Efficient Design and Analysis of a Two-Phase Study with Longitudinal Binary Outcomes

Chiara Di Gravio, Jonathan Schildcrout, Ran Tao

Vanderbilt University

June 28, 2022

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 exposure will be collected

We discuss two classes of study designs for scenarios where we have a binary longitudinal outcome and baseline covariates available on all subjects, and we need to collect information on an exposure

We introduce a semi-parametric likelihood approach to estimate model's parameters

We demonstrate how the designs and estimation procedure can be used to examine genetic association with lung function

The Lung Health Study

The Lung Health Study (LHS) is a multicenter RCT of smokers with mild chronic obstructive pulmonary disease (COPD)

Hansel et al (2013) individuated SNP rs10761570 to be a modifier of lung function decline in the LHS. The SNP is our expensive exposure

We define poor lung function based on FEV and FEV/FVC ratio. We want to study the relationship between the SNP identified by Hansel et al and poor lung function

We consider a scenario where data on outcome and confounders are available on 2,562 individuals, but data on SNP can only be collected on 600 subjects

For our analysis we use a marginalized transition and latent variable model:

$$logit(\mu_{ij}^m) = \beta_0 + \beta_t T_{ij} + \beta_x X_i + \beta_{tx} T_{ij} X_i + \boldsymbol{\beta}_z^T \boldsymbol{Z}_i$$

$$logit(\mu_{ij}^c) = \Delta_{ij} + \gamma Y_{ij-1} + \sigma U_i$$

where:

- Y_{ij-1} is the indicator for poor lung function for subject i at visit $j-1$
- X_i is an indicator for the presence of at least one copy of the allele rs10761570
- $\boldsymbol{Z}_i$ is a set of baseline covariates
- Δ_{ij} links the marginal mean μ_{ij}^m and the conditional mean μ_{ij}^c
- $U_i \sim N(0,1)$

The NSA Designs

Schildcrout et al (2008) introduced a design where informative individuals are sampled based on a summary of the outcome vector.

For each subject, compute $S_i = \sum_{j=1}^{n_i} Y_{ij}$, and classify them in one of the three strata:

- **None Stratum:** People who never experience the outcome $(S_i = 0)$
- **Some Stratum:** People who exhibit response variation $(0 < S_i < n_i)$
- **All Stratum:** People who always experience the outcome $(S_i = n_i)$

Sample from each of the three strata with different probabilities.

The Residual-Based Designs

Including information on the available confounders in the sampling scheme can improve efficiency.

We introduce a class of study designs that identifies the most informative individuals by considering all the available variables.

Step 1 Fit the marginalized transition and latent variable model

$$logit(\mu_{ij}^m) = \beta_0^* + \beta_t^* T_{ij} + \boldsymbol{\beta}_z^{*T} \boldsymbol{Z}_i$$

$$logit(\mu_{ij}^c) = \Delta_{ij}^* + \gamma^* Y_{ij-1} + \sigma^* U_i$$

Step 2 Compute $\hat{\mu}_{ij}^m = expit\left(\hat{\beta}_0^* + \hat{\beta}_t^* T_{ij} + \hat{\boldsymbol{\beta}}_z^{*T} \boldsymbol{Z}_i\right)$ and $\hat{\epsilon}_{ij} = Y_{ij} - \hat{\mu}_{ij}^m$ for each i and j.

Step 3 For each subject i compute a summary of $\hat{\epsilon}_{ij}$. Use this summary to sample informative individuals

The Proposed Method

We introduce a full-likelihood approach that combines partial data on subjects not sampled with complete data on sampled subjects.

Let V be an indicator of whether a subject has the exposure X measured.

$$\underbrace{\sum_{i=1}^{n} V_i \left\{ log P_\beta(\boldsymbol{Y}_i|X_i, \boldsymbol{Z}_i) G(X_i|\boldsymbol{Z}_i) \right\}}_{\text{Contribution of Sampled Subjects}} + \underbrace{\sum_{i=1}^{n} (1 - V_i) \left[log \int_x P_\beta(\boldsymbol{Y}_i|x, \boldsymbol{Z}_i) G(x|\boldsymbol{Z}_i) dx \right]}_{\text{Contribution of Unsampled Subjects}}$$

We estimate $P_\beta(\boldsymbol{Y}_i|X_i, \boldsymbol{Z}_i)$ parametrically using a marginalized transition and latent variable model.

We estimate $G(X_i|\boldsymbol{Z}_i)$ by discrete probability functions $G(x_1|\boldsymbol{Z}), \ldots, G(x_m|\boldsymbol{Z})$. For continuous $\boldsymbol{Z}$ this is challenging, so we use the method of sieves and extend the **Sieve Maximum Likelihood (SMLE)** from Tao et al (2017).

