A Brief History of Outcome Dependent Sampling and Two Phase Design

Chiara Di Gravio

November 24, 2020

Contents

1 Introduction

Recent technological advances together with the expansion of electronic health record (EHR) have provided researchers the opportunity to gain an understanding of complex relationships involving both phenotype and genotype. While EHR and existing cohort studies provide readily accessible data on phenotypic variables (e.g, age, sex, race, cardio metabolic profile, anthropometry, etc...), researchers might need to collect additional information on high-dimensional patient characteristics such as DNA sequencing. Frequently, collection of this type of data is expensive and limits the sample size of a study, with negative effects on both the precision and the power. Thus, it becomes important to utilise available data in order to identify the most informative subjects for whom expensive covariates will be ascertained [17, 30].

The two phase outcome dependent sampling (ODS) design is a retrospective study that assigns different probabilities of being sampled to each individual depending on their observed outcome, or their observed outcome-covariates combination. The idea behind a two phase ODS is to use available information to individuate the most informative subjects for a specific research question. Briefly, if we have a cross-sectional discrete outcome, such as disease presence/absence, the case control study with an equal number of cases and controls is the most efficient design [1, 21]; on the other hand, if we have a cross-sectional continuous outcome, such as LDL cholesterol, informative subjects are those with high and low LDL cholesterol values [7]. Finally, if we have longitudinal outcomes, such as LDL cholesterol measured over time, informative subjects can be identified after summarising individuals' longitudinal trajectories.

In every scenario, the two phase ODS can achieve higher efficiency and power than random sampling by targeting informative subjects [2, 6, 17, 19, 25, 31]. However, two phase ODS designs lead to biased samples that need to be accounted for when estimating the parameters. In this report we review existing literature on two phase ODS, and compare likelihood based methods used to estimate the parameters of interest. We start by looking at cross-sectional outcomes in Section 2, then we delve into the two phase ODS for longitudinal outcomes in Section 3. Real data examples and simulation studies are provided throughout the report.

1.1 Examples of Outcome Dependent Sampling and Two Phase Design

1.1.1 The Two Phase Case Control Study

A two phase design for the case control study was introduced by White [27]. Focusing on a rare disease and a rare binary exposure, White proposed to collect information on outcome and exposure for all subjects in a first phase, and to choose a subsample for whom information on expensive covariates would be recorded in a second phase. Specifically, subjects would be categorised into four groups based on disease and exposure status (diseased and exposed, diseased and not exposed, etc.), and the subsample would be chosen by separately sample from each group.

1.1.2 The Norwegian Mother and Child Cohort Study (MoBa)

The Norwegian Mother and Child Cohort Study (MoBa) enrolled over 90,000 pregnant women between 1999 and 2008. To understand the association between time to pregnancy and exposure to perfluorinated and polyfluorinated compounds (PFCs), a two phase ODS design was used in order to oversample women whose time to pregnancy was greater than 12 months (i.e., subfecund). Particularly, among eligible women enrolled between 2003 and 2004, the authors randomly selected 400 subfecund women in addition to 550 women with any time to pregnancy [28].

1.1.3 The National Heart, Lung, and Blood Institute Exome Sequencing Project

The National Heart, Lung, and Blood Institute Exome Sequencing Project (NHLBI ESP) brought together data from multiple cohort studies in order to identify genetic variants associated with heart, lung and blood diseases. Among the studies in the NHLBI ESP, those on body mass index (BMI), LDL cholesterol and blood pressure (BP) were designed such that only subjects with extreme values of each trait were sampled for whole-exome sequencing. For instance, the study of BMI randomly sampled 267 subjects with BMI greater than 40 kg/m^2, and 178 subjects with BMI less than 25 kg/m^2 from 11,468 individuals in the Women's Health Initiative; similarly, studies on LDL cholesterol and BP sampled subjects with the highest and lowest levels of the corresponding trait of interest after adjustments for sex, race and medication [7].

1.2 A Comparison between Two Phase ODS and Random Sampling

Before introducing available methods for conducting inference under two phase ODS sampling, we discuss two examples in order to demonstrate the practical advantages of two phase ODS over simple random sampling. Specifically, we look at how a two phase ODS can increase precision and, consequently, power. Our first example uses data from the third and fourth clinical trials of the National Wilms Tumour Study Group (NWTSG) . The data are part of the `survey` library in R [8], and have been previously used to demonstrate different aspects of a two phase ODS study [2, 3]. In this report, we aim to study the relationship between the odds of relapse after chemotherapy, tumour stage and tumour histology.

The dataset consists of 4,088 children diagnosed with Wilms tumour. Each child has information on relapse status, stage of disease (stage I to IV) and tumour histology (favourable/unfavourable). The latter was measured twice: first, by the pathologist at the patient's institution, then by the pathologist at the NWTSG laboratory. The majority of children had a favourable histology although there were discrepancies between the results from the patient's institution and the NWTSG laboratory. When analysed using logistic regression, patient's institution histology was not related to relapse once NWTSG laboratory histology was taken into account [2]. However, while the measures from the NWTSG laboratory were more accurate and predictive of relapse, they were also more expensive to collect. A two phase ODS design could reduce cost by selecting the most informative subjects for whom a more accurate measure of histology needs to be collected. Table 1 summarises the variables of interest for the entire cohort.

	N (%)
Relapse	571 (14)
Tumour stage	
I	1596 (39)
II	1069 (26)
III	960 (24)
VI	463 (11)
Tumour histology (patient's institution)	
Favourable	3677 (90)
Unfavourable	411 (10)
Tumour histology (NWTSG laboratory)	
Favourable	3623 (89)
Unfavourable	465 (11)

Table 1: Summary statistics of variables of interest for children in the third and fourth clinical trials of the National Wilms Tumour Study Group (NWTSG)

We compared three different analyses. First, we carried out a logistic regression model with stage and NWTSG laboratory histology as covariates using the entire cohort. Afterwards, we simulated a scenario where NWTSG laboratory histology was not available and we implemented two sampling schemes:

1. Two phase ODS. We sampled all cases of relapse, all controls with an unfavourable histology determined by the patient's institution, and a random sample of the remaining children from the eight strata formed by the cross-classification of institutional histology and stage such that cases and controls were the same number. This resulted in 1,142 children.

2. Random sampling. We randomly selected 1,142 children regardless of whether they relapsed.

Table 2 summarises the results from the three analyses. Even though only 28% of the children from the original cohort were sampled in the two phase ODS, the coefficients' standard errors were similar to those estimated in the full cohort. On the other hand, random sampling with sample size equal to two phase ODS sampling, led to higher standard errors. The results clearly show that sampling informative individuals would allow us to gain precision of our estimates while reducing the overall cost of the study.

