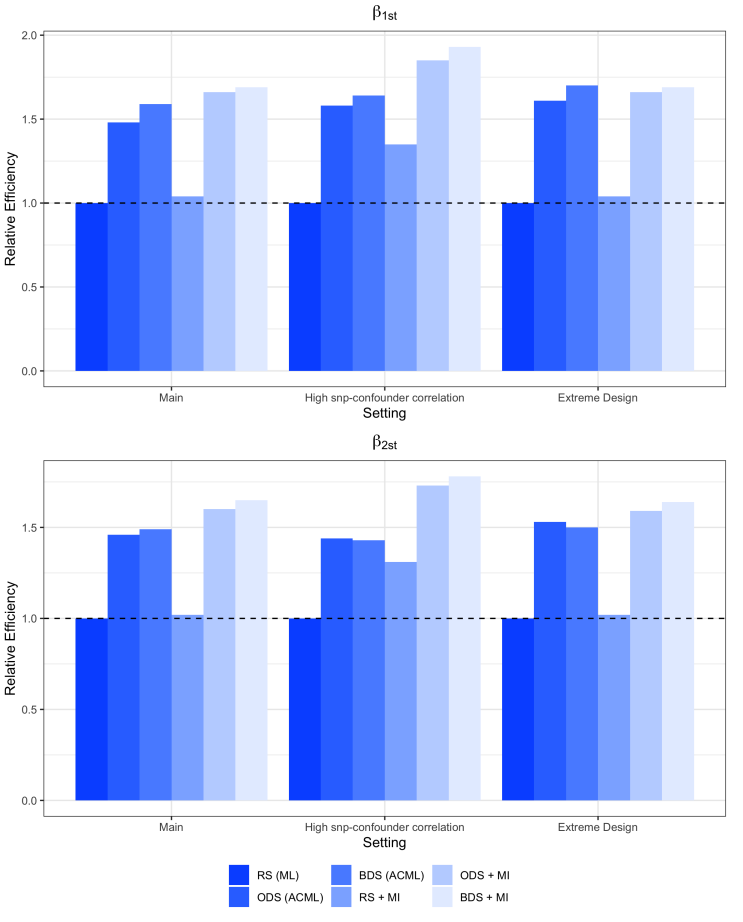

β10
β1s
β1c
β1t
β1st
β1ct
β20
β2s
β2c
β2t
β2st
β2ct
BDS + MI
ODS + MI
RS + MI
BDS + ACML
ODS + ACML
RS + ML
-10
-5
0
5
10
Percent Bias
simulation
Extreme Design
High snp-confounder correlation
Main


β10
β20
β1s
β2s
β1t
β2t
β1c
β2c
β1st
β2st
β1ct
β2ct
BDS + MI
ODS + MI
RS + MI
BDS + ACML
ODS + ACML
RS + ML
0.900
0.925
0.950
0.975
Coverage probability of the 95% confidence interval
simulation
Extreme Design
High snp-confounder correlation
Main

- We compute the relative efficiency for each design + inference procedure:

$$\frac{Var_{\text{RS+ML}}(\hat{\beta})}{Var_{\text{D+I}}(\hat{\beta})}$$

- Efficiency gain is mainly due to the design
- For the imputation approach, efficiency gains were larger when increasing the correlation between c_i and snp_i
- Efficiency gains were larger when sampling was more extreme

