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

Chiara Di Gravio, Ran Tao, Jonathan Schildcrout

Vanderbilt University, Virtual ENAR 2021

March 16, 2021

Motivation

- In longitudinal studies when exposure ascertainment cost limits the sample size, it is desirable to target a sample of informative individuals
- Different designs have been developed to efficiently select individuals for exposure ascertainment
- Analysis is usually done using only subjects in whom exposure was collected, or combining partial data on subjects not sampled with those sampled using full likelihood approaches or multiple imputation

- We discuss a class of two-phase designs for multivariate longitudinal continuous outcomes
- We introduce an analysis procedure based on an iterative imputation (IIM) approach which could be **more efficient** than conditional likelihood analysis and **easier to implement** than full likelihood approaches

Lung Health Study

- The Lung Health Study (LHS) is a multicenter RCT of smokers with mild chronic obstructive pulmonary disease
- Hansel et al (2013) individuated SNP rs177852 to be a modifier of lung function decline in the LHS. This is our expensive exposure
- We use both forced expiratory volume (FEV) and forced vital capacity (FVC) as measure of lung function. We want to study the relationship between the SNP identified by Hansel et al and rate of lung function decline over time
- We consider a scenario where data on outcomes and confounders are available on 2,563 continous smokers with at least two observations, but data on SNP can only be collected on 800 subjects

The mixed effects model used for our analyses is:

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

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

- 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 continuous baseline covariates
- $(\epsilon_{1ij}, \epsilon_{2ij}) \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ are the error terms independent of the random effects

Design of a Two-Phase Study

- Longitudinal outcome data and basic covariate data $(\boldsymbol{Y_{1i}}, \boldsymbol{Y_{2i}}, \boldsymbol{t_i}, \boldsymbol{c_i})$ are available on everyone, but resource constraints allow us to collect SNP on a third of the subjects
- We want to use available information on $(\boldsymbol{Y_{1i}}, \boldsymbol{Y_{2i}}, \boldsymbol{t_i}, \boldsymbol{c_i})$ to select the most informative individuals
- We extend the outcome dependent sampling (ODS) by Schildcrout et al (2013) and the BLUP dependent sampling (BDS) by Sun et al (2017) to multivariate outcomes.

Univariate ODS vs BDS Design

ODS Design

$E\left[Y_{ij}\right] = q_{0i} + q_{1i}t_{ij}$

- q_{0i} is the subject-specific mean of the outcome at baseline, q_{1i} is the subject-specific rate of change
- Sort values of q_{0i} and/or q_{1i} and introduce cutpoints that define sampling strata from which we sample with different probabilities

BDS Design

$Y_{ij} = \alpha_0 + \alpha_t t_{ij} + \boldsymbol{\alpha_c c_{ij}} + a_{0i} + a_{1i}t_{ij} + \epsilon_{ij}$

- Estimate a_{0i} and a_{1i} using the best linear unbiased predictor (BLUP) estimates
- Sort values of a_{0i} and/or a_{1i} and introduce cutpoints that define sampling strata from which we sample with different probabilities

Univariate ODS vs BDS Design

ODS Design

$E\left[Y_{ij}\right] = q_{0i} + q_{1i}t_{ij}$

- q_{0i} is the subject-specific mean of the outcome at baseline, q_{1i} is the subject-specific rate of change
- Sort values of q_{0i} and/or q_{1i} and introduce cutpoints that define sampling strata from which we sample with different probabilities

BDS Design

$Y_{ij} = \alpha_0 + \alpha_t t_{ij} + \boldsymbol{\alpha_c c_{ij}} + a_{0i} + a_{1i}t_{ij} + \epsilon_{ij}$

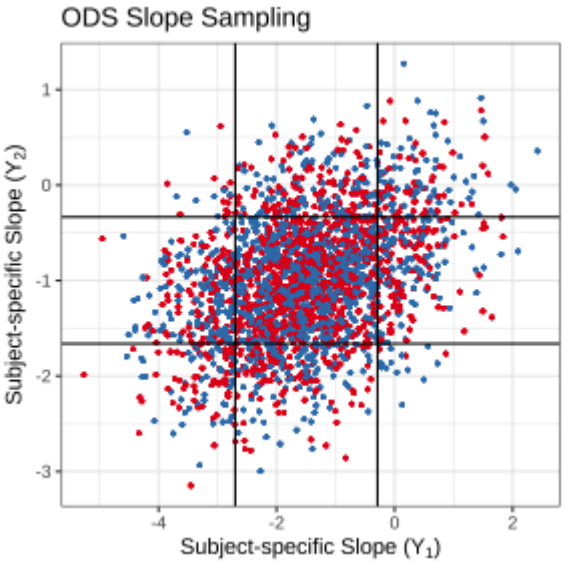
- Estimate a_{0i} and a_{1i} using the best linear unbiased predictor (BLUP) estimates
- Sort values of a_{0i} and/or a_{1i} and introduce cutpoints that define sampling strata from which we sample with different probabilities

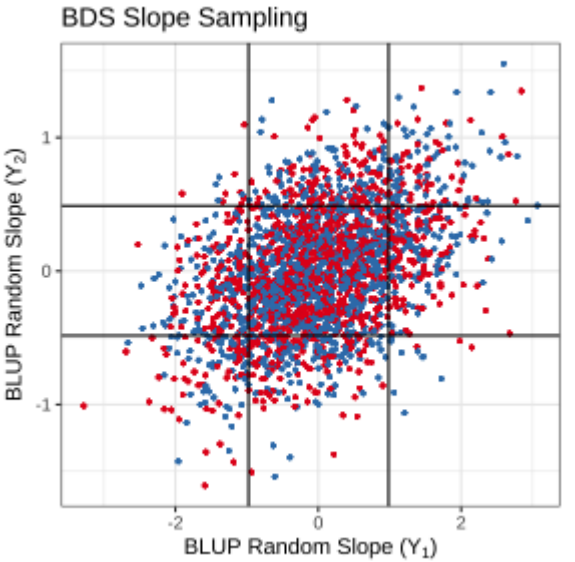
If interest is in the association between Y_{ij} with