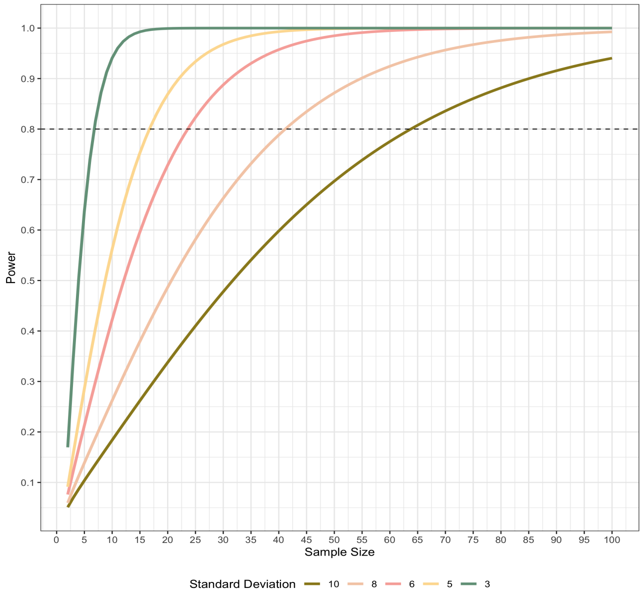
	Full Cohort	Two phase ODS	Random Sampling
Intercept	-2.71 (0.11)	-2.72 (0.11)	-2.58 (0.20)
Stage II	0.77 (0.15)	0.78 (0.15)	0.55 (0.27)
Stage III	0.77 (0.15)	0.80 (0.15)	0.83 (0.28)
Stage VI	1.05 (0.17)	1.07 (0.17)	0.66 (0.36)
Unfavourable Histology	1.31 (0.25)	1.46 (0.32)	1.52 (0.43)
Stage II * Unfavourable Histology	0.15 (0.32)	-0.05 (0.44)	-0.13 (0.59)
Stage III * Unfavourable Histology	0.59 (0.32)	0.28 (0.41)	0.02 (0.57)
Stage IV * Unfavourable Histology	1.26 (0.39)	0.91 (0.63)	1.31 (0.71)
Sample Size	4,088	1,142	1,142

Table 2: Log odds ratios and standard errors from three logistic regression models with relapse after chemotherapy as outcome, tumour stage, NWTSG laboratory histology and interaction between stage and NWTSG laboratory histology as predictors. Analyses for the two phase ODS was done using weighted likelihood in the `survey` package

1.3 Two Phase ODS Designs and Power

From the comparison between a two phase ODS and a simple random sample in Section 1.2 we concluded that, given the same sample size, we were able to gain precision (i.e., smaller standard errors) by carefully selecting informative individuals. Having a more precise estimate has direct consequences on inference: smaller standard errors reduce the width of the confidence intervals, and result in a gain in power. In this Section we discuss the effect of a more precise estimate on the power of a study. To better understand the relationship between precision and power we use a simplified example; however, the results can be extended to more complex scenarios.

We assume that a pharmaceutical company wants to produce a new, cheaper drug aimed to reduce LDL cholesterol by an extra 5 mmHg. Figure 1 shows multiple power curves computed for different levels of precision. As the level of precision decreases (i.e., the standard deviation increases), the number of subjects needed to achieve a pre-determined power level increases. Similarly, given a fixed sample size, the power of a study increases as we gain precision.



Std. Dev	Power	N
10	80%	64
8	80%	42
6	80%	24
5	80%	17
3	80%	7

Figure 1: Power curves computed for different levels of the standard deviation. The significance level for each power curve is set to 5%. The table indicates how many subjects need to be recruited in each group (two groups total) in order to achieve 80% power

1.4 General Framework and Notation

In this section we establish a general framework and notation that will be used throughout the whole report. Specifically, we consider data generated from the joint distribution:

$$f(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z};\boldsymbol{\theta})dG(\boldsymbol{X}|\boldsymbol{Z})dH(\boldsymbol{Z}) \tag{1}$$

where $\boldsymbol{Y}$ is outcome of interest, possibly multivariate, $\boldsymbol{X}$ is the vector of continuous and/or categorical expensive covariates, $\boldsymbol{Z}$ is the vector of continuous and/or categorical inexpensive covariates, $\boldsymbol{\theta}$ is the vector of coefficients linking $\boldsymbol{Y}$ with $\boldsymbol{Z}$ and $\boldsymbol{X}$, $f(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z};\boldsymbol{\theta})$, $dG(\boldsymbol{X}|\boldsymbol{Z})$ is the conditional distribution of $\boldsymbol{X}$ given $\boldsymbol{Z}$, and $dH(\boldsymbol{Z})$ is the distribution of $\boldsymbol{Z}$. Our aim is to make inference on $\boldsymbol{\theta}$.

Under a two phase ODS design, $(\boldsymbol{Y}, \boldsymbol{Z})$ are measured for N individuals in phase one, while $\boldsymbol{X}$ is recorded subsequently for a sub-sample n_V $(n_V < N)$ in phase two. Let R_i be the indicator of whether $\boldsymbol{X}$ is measured for subject i, we represent the index set of all subjects measured in phases one and two with $V = \{i : R_i = 1\}$, and the index set of all subjects measured in phase one only with $\bar{V} = \{i : R_i = 0\}$. The key assumption of a two phase ODS design is that $\boldsymbol{X}$ is missing at random, with subjects in V being sampled based on their observed values $(\boldsymbol{Y}, \boldsymbol{Z})$ only:

$$P(R_i = 1|\boldsymbol{Y_i}, \boldsymbol{X_i}, \boldsymbol{Z_i}) = P(R_i = 1|\boldsymbol{Y_i}, \boldsymbol{Z_i}) = \pi(\boldsymbol{Y_i}, \boldsymbol{Z_i}) \tag{2}$$

When phase two sampling depends on a binary outcome Y and there are not inexpensive covariates $\boldsymbol{Z}$, Prentice and Pyke [12] showed that the distribution function of $\boldsymbol{X}$, $dG(\cdot)$, and the parameter of interest $\boldsymbol{\theta}$ are orthogonal; thus, there is no need to specify the exact distribution of the covariates, and a logistic regression model based on the prospective likelihood in Equation 3 would give us correct estimates for all parameters except the intercept.

$$\prod_{i \in V} f(Y_i|\boldsymbol{X_i}; \boldsymbol{\theta}) \tag{3}$$

However, the result in [12] does not generalise for any other type of outcome [5, 10, 7], and estimators based on Equation 3 are generally biased. In the following sections, we review existing methods for conducting inference on $\boldsymbol{\theta}$ under a two phase ODS design. The methods are presented for cross-sectional and longitudinal outcomes separately, and they are further categorised based on the amount of information they include.

2 Cross-sectional Outcome

In this Section we introduce different methods for estimating $\boldsymbol{\theta}$ when we have a cross-sectional outcome. Specifically, we discuss methods that use data from subjects sampled in phase two only, and methods that account for phase one outcome and covariates data. Additionally, we discuss possible extensions to multivariate cross-sectional outcomes, and look at how we can re-use data from a two phase ODS sampling for a secondary analysis.

2.1 Conducting Inference: Methods Using Phase Two Data Only

The parameter $\boldsymbol{\theta}$ can be estimated by explicitly conditioning on a subject being sampled in phase two. In this case, the likelihood function, so-called *complete data likelihood* [6], can be written as:

$$\begin{aligned} L_{C0}(\boldsymbol{\theta}, G) &= \prod_{i \in V} P(Y_i, \boldsymbol{X_i}|\boldsymbol{Z_i}, R_i = 1) = \prod_{i \in V} \left[\frac{f(Y_i|\boldsymbol{X_i}, \boldsymbol{Z_i}; \boldsymbol{\theta}) dG(\boldsymbol{X_i}|\boldsymbol{Z_i}) \pi(Y_i, \boldsymbol{Z_i})}{P(R_i = 1|\boldsymbol{Z_i}; \boldsymbol{\theta})} \right] \\ &\propto \prod_{i \in V} f(Y_i|\boldsymbol{X_i}, \boldsymbol{Z_i}; \boldsymbol{\theta}) \prod_{i \in V} dG(\boldsymbol{X_i}|\boldsymbol{Z_i}) \prod_{i \in V} \left[P(R_i = 1|\boldsymbol{Z_i}; \boldsymbol{\theta})\right]^{-1} \end{aligned} \tag{4}$$