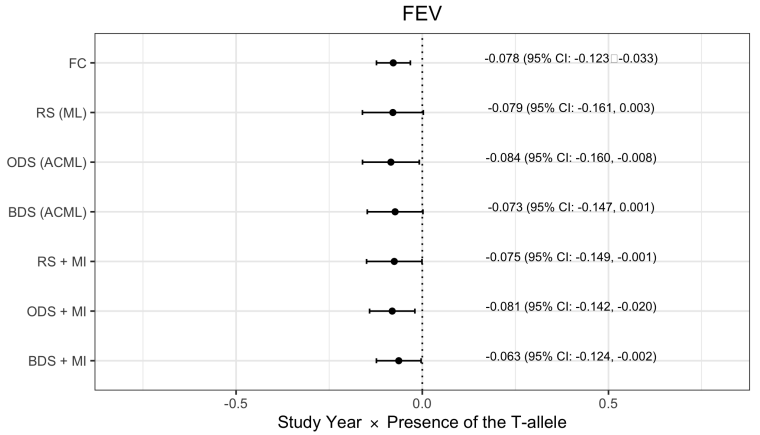


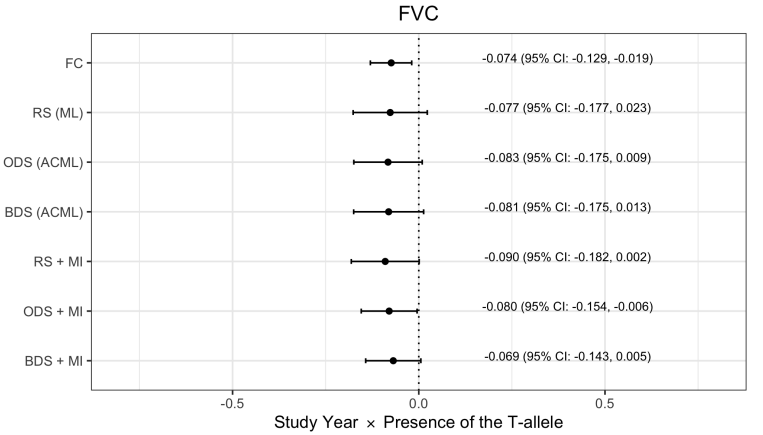
- We re-use data from a previous two-phase ODS design where we sampled 400 subjects (out of 2,400) using ODS slope based on Y_{1ij}
- We are interested in the association between the expensive exposure and Y_{2ij}

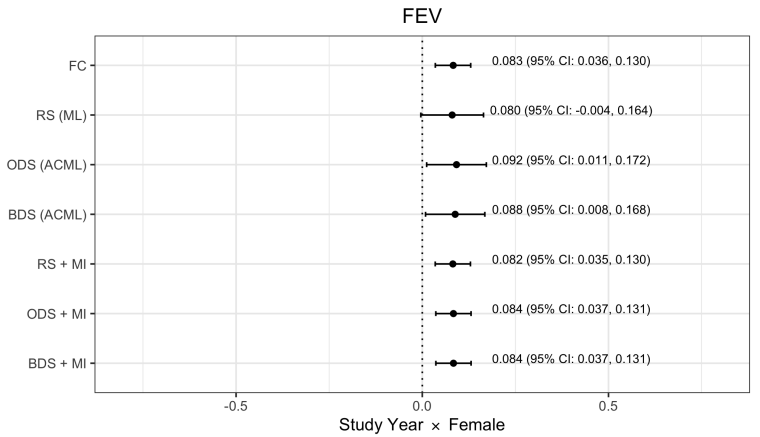
	Y_1						Y_2					
	β_{10}	β_{1s}	β_{1c}	β_{1t}	β_{1st}	β_{1ct}	β_{20}	β_{2s}	β_{2z}	β_{2t}	β_{2st}	β_{2ct}
Linear Mixed Effects Model												
% Bias							−0.016	17.744	11.813	0.625	19.081	17.482
SE							0.189	0.333	0.138	0.037	0.068	0.026
SEE							0.190	0.342	0.139	0.038	0.068	0.028
CP							0.943	0.950	0.597	0.951	0.939	0.180
ODS Slope Sampling with ACML												
% Bias	0.001	−4.042	−0.050	0.133	−0.085	0.214	0.006	3.816	0.187	0.131	−0.016	−0.231
SE	0.268	0.515	0.212	0.055	0.098	0.050	0.186	0.327	0.146	0.035	0.063	0.027
SEE	0.280	0.506	0.215	0.055	0.099	0.050	0.185	0.334	0.143	0.035	0.064	0.028
CP	0.955	0.948	0.956	0.947	0.957	0.944	0.943	0.958	0.945	0.945	0.956	0.962
ODS Slope Sampling with MI												
% Bias	0.010	−5.940	−0.089	−0.021	0.032	−0.071	0.011	3.361	0.000	−0.009	−0.858	−0.127
SE	0.179	0.509	0.094	0.036	0.096	0.024	0.118	0.326	0.067	0.022	0.062	0.012
SEE	0.176	0.487	0.097	0.037	0.095	0.024	0.117	0.321	0.065	0.022	0.061	0.013
CP	0.951	0.939	0.954	0.959	0.951	0.938	0.944	0.949	0.935	0.951	0.943	0.958

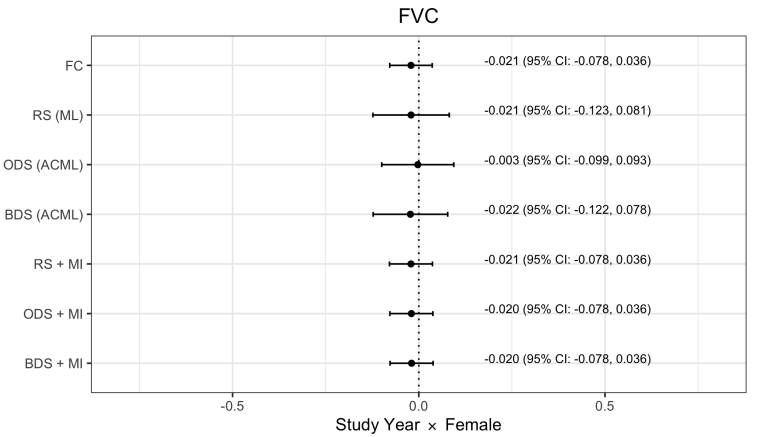
The Lung Health Study

- Subjects were followed-up over a 5 years period (66% of individuals had outcome data measured at each follow-up time)
- 56% of individuals had at least one copy of the T-allele at rs177852
- We sample 800 subjects and examine three design: Random Sampling (RS), ODS Slope and BDS Slope
- For ODS Slope and BDS Slope we sample 400 people based on FEV and 400 people based on FVC