To estimate $G(X|\boldsymbol{Z})$ we use B-spline basis to construct the approximating function. If $B_j^q(\boldsymbol{Z}_i)$ is the jth B-spline of order q then:

$$logG(X_i|\boldsymbol{Z}_i) \approx \sum_{k=1}^{m} I(\boldsymbol{X}_i = x_k) \sum_{j=1}^{s_n} B_j^q(\boldsymbol{Z}_i) logp_{kj}$$

$$G(x_i|\boldsymbol{Z}_i) \approx \sum_{i=1}^{m} I(\boldsymbol{X}_i = x_k) \sum_{j=1}^{s_n} B_j^q(\boldsymbol{Z}_i) p_{kj}$$

- s_n is the total number of functions in the B-spline basis
- p_{kj} is the coefficient associated with the B-spline term $B_j^q(\boldsymbol{Z}_i)$ at $X = x_k$

The Observed Data Log-Likelihood

$$\sum_{i=1}^{n} V_i \left[logP_\beta(\boldsymbol{Y}_i|X_i,\boldsymbol{Z}_i) + \sum_{k=1}^{m}\sum_{k=1}^{m} I(\boldsymbol{X}_i = x_k) \sum_{j=1}^{s_n} B_j^q(\boldsymbol{Z}_i)logp_{kj}\right] +$$

$$\sum_{i=1}^{n}(1-V_i)\left[log\left(\sum_{i=1}^{m} I(\boldsymbol{X}_i = x_k)P_\beta(\boldsymbol{Y}_i|x_k,\boldsymbol{Z}_i)\sum_{j=1}^{s_n} B_j^q(\boldsymbol{Z}_i)p_{kj}\right)\right]$$

Direct maximization of this likelihood is difficult.

We introduce a latent variable $W \in \{1/s_n, \ldots .1\}$ such that the second term can be interpreted as the log-likelihood of $(Y_i, \boldsymbol{Z}_i)$ assuming that the complete data consist of $(Y_i, X_i\boldsymbol{Z}_i, W_i)$ but X_i and W_i are missing.

We estimate the parameters $\boldsymbol{\beta}$ using the EM algorithm.

We estimate $Cov(\boldsymbol{\beta})$ using the profile likelihood method from Murphy et al (2000).

The Lung Health Study

During the follow-up period, 1570 never experienced the outcome, 602 exhibited response variation and 390 always experienced the outcome. Prevalence of the outcome across all times and subjects was 27%

We sample 600 subjects and examine three designs: SRS, NSA[90,420,90], mR[600]

	Full Cohort	**SRS + ML**	**SRS + SMLE**	**NSA[90,420,90] + SMLE**	**mR[600] + SMLE**
SNP	-0.32 (0.20)	-0.36 (0.43)	-0.32 (0.31)	-0.24 (0.28)	-0.35 (0.26)
SNP $\times$ Visit	0.03 (0.05)	0.03 (0.11)	0.02 (0.10)	0.04 (0.07)	0.04 (0.07)
Visit	0.52 (0.05)	0.53 (0.11)	0.52 (0.05)	0.52 (0.05)	0.52 (0.05)
Sex	0.06 (0.16)	0.11 (0.33)	0.07 (0.16)	0.06 (0.16)	0.07 (0.16)
Age (per 10 years)	0.41 (0.12)	0.41 (0.26)	0.40 (0.14)	0.40 (0.12)	0.40 (0.13)
BMI (per 5 kg/m^2)	-0.15 (0.06)	-0.14 (0.13)	-0.15 (0.06)	-0.15 (0.06)	-0.15 (0.06)
Cigarettes (per 20 cigs)	0.12 (0.04)	0.12 (0.08)	0.12 (0.04)	0.12 (0.04)	0.12 (0.04)
SNP $\times$ Age	-0.06 (0.03)	-0.06 (0.06)	-0.06 (0.03)	-0.07 (0.03)	-0.06 (0.03)
SNP $\times$ Sex	0.08 (0.04)	0.07 (0.08)	0.08 (0.04)	0.08 (0.04)	0.08 (0.04)
γ	0.58 (0.16)	0.53 (0.34)	0.56 (0.16)	0.57 (0.16)	0.57 (0.16)
$log(\sigma)$	0.94 (0.06)	0.93 (0.12)	0.94 (0.06)	0.94 (0.06)	0.94 (0.06)

Summary

We discussed two classes of designs for a binary longitudinal outcome, and proposed a semi-parametric approach to estimate the parameters

We demonstrated how the design and estimation procedure can be used to examine genetic associations with lung function

We are planning to extend the designs and methods to a scenario where we have ordinal longitudinal outcomes

Reference

Dempster AP, Laird NM, and Rubin DB (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B, 39, 1-38.