Maximisation of Equation 4 leads to unbiased estimates of $\boldsymbol{\theta}$. However, the maximisation procedure is not straightforward. First, under a two phase sampling, the scaling factor $P(R_i = 1|\boldsymbol{Z_i}; \boldsymbol{\theta})$ involves $\boldsymbol{\theta}$. Second, even though the marginal distribution $dG(\boldsymbol{X_i}|\boldsymbol{Z_i})$ is unrelated to $\boldsymbol{\theta}$, $dG(\boldsymbol{X_i}|\boldsymbol{Z_i})$ is involved in the scaling factor, and cannot be ignored. Thus, to correctly estimate $\boldsymbol{\theta}$ we would need to either specify the parametric form of $dG(\boldsymbol{X_i}|\boldsymbol{Z_i})$, or use a non parametric method to estimate $dG(\boldsymbol{X_i}|\boldsymbol{Z_i})$. Alternatively, we can remove the dependence on $dG(\boldsymbol{X_i}|\boldsymbol{Z_i})$ by further conditioning on $\boldsymbol{X_i}$:

$$L_{C1}(\boldsymbol{\theta}) = \prod_{i \in V} P(Y_i|\boldsymbol{X_i}, \boldsymbol{Z_i}, R_i = 1) = \prod_{i \in V} \left[\frac{f(Y_i|\boldsymbol{X_i}, \boldsymbol{Z_i}; \boldsymbol{\theta})\pi(Y_i, \boldsymbol{Z_i})}{P(R_i = 1|\boldsymbol{X_i}, \boldsymbol{Z_i}; \boldsymbol{\theta})} \right] \tag{5}$$

It can be shown that if we fix $\boldsymbol{\theta}$ and maximise $L_{C0}(\boldsymbol{\theta}, G)$ with respect of G over the space of all discrete distributions whose support include $\boldsymbol{X}$, we obtain the likelihood $L_{C1}(\boldsymbol{\theta})$ [6]. Hence, $dG(\boldsymbol{X_i}|\boldsymbol{Z_i})$ does not contain any information on the parameter $\boldsymbol{\theta}$, and we would be able to correctly and efficiently estimate $\boldsymbol{\theta}$ without having to specify the form of $dG(\boldsymbol{X_i}|\boldsymbol{Z_i})$.

2.1.1 Weighted Pseudolikelihood

If every variable was observed for all N subjects, the parameter $\boldsymbol{\theta}$ would have been calculated from the likelihood $L_N(\boldsymbol{\theta}) = \prod_{i=1}^{N} f(Y_i|\boldsymbol{X_i}, \boldsymbol{Z_i}; \boldsymbol{\theta})$. The *weighted pseudolikelihood* uses an estimate of $L_N(\boldsymbol{\theta})$ to compute $\boldsymbol{\theta}$. Specifically, the weighted pseudolikelihood estimates $\boldsymbol{\theta}$ by considering subjects observed in phase two only, and weighting their contribution to the likelihood according to the inverse of their selection probability:

$$L_W(\boldsymbol{\theta}) = \prod_{i \in V} \frac{1}{\pi(Y_i, \boldsymbol{Z_i})} f(Y_i|\boldsymbol{X_i}, \boldsymbol{Z_i}; \boldsymbol{\theta}) \tag{6}$$

The weights $\pi(Y_i, \boldsymbol{Z_i})$ can either be the pre-specified selection probabilities or the observed selection probabilities with the latter leading to more efficient estimates. However, regardless of the chosen weights, the weighted pseudolikelihood leads to inefficient estimates when the selection probabilities are highly variable (which is usually the case in a two phase design). Consequently, the method is expected to be less efficient than any other likelihood that considers phase two data only [6].

2.1.2 Semiparametric Empirical Likelihood Estimator

The *semiparametric empirical likelihood estimator* (SELE) was proposed by Zhou et al [31] in order to estimate $\boldsymbol{\theta}$ in settings with continuous Y. Under the assumption of no inexpensive covariates $\boldsymbol{Z}$, Zhou et al suggested to categorise Y in K strata $\mathcal{S}_k$ $(k = 1, ..., K)$ based on known constant $-\infty = a_0 < a_1 < \cdots < a_K = \infty$, and to supplement a simple random sample of n_0 subjects with samples of size n_k taken from each stratum $\mathcal{S}_k$. Under the proposed scheme, the likelihood function could be re-written as the product of two terms representing the contribution of subjects in the simple random sample and in the supplemental sample respectively [30].

$$\begin{aligned}
L(\boldsymbol{\theta}, G) &= \prod_{i=1}^{n_0} f(Y_i|\boldsymbol{X_i}; \boldsymbol{\theta}) dG(\boldsymbol{X_i}) \prod_{k=1}^{K} \prod_{i=1}^{n_k} f(Y_i, \boldsymbol{X_i}|Y_i \in \mathcal{S}_k; \boldsymbol{\theta}) \\
&= \prod_{i=1}^{n_0} f(Y_i|\boldsymbol{X_i}; \boldsymbol{\theta}) dG(\boldsymbol{X_i}) \prod_{k=1}^{K} \prod_{i=1}^{n_k} \mathbf{I}_{\{Y_i \in \mathcal{S}_k\}} \frac{f(Y_i|\boldsymbol{X}_i; \boldsymbol{\theta}) dG(\boldsymbol{X_i})}{P(Y_i \in \mathcal{S}_k)} \\
&= \prod_{k=0}^{K} \prod_{i=1}^{n_k} f(Y_i|\boldsymbol{X_i}; \boldsymbol{\theta}) \prod_{k=0}^{K} \prod_{i=1}^{n_k} dG(\boldsymbol{X_i}) \prod_{k=1}^{K} P(Y_i \in \mathcal{S}_k)^{-n_i}
\end{aligned} \tag{7}$$

Similarly to what observed in the complete data likelihood case, the scaling factor $P(Y_i \in \mathcal{S}_k)$ is a function of $dG(\boldsymbol{X_i})$; thus, the distribution of the expensive covariates $\boldsymbol{X}$ needs to be included in the maximisation of Equation 7. Zhou et al modelled $dG(\boldsymbol{X_i})$ non parametrically by 1) fixing $\boldsymbol{\theta}$ and solving for the empirical likelihood of $d\widehat{G(\boldsymbol{X_i})}$ over all discrete distributions whose support contain $\boldsymbol{X}$ observed in phase two, 2) plugging $d\widehat{G(\boldsymbol{X_i})}$ into Equation 7, and 3) using Newton-Rapson algorithm to estimate $\boldsymbol{\theta}$.

It is important to observe that Equation 7 and the complete data likelihood in Equation 4 share a similar structure: both likelihoods use observed data only, and can be seen as the product of the specified regression model, the distribution of $\boldsymbol{X}$ and a third term that reflects the biased nature of the design. However, in the complete data likelihood this last term is based on the probability of being sampled, whereas in the SELE the last term is computed from the probability of being in stratum $\mathcal{S}_k$.

2.2 Conducting Inference: Methods Using Phase One and Two Data

If data collected during phase one are available on all subjects, including them into the likelihood can lead to efficiency improvement. In this Section we introduce methods that account for unsampled subjects' outcome and covariate data. First, we focus on categorical outcome and inexpensive covariates; we then consider continuous data. All methods are based on likelihood functions comprising of two main factors: the first factor summarises the contribution of subjects who are sampled in phase two, the second factor summarises the contribution of subjects who are not sampled in phase two.