FEV
FC
RS (ML)
ODS (ACML)
BDS (ACML)
RS + MI
ODS + MI
BDS + MI
-0.078 (95% CI: -0.123 -0.033)
-0.079 (95% CI: -0.161, 0.003)
-0.084 (95% CI: -0.160, -0.008)
-0.073 (95% CI: -0.147, 0.001)
-0.075 (95% CI: -0.149, -0.001)
-0.081 (95% CI: -0.142, -0.020)
-0.063 (95% CI: -0.124, -0.002)
-0.5
0.0
0.5
Study Year × Presence of the T-allele


FVC
FC
RS (ML)
ODS (ACML)
BDS (ACML)
RS + MI
ODS + MI
BDS + MI
-0.074 (95% CI: -0.129, -0.019)
-0.077 (95% CI: -0.177, 0.023)
-0.083 (95% CI: -0.175, 0.009)
-0.081 (95% CI: -0.175, 0.013)
-0.090 (95% CI: -0.182, 0.002)
-0.080 (95% CI: -0.154, -0.006)
-0.069 (95% CI: -0.143, 0.005)
-0.5
0.0
0.5
Study Year × Presence of the T-allele


FEV
FC
RS (ML)
ODS (ACML)
BDS (ACML)
RS + MI
ODS + MI
BDS + MI
0.083 (95% CI: 0.036, 0.130)
0.080 (95% CI: -0.004, 0.164)
0.092 (95% CI: 0.011, 0.172)
0.088 (95% CI: 0.008, 0.168)
0.082 (95% CI: 0.035, 0.130)
0.084 (95% CI: 0.037, 0.131)
0.084 (95% CI: 0.037, 0.131)
-0.5
0.0
0.5
Study Year × Female


FVC
FC
RS (ML)
ODS (ACML)
BDS (ACML)
RS + MI
ODS + MI
BDS + MI
-0.021 (95% CI: -0.078, 0.036)
-0.021 (95% CI: -0.123, 0.081)
-0.003 (95% CI: -0.099, 0.093)
-0.022 (95% CI: -0.122, 0.078)
-0.021 (95% CI: -0.078, 0.036)
-0.020 (95% CI: -0.078, 0.036)
-0.020 (95% CI: -0.078, 0.036)
-0.5
0.0
0.5
Study Year × Female

Summary

- We introduced two-phase ODS design for a single outcome and discuss possible study designs and inference procedures
- We discussed extensions of the two-phase ODS to multivariate longitudinal outcome and introduce two methods that can be used to estimate the parameters
- We demonstrated how considering a multivariate outcome allows us to re-use data from a previously conducted two-phase ODS with a single longitudinal outcome
- We examined finite sampling operating characteristics of the proposed approaches and demonstrated how the proposed designs and estimation procedure can be used to examine genetic associations with lung function decline

Thank you!

References

Di Gravio C, Tao R, Schildcrout J. Design and analysis of two-ohase studies with multivariate longitudinal data. *Submitted*

Hansel, N. et al (2013) Genome-wide study identifies two loci associated with lung function decline in mild to moderate COPD. Human Genetics. 132, 79–90.

Lin, D et al (2013). Quantitative trait analysis in sequencing studies under trait-dependent sampling. PNAS. 110, 12247–12252.

Schildcrout, J. et al (2013). Outcome vector dependent sampling with longitudinal continuous response data: stratified sampling based on summary statistics. Biometrics. 69, 405–16.

Schildcrout, J. et al (2020). Two-phase, generalized case-control designs for the study of quantitative longitudinal outcomes. American Journal of Epidemiology. 189(2), 81-90.

Sun, Z. et al (2017). Exposure enriched outcome dependent designs for longitudinal studies of gene-environment interaction. Statistics in Medicine. 36, 2947–2960.

van Buuren, S (2018). Flexible imputation of missing data, second edition. Chapman and Hall/CRC

Zhou, H et al (2002) A semiparametric empirical likelihood method for data from an outcome-dependent sampling scheme with a continuous outcome. Biometrics. 58, 413-421