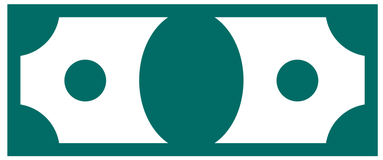
Efficient Study Designs and Analysis Methods for Binary Longitudinal Data: An Application to the Lung Health Study

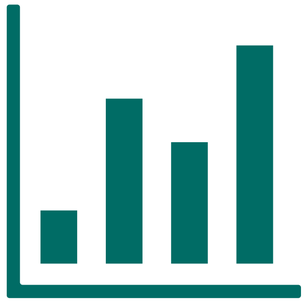
Chiara Di Gravio, Jonathan Schildcrout, Ran Tao
Vanderbilt University
ENAR 2023

Motivation

Electronic health records and clinical trials provide readily accessible data on longitudinal outcome and covariates



Researchers might be interested in an exposure that is unavailable and expensive to collect



We want to use available data to identify the most informative subjects for whom the expensive exposure will be collected

We discuss a class of study designs for settings 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) was a multicenter RCT of smokers with mild chronic obstructive pulmonary disease (COPD)

Hansel et al individuated SNP rs10761570 to be a modifier of lung function decline in the LHS. The presence/absence of the SNP is our expensive exposure

We define poor lung function as forced expiratory volume (FEV) less than 70% and FEV over forced vital capacity (FEV/FVC) less than 0.70. 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 inexpensive covariates 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 (Schildcrout & Heagerty 2007):

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

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

where:

- Y_{ij-1} is the indicator of lung function decline for subject i at time j
- X_i is an indicator of the presence of at least one copy of the allele rs10761570
- $\boldsymbol{Z}_i$ is a set of baseline covariates (age, BMI, sex, cigarettes smoked per year, etc.)
- $U_i \sim N(0,1)$
- Δ_{ij} links the marginal mean $\mu_{ij}^{m} \equiv E[Y_{ij} \mid T_{ij}, X_i, Z_i]$ and the conditional mean $\mu_{ij}^{c} \equiv E[Y_{ij} \mid U_i, Y_{ij-1}]$

Two-Phase Outcome Dependent Sampling (ODS) Designs

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 the observed outcome

Outcome dependent sampling aims to maximize response variability

For a cross-sectional binary outcome:

- the case-control study with the same number of cases and controls is the most efficient design
- efficiency can increase when one further stratifies on phase one inexpensive covariates
- sampling based on the residuals led to efficiency gains (Tao et al 2020)

We want to extend the case-control study to a longitudinal binary outcome while including in the sampling scheme the available information on outcome and covariates

Residual Dependent Sampling (RDS)

Step 1. Fit a marginalized transition and latent variable model that excludes the expensive exposure

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

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

Step 2. Compute the predicted outcome and calculate the residual $\widehat{\epsilon}_{ij} = Y_{ij} - \widehat{\mu_{ij}^{m*}}$

Step 3. Select informative individuals based on a subject-specific summary of $\widehat{\epsilon}_{ij}$

- The choice of summary will depend on the inferential target of interest

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

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

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

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

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

$$logG(X_i|\boldsymbol{Z}_i) \approx \sum_{k=1}^{m} I(X_i = x_k) \sum_{l=1}^{s_n} B_l^q(\boldsymbol{Z}_i) logp_{lj}$$

$$G(x_i|\boldsymbol{Z}_i) \approx \sum_{k=1}^{m} I(X_i = x_k) \sum_{l=1}^{s_n} B_l^q(\boldsymbol{Z}_i) p_{kl}$$

where

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

$$\sum_{i=1}^{n} V_i \left\{ log P_\beta(\mathbf{Y}_i \,|\, X_i, \mathbf{Z}_i) + \sum_{k=1}^{m} \sum_{l=1}^{s_n} I(X_i = x_k) B_l^q(\mathbf{Z}_i) log p_{kl} \right\} +$$

$$\sum_{i=1}^{n} (1 - V_i) \left[log \left(\sum_{k=1}^{m} I(X_i = x_k) P_\beta(\mathbf{Y}_i \,|\, x_k, \mathbf{Z}_i) \sum_{l=1}^{s_n} B_l^q(\mathbf{Z}_i) \right) \right]$$

Direct maximisation of this likelihood is difficult

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

We estimate the parameters β using the EM algorithm

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

The Lung Health Study

2,562 participants in the LHS with at least two follow-up visits were included in the analysis

We sampled 600 subjects and examine two designs: simple random sampling (SRS) and a RDS design where we selected subjects with the highest/lowest mean of the residuals

We are interested in the difference in lung function between those with and without the SNP of interest at visit 1 (the first follow-up time) and visit 5 (the last follow-up time)

We have SNP data on everyone, so we can compare our design/method with the full cohort analysis

Covariates included in the model: SNP, age, sex, BMI, FEV at baseline, time, number of cigarettes smoked, and the interactions of time with age, sex, SNP and FEV at baseline

Estimated log-odds ratio with standard errors from the analysis of the LHS data

	Difference in Lung Function at Visit 1	Difference in Lung Function at Visit 5
Full Cohort	-0.32 (0.20)	-0.15 (0.15)
SRS	-0.28 (0.30)	-0.16 (0.38)
RDS	-0.37 (0.25)	-0.13 (0.21)

Summary

We discussed a class of study designs for a binary longitudinal outcome, and proposed a semiparametric approach to estimate the parameters

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

The RDS design and method presented today are implemented in the R package `dames` available on Github

Reference

Dempster AP, Laird NM, 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, 79-80.

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

Tao R, Zeng D, 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.

Tao R, Zeng D, Lin DY (2020). Optimal designs of two-phase studies. Journal of the American Statistical Association, 115, 1946-1959

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

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