2.2.1 Mean Score Estimator

The *mean score estimator* was introduced by Reilly and Pepe [13] to estimate $\boldsymbol{\theta}$ in scenarios with categorical outcome and inexpensive covariates. Specifically, the method evaluate $\boldsymbol{\theta}$ by solving the estimating equation:

$$\frac{\partial}{\partial\boldsymbol{\theta}}l(\boldsymbol{\theta}) = \sum_{i\in V} S(Y_i|\boldsymbol{X_i},\boldsymbol{Z_i};\boldsymbol{\theta}) + \sum_{i\in\bar{V}} E\left[S(Y_i|\boldsymbol{X},\boldsymbol{Z_i})|Y_i,\boldsymbol{Z_i}\right] = 0 \tag{8}$$

where $S(Y_i|\boldsymbol{X_i},\boldsymbol{Z_i})$ is the score function, and $E\left[S(Y_i|\boldsymbol{X},\boldsymbol{Z_i})|Y_i,\boldsymbol{Z_i}\right]$ is the contribution of subjects non sampled for phase two. Reilly and Pepe suggested to estimate $E\left[S(Y_j|\boldsymbol{X},\boldsymbol{Z_j})|Y_j,\boldsymbol{Z_j}\right]$ using the empirical distribution of $\boldsymbol{X}$ given $\boldsymbol{Z}$ and $\boldsymbol{Y}$, $\hat{P}(\boldsymbol{X}|\boldsymbol{Y},\boldsymbol{Z})$:

$$\hat{E}\left[S(Y_j|\boldsymbol{X},\boldsymbol{Z_j})|Y_j,\boldsymbol{Z_j}\right] = \int_{\mathcal{X}} S(Y|\boldsymbol{X},\boldsymbol{Z})d\hat{P}(\boldsymbol{X}|,\boldsymbol{Y},\boldsymbol{Z}) = \sum_{i\in V_{Z_jY_j}} \frac{S(Y_j|\boldsymbol{X_i},\boldsymbol{Z_j})}{n_{V_{Z_jY_j}}} \tag{9}$$

where $V_{Z_jY_j}$ denotes the set of all subjects sampled in phase two with $\boldsymbol{Z}=\boldsymbol{Z_j}$ and $Y=Y_j$, and $n_{V_{Z_jY_j}}$ is the number of such cases. The mean score estimator is a valid approach since $P(\boldsymbol{X}|\boldsymbol{Z},Y) = P(\boldsymbol{X}|,\boldsymbol{Z},Y,R=1)$; however, it cannot be extended to continuous data and requires the presence of at least one subject in each stratum defined by the combinations of Y and $\boldsymbol{Z}$.

2.2.2 Semiparametric Maximum Likelihood Estimator and Estimated Pseudolikelihood

The *semiparametric maximum likelihood estimator* (SPMLE) was introduced by Lawless et al in order to make inference on $\boldsymbol{\theta}$ when the outcome Y and the inexpensive covariates $\boldsymbol{Z}$ are continuous variables [6]. Particularly, Lawless proposed to discretise $\boldsymbol{Z}$ and Y, classify subjects in a small number of strata $\mathcal{S}_k$ $(k=1,...,K)$ based on $(Y,\boldsymbol{Z})$, and use stratum membership to select subjects for phase two. Under the assumption that only stratum membership is retained for subjects not sampled in phase two, the likelihood is given by:

$$L_F(\boldsymbol{\theta}, G) = \prod_{k=1}^{K} \left[\prod_{i \in V} f(Y_i | \boldsymbol{X_i}, \boldsymbol{Z_i}; \boldsymbol{\theta}) dG(\boldsymbol{X_i} | \boldsymbol{Z_i}) \right] Q_k(\boldsymbol{\theta}, G)^{N_k - n_{V_k}} \tag{10}$$

where N_k is the total number of subjects in stratum $\mathcal{S}_k$, n_{V_k} is the number of subjects in $\mathcal{S}_k$ that are sampled in phase two, and $Q_k(\boldsymbol{\theta}, G) = Pr\{(Y, \boldsymbol{X}, \boldsymbol{Z}) \in \mathcal{S}_k\}$. To estimate $\boldsymbol{\theta}$ Lawless et al used the semiparametric approach discussed in Section 2.1.2 where the distribution of the expensive covariates is modelled non parametrically considering only those values of $\boldsymbol{X}$ observed in phase two, and $\boldsymbol{\theta}$ is subsequently found through a Newton-Rapson approach.

An alternative way to deal with the likelihood function in Equation 10 is given by the *estimated pseudolikelihood*. The estimated pseudolikelihood maximises $L_F(\theta, \tilde{G})$ where $\tilde{G}(\cdot)$ is a consistent empirical estimate of $G(\cdot)$. Lawless et al [6] suggested the use of:

$$\tilde{G}(\boldsymbol{X} | \boldsymbol{Z} = \boldsymbol{z}) = \sum_{k=1}^{K} \hat{G}_k(\boldsymbol{X} | \boldsymbol{Z} = \boldsymbol{z}) \frac{N_k(\boldsymbol{z})}{N(\boldsymbol{z})} \tag{11}$$

where $\hat{G}_k(\boldsymbol{X} | \boldsymbol{Z} = \boldsymbol{z})$ is the empirical conditional distribution function for the $\mathcal{S}_k$ stratum, $N_k(\boldsymbol{z})$ is the number of subjects in stratum $\mathcal{S}_k$ with inexpensive covariate $\boldsymbol{z}$ and $N(\boldsymbol{z})$ is the total number of subjects with inexpensive covariate $\boldsymbol{z}$.

2.2.3 The Pseudoscore Estimator

The *pseudoscore estimator* was introduced by Chatterjee et al [4, 3] in order to make inference on the parameter $\boldsymbol{\theta}$ in scenarios where the outcome Y is continuous, and information on the inexpensive covariates $\boldsymbol{Z}$ is available for all subjects. Under the assumption of categorical inexpensive covariates $\boldsymbol{Z}$, Chatterjee et al re-wrote the full likelihood in Equation 10 as:

$$L(\boldsymbol{\theta}, G) = \prod_{i \in V} f(Y_i | \boldsymbol{X_i}, \boldsymbol{Z_i}; \boldsymbol{\theta}) dG(\boldsymbol{X_i} | \boldsymbol{Z_i}) \prod_{j \in \bar{V}} \int_{\mathcal{X}} f(Y_j | \boldsymbol{x}, \boldsymbol{Z_j}) dG(\boldsymbol{x} | \boldsymbol{Z_j}) \tag{12}$$

where the first and second product terms represent the contribution to the likelihood of subjects sampled and not sampled in phase two respectively. Differently from the SPMLE, the pseudoscore estimator does not require the discretisation of the outcome and includes the observed phase one data in the likelihood. The score function associated with the likelihood in Equation 12 is:

$$\begin{aligned} S(\boldsymbol{\theta}, G) &= \frac{\partial log L(\boldsymbol{\theta}, G)}{\partial \boldsymbol{\theta}} \\ &\propto \sum_{i \in V} \frac{\partial log(f(Y_i | \boldsymbol{X_i}, \boldsymbol{Z_i}; \boldsymbol{\theta}))}{\partial \boldsymbol{\theta}} + \sum_{j \in \bar{V}} \frac{\partial log(\int_{\mathcal{X}} f(Y_j | \boldsymbol{x}, \boldsymbol{Z_j}) dG(\boldsymbol{x} | \boldsymbol{Z_j}))}{\partial \boldsymbol{\theta}} \\ &= \sum_{i \in V} S(Y_i | \boldsymbol{X_i}, \boldsymbol{Z_i}; \boldsymbol{\theta}) + \sum_{j \in \bar{V}} \frac{\int_{\mathcal{X}} S(Y_j | \boldsymbol{x}, \boldsymbol{Z_j}; \boldsymbol{\theta}) f(Y_j | \boldsymbol{x}, \boldsymbol{Z_j}) dG(\boldsymbol{x} | \boldsymbol{Z_j})}{\int_{\mathcal{X}} f(Y_j | \boldsymbol{x}, \boldsymbol{Z_j}) dG(\boldsymbol{x} | \boldsymbol{Z_j})} \end{aligned} \tag{13}$$