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.

Murphy SA, and van der Vaart AW (2000). On profile likelihood. Journal of the American Statistical Association, 95, 449-465.

Tao R, Zeng D, and Lin D (2017). Efficient semiparametric inference under two-phase sampling with applications to genetic association studies. Journal of the American Statistical Association, 112, 1468-1476.

Schildcrout JS, Heagerty PJ (2007). Marginalized models for moderate to long series of longitudinal binary response data. Biometrics, 63, 322–333

Schildcrout JS, and Heagerty PJ (2008). On outcome-dependent sampling designs for longitudinal binary response data with time-varying covariates. Biostatistics, 9, 735-749.

Thank you!

Appendix: The Latent Variable W

- $W \in \{1/s_n, \ldots, 1\}$
- $B_q^j(\boldsymbol{Z}) = P(W = j/s_n | \boldsymbol{Z})$
- $p_{kj} = P(\boldsymbol{X} = \boldsymbol{x}_k | \boldsymbol{Z}, W = j/s_n) = P(\boldsymbol{X} = \boldsymbol{x}_k | W = j/s_n)$
- $P(\boldsymbol{Y} | \boldsymbol{X}, \boldsymbol{Z}, W) = P(\boldsymbol{Y} | \boldsymbol{X}, \boldsymbol{Z})$

Simulation Study

We generate data from a marginalized transition model

$$logit(\mu_{ij}^m) = \beta_0 + \beta_t T_{ij} + \beta_x X_i + \beta_{tx} T_{ij} X_i + \beta_z Z_i$$

$$logit(\mu_{ij}^c) = \Delta_{ij} + \gamma Y_{ij-1}$$

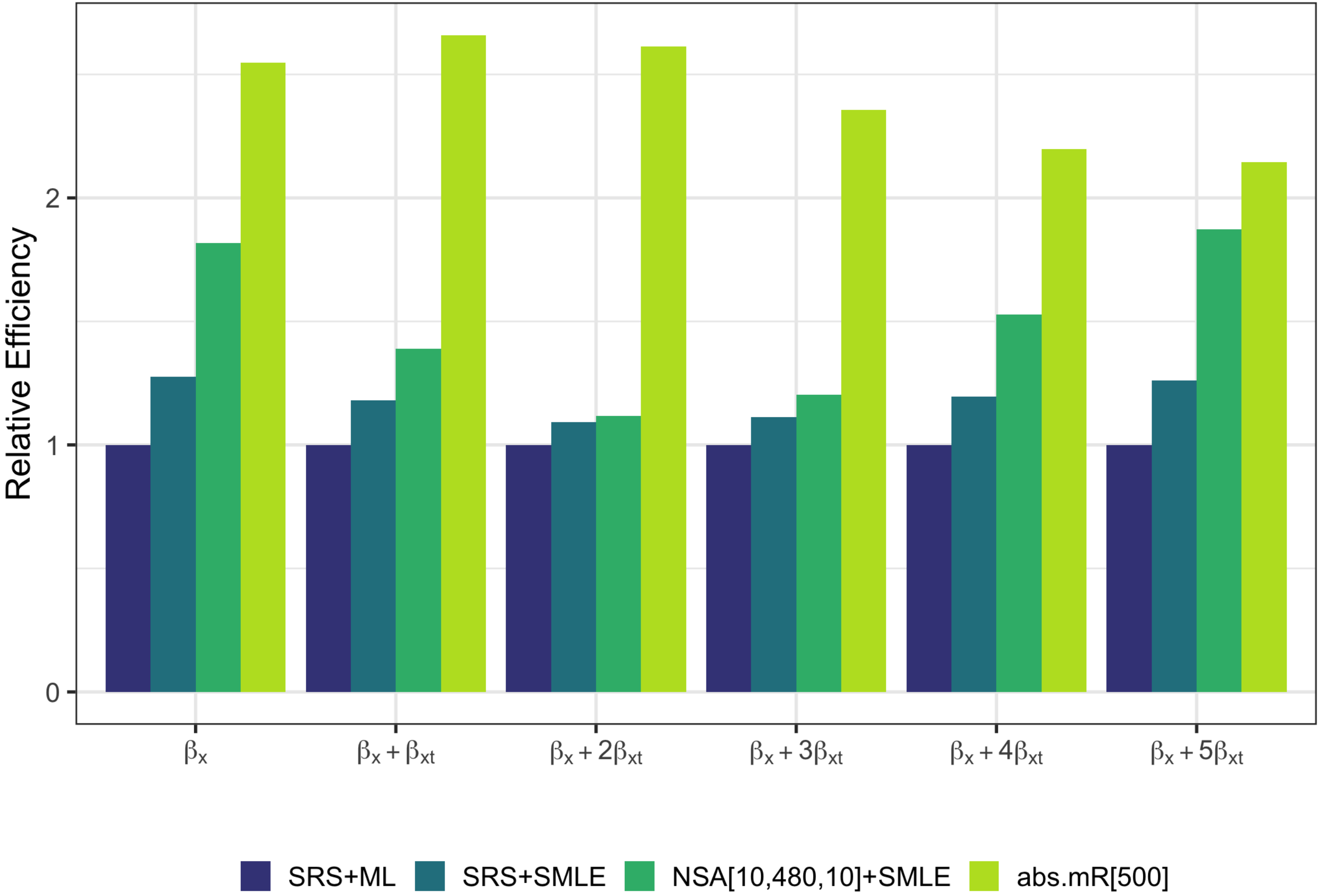
Assuming that:

- each subject i $(i = 1, \ldots, 2000)$ is observed three to six times
- one baseline confounder Z_i such that $P(Z_i = 1) = 0.3$
- one binary expensive covariate X_i such that $logit(P(X_i = 1|Z_i)) = -2.20 + 2Z_i$
- prevalence of the outcome across all times and subjects is 14%

We are interested in estimating $\beta_x + T_{ij}\beta_{xt}$ for $T_{ij} = \{0, 1, \ldots, 5\}$

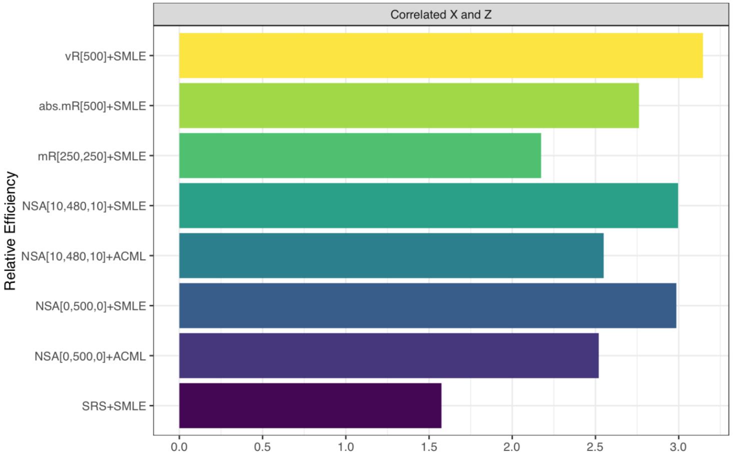
We going to sample 500 people using three different designs

- Estimated coefficients and standard errors were unbiased
- Relative efficiency compared to a simple random sample where model's parameters were estimated using the sampled subject only

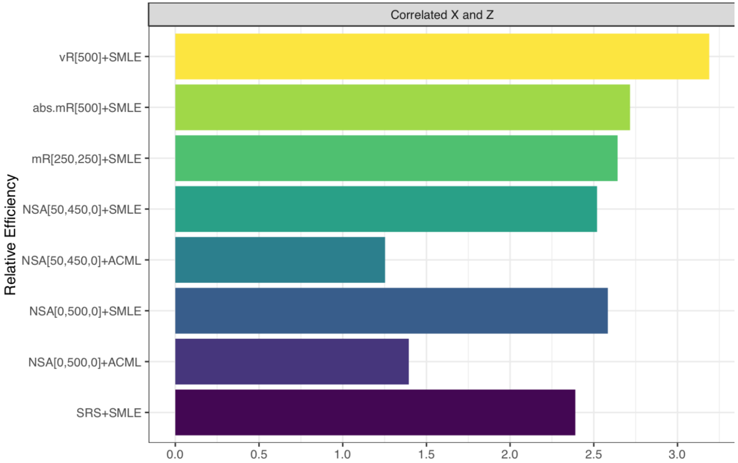


Efficiency for the Coefficients Associated with Time-Varying Covariates

Small Cluster Size



Large Cluster Size


Correlated X and Z
Relative Efficiency
vR[500]+SMLE
abs.mR[500]+SMLE
mR[250,250]+SMLE
NSA[50,450,0]+SMLE
NSA[50,450,0]+ACML
NSA[0,500,0]+SMLE
NSA[0,500,0]+ACML
SRS+SMLE