- time-fixed covariates we oversample extreme values of q_{0i} or a_{0i}
- time-varying covariates we oversample extreme values of q_{1i} or a_{1i}

Extension to Multivariate Outcomes

- For either ODS or BDS we assign a fraction of individuals to be sampled based on $\boldsymbol{Y_{1i}}$ and the remaining based on $\boldsymbol{Y_{2i}}$
- Since we are interested in a time-varying covariate, we oversample subjects with extreme subject-specific slope or BLUP estimate of the random slope





Analysis of a Two-Phase Study: Iterative Imputation

- ODS and BDS allow us to select the most informative individuals for whom SNP will be collected
- We introduce an iterative imputation approach that uses all available data and impute SNP in individuals not sampled $(S_i = 0)$. The algorithm fills-in missing data by drawing from $[snp_i|\boldsymbol{Y_{1i}}, \boldsymbol{Y_{2i}}, \boldsymbol{t_i}, \boldsymbol{c_i}, S_i = 0]$
- Because sampling depends 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 costruct the imputation model using available data on all subjects. By Bayes' theorem:

$$\frac{pr(snp_i = 1|\boldsymbol{Y_{1i}}, \boldsymbol{Y_{2i}}, \boldsymbol{t_i}, \boldsymbol{c_i}, S_i = 0)}{pr(snp_i = 0|\boldsymbol{Y_{1i}}, \boldsymbol{Y_{2i}}, \boldsymbol{t_i}, \boldsymbol{c_i}, S_i = 0)} = \frac{f(\boldsymbol{Y_{1i}}, \boldsymbol{Y_{2i}}|snp_i = 1, \boldsymbol{t_i}, \boldsymbol{c_i})}{f(\boldsymbol{Y_{1i}}, \boldsymbol{Y_{2i}}|snp_i = 0, \boldsymbol{t_i}, \boldsymbol{c_i})} \times \frac{P(snp_i = 1|\boldsymbol{t_i}, \boldsymbol{c_i})}{P(snp_i = 0|\boldsymbol{t_i}, \boldsymbol{c_i})}$$

- 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 $\boldsymbol{V_i} = Cov(\boldsymbol{Y_i}|snp_i, \boldsymbol{t_i}, \boldsymbol{c_i})$.

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

$$\boldsymbol{Y_i^t V_i^{-1}(\mu_{1,i} - \mu_{0,i})} - \frac{1}{2}(\boldsymbol{\mu_{1,i}^T V_i^{-1} \mu_{1,i} - \mu_{0,i}^T V_i^{-1} \mu_{0,i}}) + log\left[\frac{P(snp_i = 1|\boldsymbol{t_i}, \boldsymbol{c_i})}{P(snp_i = 0|\boldsymbol{t_i}, \boldsymbol{c_i})}\right]$$

- The imputation model is an offsetted logistic regression

Estimate $\boldsymbol{V}^{-1}$ and $\boldsymbol{\mu}_{x,i}$ and the offset iteratively

1) Fit the linear mixed effects model of interest on sampled subjects only

2) Take the estimated parameters and compute the offset

Estimate $\boldsymbol{V}^{-1}$ and $\boldsymbol{\mu}_{x,i}$ and the offset iteratively

1) Fit the linear mixed effects model of interest on sampled subjects only

2) Take the estimated parameters and compute the offset

3) Fit the logistic imputation model using the offset in calculated 2)

4) For unsampled subjects impute SNP using the results of the model in 3)

Estimate $\boldsymbol{V}^{-1}$ and $\boldsymbol{\mu}_{x,i}$ and the offset iteratively

1) Fit the linear mixed effects model of interest on sampled subjects only

2) Take the estimated parameters and compute the offset

3) Fit the logistic imputation model using the offset in calculated 2)

4) For unsampled subjects impute SNP using the results of the model in 3)

5) Fit the linear mixed effects model of interest on everyone

6) Repeat steps 2) to 5), and discard the first few iterations

7) Combine the results using Rubin's rule

The Lung Health Study

- Subjects were followed-up over a 5 years period (66% of individuals had outcome data at every follow-up time)
- 56% of individuals had at least one copy of the T-allele at rs177852
- Median age at baseline was 48 years (interdecile range: 39 - 57 years)
- Even though genetic data are available for all the 2,563 subjects, we consider a scenario where information on the expensive exposure is unavailable and can be measured for a third of the subjects due to financial constraints

- We sample approximately 800 subjects and examine three designs: random sampling (RS), ODS and BDS
- We compare our results with a full cohort analysis (FC)

Outcome	Sampling	Estimate	95% CI
FEV	FC	-0.08	(-0.12, -0.04)
FEV	RS	-0.07	(-0.15, 0.00)
FEV	ODS	-0.08	(-0.14, -0.02)
FEV	BDS	-0.07	(-0.13, -0.01)
--------	----------	----------	----------
FVC	FC	-0.07	(-0.13, -0.02)
FVC	RS	-0.09	(-0.18, 0.00)
FVC	ODS	-0.08	(-0.15, -0.01)
FVC	BDS	-0.06	(-0.13, 0.02)

Summary

- We extended the two-phase design to multivariate longitudinal data, and we introduced different designs
- We developed an iterative imputation algorithm that is easy to implement
- We demonstrated how the proposed designs and estimation procedure can be used to examine genetic associations with lung function

Reference

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.

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

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

Thank you!