The idea behind the pseudoscore estimator is to evaluate $\boldsymbol{\theta}$ by substituting $dG(\boldsymbol{x} | \boldsymbol{Z_j})$ in Equation 13 with a consistent estimate computed based on the conditional distribution $P(\boldsymbol{x} | \boldsymbol{Z}, R = 1)$. Specifically, from Bayes' theorem the cumulative distribution can be written as:

$$dG^*(\boldsymbol{x}|\boldsymbol{Z}) = \frac{dP(\boldsymbol{X} \le \boldsymbol{x}|\boldsymbol{Z}, R=1)P(R=1|\boldsymbol{Z})}{P(R=1|\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{Z})} \tag{14}$$

which can be approximated by its empirical estimate:

$$G_N(\boldsymbol{x}|\boldsymbol{Z}) = \frac{\sum_i I_{[\boldsymbol{X_i} \le \boldsymbol{x}, \boldsymbol{Z_i} = \boldsymbol{Z}, R_i=1]}}{\sum_i I_{[\boldsymbol{Z_i} = \boldsymbol{Z}, R_i=1]}} \tag{15}$$

By substituting Equations 14 and 15 into $S(\boldsymbol{\theta}, G)$, Chatterjee et al obtained the pseudoscore function:

$$\begin{aligned} S_{PS}(\boldsymbol{\theta}; G_N, \pi) &= \sum_{i \in V} S(Y_i|\boldsymbol{X_i}, \boldsymbol{Z_i}; \boldsymbol{\theta}) + \sum_{j \in \bar{V}} \frac{\int_{\mathcal{X}} S(Y_j|\boldsymbol{x}, \boldsymbol{Z_j}; \boldsymbol{\theta}) h_\theta^\pi(Y_j, \boldsymbol{x}, \boldsymbol{Z_j}) dG^*(\boldsymbol{x}|\boldsymbol{Z_j})}{\int_{\mathcal{X}} h_\theta^\pi(Y_j, \boldsymbol{x}, \boldsymbol{Z_j}) dG^*(\boldsymbol{x}|\boldsymbol{Z_j})} \\ &= \sum_{i \in V} S(Y_i|\boldsymbol{X_i}, \boldsymbol{Z_i}; \boldsymbol{\theta}) + \sum_{j \in \bar{V}} \sum_{i \in V} \frac{S(Y_j|\boldsymbol{X_i}, \boldsymbol{Z_j}; \boldsymbol{\theta}) h_\theta^\pi(Y_j, \boldsymbol{X_i}, \boldsymbol{Z_j}) I(\boldsymbol{Z_j} = \boldsymbol{Z_i})}{\sum_{l \in V} h_\theta^\pi(Y_j, \boldsymbol{X_l}, \boldsymbol{Z_j}) I(\boldsymbol{Z_j} = \boldsymbol{Z_l})} \end{aligned} \tag{16}$$

where $h_\theta^\pi(Y_j, \boldsymbol{X_l}, \boldsymbol{Z_j}) = \frac{f(Y_j|\boldsymbol{X_l}, \boldsymbol{Z_j}; \boldsymbol{\theta})}{P(R=1|\boldsymbol{X_l}, \boldsymbol{Z_j})}$.

The parameter $\boldsymbol{\theta}$ is estimated by solving $S_{PS}(\boldsymbol{\theta}; G_N, \pi) = 0$ using the iterated reweighting algorithm [4].

Algorithm 1: Iterated Reweighting Algorithm

```
1 Let $\boldsymbol{\theta}^{(m-1)}$ be the value of the parameter at step $m-1$, then at step $m$:
2 For each subject $j$ not sampled, build the filled-in data $\{(\boldsymbol{Y_j}, \boldsymbol{X_i}, \boldsymbol{Z_j})\}$ using all observed combinations of
    $(\boldsymbol{X_i}, \boldsymbol{Z_i})$ with $\boldsymbol{Z_i} = \boldsymbol{Z_j}$.
3 For each filled-in observation $\{(\boldsymbol{Y_j}, \boldsymbol{X_i}, \boldsymbol{Z_j})\}$ calculate its associated weight:
    $\omega_{ij}\left(\boldsymbol{\theta}^{(m)}\right) = \frac{h^{\hat{\pi}}_{\theta^{(m)}}(\boldsymbol{Y_j}, \boldsymbol{X_i}, \boldsymbol{Z_j})}{\sum_l h^{\hat{\pi}}_{\theta^{(m)}}(\boldsymbol{Y_j}, \boldsymbol{X_l}, \boldsymbol{Z_j})}$
4 Obtain a new estimate $\boldsymbol{\theta}^{(m)}$ by fitting a parametric regression model. Assign weight one to subjects sampled
    in phase two and $\omega_{ij}\left(\boldsymbol{\theta}^{(m)}\right)$ to those not sampled in phase two.
5 Repeat (3) and (4) until convergence.
```

The pseudoscore estimator can be used when some subjects have zero probability of being sampled (i.e., case-only sampling or control-only sampling). Simulation studies presented in [4] showed that the pseudoscore estimator often achieved the same efficiency as the SPMLE, and was always more efficient than the estimated pseudolikelihood. However, as the regression effect increased, the efficiency of the pseudoscore estimator decreased and the SPMLE was substantially more efficient.

The method presented above was subsequently extended by the same author in order to account for the presence of continuous inexpensive covariates $\boldsymbol{Z}$. Specifically, let $\boldsymbol{Z}_c$ and $\boldsymbol{Z}_d$ be the inexpensive continuous and categorical covariates respectively, $K(\cdot)$ be a symmetric kernel function, and h the bandwidth, the pseudoscore estimator in Equation 16 can be re-written as:

$$\begin{aligned} S_{PS}(\boldsymbol{\theta}; G_N, \pi) &= \sum_{i \in V} S(Y_i | \boldsymbol{X_i}, \boldsymbol{Z_i}; \boldsymbol{\theta}) \\ &+ \sum_{j \in \bar{V}} \sum_{i \in V} \frac{S(Y_j | \boldsymbol{X_i}, \boldsymbol{Z_j}; \boldsymbol{\theta}) h_\theta^\pi (Y_j, \boldsymbol{X_i}, \boldsymbol{Z_j}) K\left(\frac{\boldsymbol{Z_{c_i}} - \boldsymbol{Z_{c_j}}}{h}\right) I(\boldsymbol{Z_{d_j}} = \boldsymbol{Z_{d_i}})}{\sum_{l \in V} h_\theta^\pi (Y_j, \boldsymbol{X_l}, \boldsymbol{Z_j}) K\left(\frac{\boldsymbol{Z_{c_l}} - \boldsymbol{Z_{c_j}}}{h}\right) I(\boldsymbol{Z_{d_j}} = \boldsymbol{Z_{d_l}})} \end{aligned} \tag{17}$$

Similarly to the case of categorical inexpensive covariates, the parameter $\boldsymbol{\theta}$ is estimated by solving $S_{PS}(\boldsymbol{\theta}; G_N, \pi) = 0$. While the proposed method can be applied regardless of how many continuous inexpensive covariates are available, the more are the $\boldsymbol{Z}_c$, the higher is the dimension of the kernel $K(\cdot)$. Thus, to simplify the problem, Chatterjee et al suggested to reduce the dimension $\boldsymbol{Z}_c$ before estimating $\boldsymbol{\theta}$ [3].

2.2.4 Maximum Estimated Likelihood Estimator

An alternative method to estimate $\boldsymbol{\theta}$ that accounts for phase one and two data is the *maximum estimated likelihood estimator* (MELE) introduced by Weaver et al [26]. The MELE can be implemented when we have a random sample supplemented by stratified sampling based on Y, as well as in settings where we only perform stratified sampling. Similarly to the pseudoscore estimator, the MELE can accommodate a continuous outcome Y and a set of inexpensive categorical covariates $\boldsymbol{Z}$. However, the MELE cannot be extended to continuous $\boldsymbol{Z}$, and requires all subjects to have a non zero probability of being sampled in phase two.

The MELE substitutes $dG(\boldsymbol{x}|\boldsymbol{Z})$ in Equation 12 with the consistent empirical estimate described in Equation 11, and computes the parameter $\boldsymbol{\theta}$ by maximising the resulting likelihood:

$$\hat{L}(\boldsymbol{\theta}) = \prod_{i \in V} f(Y_i | \boldsymbol{X_i}, \boldsymbol{Z_i}; \boldsymbol{\theta}) \prod_{j \in \bar{V}} \left[\sum_{k=1}^{K} \frac{N_k(\boldsymbol{z})}{N(\boldsymbol{z})(n_k(\boldsymbol{z}) + n_{0,k}(\boldsymbol{z}))} \sum_{i \in V_k} f(Y_j | \boldsymbol{X_i}, \boldsymbol{Z_j}; \boldsymbol{\theta}) \right] \tag{18}$$

where V_k is the index set of all subjects in stratum k sampled in phase two. The MELE can be compared to the estimated pseudolikelihood in [6] since both methods evaluate $\boldsymbol{\theta}$ after replacing $dG(\boldsymbol{X}|\boldsymbol{Z})$ with its empirical estimate. However, the MELE and the estimated pseudolikelihood start from two different likelihood functions: the MELE accounts for the observed phase one data by using the full likelihood in Equation 12, the estimated pseudolikelihood considers only stratum membership for those subjects not sampled by using the likelihood in Equation 10. Simulation studies presented in [26] did not compare the MELE with the estimated pseudolikelihood; however, results showed that, the MELE was more efficient than the SELE but less efficient than the pseudoscore estimator.

2.2.5 Fully Efficient Estimators

The MELE and the pseudoscore estimator are based on approximations of the full likelihood computed using a consistent estimate of $dG(\boldsymbol{X}|\boldsymbol{Z})$. Hence, both methods are not fully efficient. Fully efficient methods that can be implemented when there are no inexpensive covariates were developed by Song et al [22] and Lin et al [7]. Starting from the full likelihood:

$$\prod_{i \in V} f(Y_i | \boldsymbol{X_i}; \boldsymbol{\theta}) dG(\boldsymbol{X_i}) \prod_{j \in \bar{V}} \int_{\mathcal{X}} f(Y_j | \boldsymbol{x}) dG(\boldsymbol{x}) \tag{19}$$

both authors estimated $dG(\boldsymbol{x})$ using the discrete probabilities at each $\boldsymbol{x}$ observed in phase two. However, while Lin et al implemented the EM algorithm to estimate $\boldsymbol{\theta}$, Song et al used a mixed Newton method.

Algorithm 2: EM Algorithm for Maximum Likelihood Estimator with No Inexpensive Covariates

```
1 Let $m$ be the number of distinct $x_j$ observed in the second phase, and $p_j$ the associated probability. Start with
  $\boldsymbol{\theta} = 0$, $p_j = 1/m$, and $\sigma_1 =$ sample variance of y In the E-step, set $\psi_{ij} = I(\boldsymbol{X_i} = \boldsymbol{x_j})$ for sampled subjects.
  For unsampled subjects set:
      $$\psi_{ij} = \frac{f(Y_i|\boldsymbol{X_i};\boldsymbol{\theta})p_j}{\sum_{k=1}^{m} f(Y_i|\boldsymbol{X_k};\boldsymbol{\theta})p_k}$$
2 In the M-step, update the parameter values as follows:
  $\boldsymbol{\theta} = \left(\sum_{i=1}^{n}\sum_{j=1}^{m}\psi_{ij}\boldsymbol{X_j'}\boldsymbol{X_j}\right)^{-1}\left(\sum_{i=1}^{n}Y_i\sum_{j=1}^{m}\psi_{ij}\boldsymbol{X_j}\right)$,   $\sigma_1 = n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}(Y_i - \boldsymbol{\theta'}\boldsymbol{X_j})^2$ and
  $p_j = n^{-1}\sum_{i=1}^{n}\psi_{ij}$
3 Repeat (2) and (3) until convergence.
```

Algorithm 3: Mixed Newton Algorithm for Restricted Maximum Likelihood

```
1 Start with initial estimates for $\boldsymbol{\theta}^{(0)}$ and $\widehat{dG_i(\boldsymbol{X})}^{(0)}$, $i \in V$
2 Insert the estimate in the restricted maximum likelihood estimator of $dG(\boldsymbol{x})$
      $$\widehat{dG_i(\boldsymbol{X})} = n - \sum_{j\in\bar{V}} \frac{f(Y_j|\boldsymbol{X_i};\boldsymbol{\theta})}{\sum_{k\in V}\widehat{dG_k(\boldsymbol{X})}f(Y_j|\boldsymbol{X_k};\boldsymbol{\theta})}$$
  solve the equation iteratively using the fixed point algorithm until it converges to a solution $\widehat{dG_i(\boldsymbol{X})}^{(1)}$
3 Insert $\widehat{dG_i(\boldsymbol{X})}^{(1)}$ into the full likelihood in Equation 19, and Newton's method to update $\boldsymbol{\theta}^{(1)}$
4 Repeat (2) and (3) until convergence.
```

Simulation studies presented in [22] highlighted how by estimating $\boldsymbol{\theta}$ and $dG(\boldsymbol{X})$ simultaneously, the maximum likelihood estimator achieved higher efficiency than the MELE and the pseudoscore estimator. However, if inexpensive covariates are available in phase one, the two methods discussed above incur in efficiency loss as the information on inexpensive covariates for non sampled subjects is not included in the estimation of $\boldsymbol{\theta}$. Moreover, if sampling for phase two is related to any inexpensive covariate, the methods do not reflect the correct sampling mechanism leading to biased results [25].

A fully efficient method that accounts for the presence of categorical and continuous inexpensive covariates and allows phase two sampling to depend on phase one data in any manner, is the *semiparametric maximum likelihood estimator* (SMLE) proposed by Tao et al [25]. Similarly to the methods discussed above, Tao et al started from the full likelihood in Equation 12, and estimated $\boldsymbol{\theta}$ and $dG(\boldsymbol{X}|\boldsymbol{Z})$ simultaneously. For each observed z, the authors evaluated $dG(\boldsymbol{X}|z)$ on the distinct observed values of $\boldsymbol{X}$, $(x_1, ..., x_m)$. To account for continuous variables and the possibility that $\boldsymbol{Z}$ might be of infinite dimension, the approximation of $dG(\boldsymbol{X}|\boldsymbol{Z})$ was carried out using the method of sieves with B-spline basis. The idea behind the use of B-spline was to facilitate the estimation of $dG(\boldsymbol{X}|\boldsymbol{Z})$ by "borrowing information" from a neighbourhood of the observed z. Specifically, multivariate B-spline basis were constructed over the support of $\boldsymbol{Z}$ as:

$$\left\{B^{q}_{\boldsymbol{l}}(\boldsymbol{Z}) = N_{l_1}(Z_1)\cdots N_{l_{d_z}}(Z_{d_z}), \boldsymbol{l} = (l_1, ..., l_{d_z}), l_1, \dots, l_{d_z=-q+1,\dots b_n}\right\} \tag{20}$$

where b_n is the number of knots, q is the order of the B-spline, and Z_i, $i = 1, ..., d_z$, are the inexpensive covariates. By indicating the $s_n = (b_n + q)^{d_z}$ B-spline basis in 20 as $\{B_j(\boldsymbol{Z}), j = 1, ..., (b_n + q)^{d_z}\}$, the full likelihood in Equation 12 can be written as:

$$\begin{aligned} l(\boldsymbol{\theta}, \{p_{kj}\}) = &\sum_{i \in V} \left[log f(Y_i|\boldsymbol{X_i}, \boldsymbol{Z_i}; \boldsymbol{\theta}) + \sum_{k=1}^{m} \sum_{j=1}^{s_n} I(\boldsymbol{X_i} = \boldsymbol{x_k}) B_j^q(\boldsymbol{Z_i}) log(p_{kj}) \right] \\ &+ \sum_{i \in \bar{V}} log \left[\sum_{k=1}^{m} f(Y_i|\boldsymbol{X_k}, \boldsymbol{Z_i}; \boldsymbol{\theta}) \sum_{j=1}^{s_n} B_j^q(\boldsymbol{Z_i}) p_{kj} \right] \end{aligned} \tag{21}$$

where $p_{kj} = s_n \int P(\boldsymbol{X_k}|\boldsymbol{z}) B_j^q(\boldsymbol{z}) d\boldsymbol{z}$.

To calculate the SMLE, Tao et al maximised Equation 21 using the EM algorithm. To be able to implement the algorithm, the authors introduced a discrete latent variable U such that the second term of Equation 21 could be interpreted as a log-likelihood. The resulting estimate for $\boldsymbol{\theta}$ is unbiased, consistent and asymptotically normal.

Algorithm 4: EM Algorithm for Semiparametric Maximum Likelihood

```
1 In the E-step calculate the expectations of $I(\boldsymbol{X_i} = \boldsymbol{x_k}, U_i = j/s_n)$ and $I(\boldsymbol{X_i} = \boldsymbol{x_k})$ given $(Y_i, \boldsymbol{Z_i})$ for a non
  sampled subject as: $\hat{\psi}_{kji} = P(\boldsymbol{X} = x_k, U = j/s_n|\boldsymbol{Z_i})$ and $\hat{q}_{ik} = \sum_{j'=1}^{s_n} \hat{\psi}_{kj'i}$ where
      $P(\boldsymbol{X} = \boldsymbol{x_k}, U = j/s_n|\boldsymbol{Z_i}) = \frac{P(\boldsymbol{X} = \boldsymbol{x_k}, U = j/s_n, Y, \boldsymbol{Z})}{P(Y, \boldsymbol{Z})} = \frac{f(Y|\boldsymbol{x_k}, \boldsymbol{Z}; \boldsymbol{\theta}) B_j^q(\boldsymbol{Z_i}) p_{kj}}{\sum_{k'=1}^{m} f(Y|\boldsymbol{x_k}, \boldsymbol{Z}; \boldsymbol{\theta}) \sum_{j'=1}^{s_n} B_j'^q(\boldsymbol{Z_i}) p_{k'j'}}$
2 In the M-step, update $\boldsymbol{\theta}$ by maximising the weighted sum of the likelihood function for the regression model
  $f(y|\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta})$:
      $\sum_{i \in V} log(f(Y_i|\boldsymbol{X_i}, \boldsymbol{Z_i}; \boldsymbol{\theta})) + \sum_{i \in \bar{V}} \sum_{k=1}^{m} \hat{q}_{ik} log(f(Y_i|\boldsymbol{x_k}, \boldsymbol{Z_i}; \boldsymbol{\theta}))$
  and update $p_{kj}$ by maximising:
      $\sum_{i \in V} \sum_{k=1}^{m} \sum_{j=1}^{s_n} I(\boldsymbol{X_i} = \boldsymbol{x_k}) B_j^q(\boldsymbol{Z_i}) log(p_{kj}) + \sum_{i \in \bar{V}} \sum_{k=1}^{m} \sum_{j=1}^{s_n} \widehat{\psi_{kji}} log(p_{kj})$
  under the constraint that $p_{kj}$ is a probability mass function.
3 Repeat (1) and (2) until convergence.
```

Simulation studies presented in [25] showed that when sampling depended on the outcome only, the SMLE was more efficient than the maximum likelihood estimators introduced by Song et al and Lin et al, with the gain in efficiency due to the inclusion of information on inexpensive covariates measured on subjects not sampled for phase two. Moreover, when sampling was a function of both outcome and inexpensive covariates, the SMLE performed better than the pseudoscore estimator due to the former estimating $dG(\boldsymbol{X}|\boldsymbol{Z})$ together with $\boldsymbol{\theta}$. The SMLE performed well even when two continuous inexpensive covariates were considered; however, similar to the pseudoscore estimator, the SMLE is subjected to the curse of dimensionality: the B-spline basis cannot handle too many continuous variables without becoming unstable and, as the neighbourhood of z becomes sparser, it is harder to borrow information to approximate $dG(\boldsymbol{X}|\boldsymbol{Z})$ precisely. Hence, when many continuous variables are available, the dimension of $\boldsymbol{Z}$ should be reduced before estimating the parameters.

2.3 Methods for Multivariate Outcomes

The methods discussed in Sections 2.1 and 2.2 focused on a single outcome Y. However, many biomedical applications aim to study multiple (potentially correlated) outcomes. For instance, to examine patients' cardiometabolic profile, both blood pressure and LDL cholesterol are of interest. When different outcomes are available, it becomes possible to sample informative individuals sequentially: starting from one outcome, we would pick subjects with high and/or low values for that outcome; then, we would repeat the same procedure for all the outcomes of interest one after another [24]. To account for multivariate quantitative outcomes in the absence of inexpensive covariates, an extension of the maximum likelihood estimators presented in Section 2.2.5 was proposed by Tao et al [24]. The method allows for components of the outcome vector $\boldsymbol{Y_i}$ to be partially missing, and estimated the parameters of interest starting from the following observed data likelihood:

$$\prod_{i \in V} f(\boldsymbol{Y_i^{obs}} | \boldsymbol{X_i}; \boldsymbol{\theta}) dG(\boldsymbol{X_i}) \prod_{j \in \bar{V}} \int_{\mathcal{X}} (\boldsymbol{Y_j^{obs}} | \boldsymbol{x}; \boldsymbol{\theta}) dG(\boldsymbol{x}) \tag{22}$$

where $\boldsymbol{Y_i^{obs}}$ indicates the observed part of $\boldsymbol{Y_i}$. To estimate $\boldsymbol{\theta}$, Tao et al implemented a modified version of Algorithm 2 that considers $\boldsymbol{Y_i}$ to be a vector, and calculate the conditional expectation of the missing components given the observed data in the E-step. Simulation studies presented in [24] showed that the resulting estimator is unbiased, consistent and asymptotically normal.

If inexpensive covariates are present, they can be accounted for by extending the SMLE in Equation 21 to include the vector $\boldsymbol{Y_i}$ instead of the single Y_i. Similarly to the case with no inexpensive covariates, the SMLE can still be implemented when $\boldsymbol{Y_i}$ has missing components by modifying Algorithm 4 to add the calculation of the conditional expectation of the missing components given the observed data in the E-step, and replace the missing $\boldsymbol{Y_i}$ with their conditional expectations in the M-step [25].

2.4 Methods for Re-Using Data from an ODS Two Phase Sampling Scheme

Given the cost associated with collecting information, researchers might want to re-use data from a two phase ODS sampling to study the association between a secondary outcome (a variable not used for sampling) and the expensive covariates. Since data collected under a two phase ODS design are not a random sample from the entire population, analyses of secondary outcomes need to account for the biased nature of the sampling. Ignoring the sampling scheme would yield invalid results unless the secondary outcome is independent of the primary outcome used for sampling.

When both primary and secondary outcomes are continuous, the maximum likelihood estimator and the SMLE introduced by Tao et al [24, 25] can be easily extended to include a secondary outcome by assuming a multivariate regression model. A similar assumption is made by Pan et al who proposed to calculate the parameters of interest using two estimating equation approaches [11]. The first approach accounts for subjects selected for phase two only, and finds the parameter of interest by iteratively solving the estimating equation:

$$\sum_{i \in V} \frac{1}{\hat{\pi}_i} \left(\frac{\boldsymbol{\partial X_i}}{\boldsymbol{\partial \theta}} \right)^T \boldsymbol{\hat{Q}}^{-1} (\boldsymbol{Y_i} - \boldsymbol{X_i \theta}) \tag{23}$$

where $\hat{\pi}_i$ is the observed probability of subject i being sampled in the two phase ODS, and $\boldsymbol{\hat{Q}}$ is the sample covariance matrix derived from the full cohort.

The second approach includes the full cohort and is based on solving:

$$\sum_{i \in V} \frac{1}{\hat{\pi}_i} \left(\frac{\partial \boldsymbol{X_i}}{\partial \boldsymbol{\theta}} \right)^T \hat{\boldsymbol{Q}}^{-1} (\boldsymbol{Y_i} - \boldsymbol{X_i \theta}) + \sum_{i \in \bar{V}} \left(1 - \frac{1}{\boldsymbol{\pi_i}} \right) E_{\mathbf{X}|\mathbf{Y},\boldsymbol{Z}} \left[\left(\frac{\partial \boldsymbol{X_i}}{\partial \boldsymbol{\theta}} \right)^T \hat{\boldsymbol{Q}}^{-1} (\boldsymbol{Y_i} - \boldsymbol{X_i \theta}) \right] \tag{24}$$

Simulation studies presented in [11] showed that both estimators are valid, with the estimator derived from Equation 24 having higher efficiency.

3 Longitudinal Outcome

Two phase ODS designs can be extended to longitudinal binary and continuous outcomes. In the following Sections, we introduce methods for sampling informative individuals and for conducting inference. Additionally, we propose a data augmentation algorithm to estimate the parameters of interest and we discuss how the methods presented can be extended to multivariate longitudinal continuous outcomes. Finally, we discuss possible ways to re-use data that have been previously collected under a two phase ODS design.

3.1 Selecting the Most Informative Individuals

Differently from the cross-sectional case where the outcome Y_i for a subject i is measured once, the outcome in a longitudinal study is repeatedly collected over time. Thus, for one subject i, the outcome is a vector $\boldsymbol{Y_i} = (Y_{i1}, ..., Y_{im_i})$ where m_i is the number of recorded observations. For instance, if we are looking at disease status over time, then $\boldsymbol{Y_i}$ can be represented as a sequence of 0s and 1s depending on whether subject i has the disease at each time point; on the other hand, if we are studying a continuous outcome such as forced expiratory volume (FEV), then $\boldsymbol{Y_i}$ is a sequence of FEV values recorded at each time point. The multiple measures of outcomes and covariates allow researchers to separate changes over time within individuals from changes between individuals. To give an example, if we are interested in looking at FEV, a longitudinal study permits us to understand whether subjects have different FEV levels, as well as what is the subject-specific FEV trend over time. Consequently, selection of the most informative subjects will be highly dependent on whether our interest lies in changes between individuals or changes over time within individuals.

3.1.1 Binary Outcomes

Since individuals have multiple measures of the outcome, the sampling in a two phase ODS design needs to be based on low dimensional summaries of $\boldsymbol{Y_i}$. In settings with a binary outcome, Schildcrout and Heagerty [19, 16] proposed to stratify individuals into three groups: those who did not experience the outcome ($\boldsymbol{Y_i}$ is a vector of all 0s), those who experienced the outcome ($\boldsymbol{Y_i}$ is a vector of all 1s), and those who exhibited response variation ($\boldsymbol{Y_i}$ is a vector of both 0s and 1s). Individuals were sampled with different probabilities depending on their respective group. When the research question focused on expensive time-varying covariates, the authors showed that sampling individuals with response variation resulted in estimates as efficient as the ones from the full cohort analysis [19].

However, when interest lied in both time-variant and time-invariant covariates, individuating the most informative individuals was as not straightforward. In this last scenario, estimates were inefficient if only individuals who exhibited response variation were selected [16]. Hence, Schildcrout and Heagerty suggested to additionally sample those who did not exhibit response variation, and compare multiple sampling schemes before deciding whom to sample. Specifically, the authors considered the expensive covariates $\boldsymbol{X}$ needed to be ascertained as missing, and implemented an EM algorithm to impute the missing $\boldsymbol{X}$ conditional on the observed outcome and inexpensive covariates. By making assumptions on the marginal distribution of $\boldsymbol{X}$ and on the relationship between $\boldsymbol{X}$ and the outcome, the algorithm generated multiple potential datasets that could be used to simulate any two phase ODS design and study its efficiency [16].

3.1.2 Continuous Outcomes

For continuous outcomes, Schildcrout et al [17] proposed to stratify individuals in groups by 1) estimating subject-specific intercept and slope from the simple linear regression analyses $E[Y_{ij}] = q_{0i} + q_{1i}t_{ij}$ where t_{ij} is a time-variant inexpensive covariate (i.e., time), 2) sorting the values of q_{0i} and/or q_{1i} and 3) introducing $K-1$ cut-points to define K sampling strata $\mathcal{S}_k$ $(k = 1, ..., K)$. Depending on whether the expensive variables of interest were time-invariant or time-variant, sampling strata were defined based on q_{0i}, or q_{1i} or (q_{0i}, q_{1i}). Informative subjects would be those in the extreme strata $\mathcal{S}_1$ and $\mathcal{S}_K$; thus, researchers should assign higher sampling probabilities to individuals in the extreme strata.

Simulation studies in [17] showed that the efficiency of the estimated coefficients was improved when sampling on the summary measure related to the inferential target. Hence, when time-variant expensive covariates were of interest, informative individuals were those with extreme values of q_{1i}; on the other hand, if interest lied on expensive time-invariant covariates, informative individuals were those with extreme q_{0i}. Finally, if interest was on a combination of time-invariant and time-variant covariates, then informative individuals could be found either by sampling the extreme of the joint distribution (q_{0i}, q_{1i}) or by assigning a portion of individuals to be sampled based on extreme q_{0i} and the rest to be sampled based on extreme q_{1i} [18]. Figure 2 summarises three possible designs.

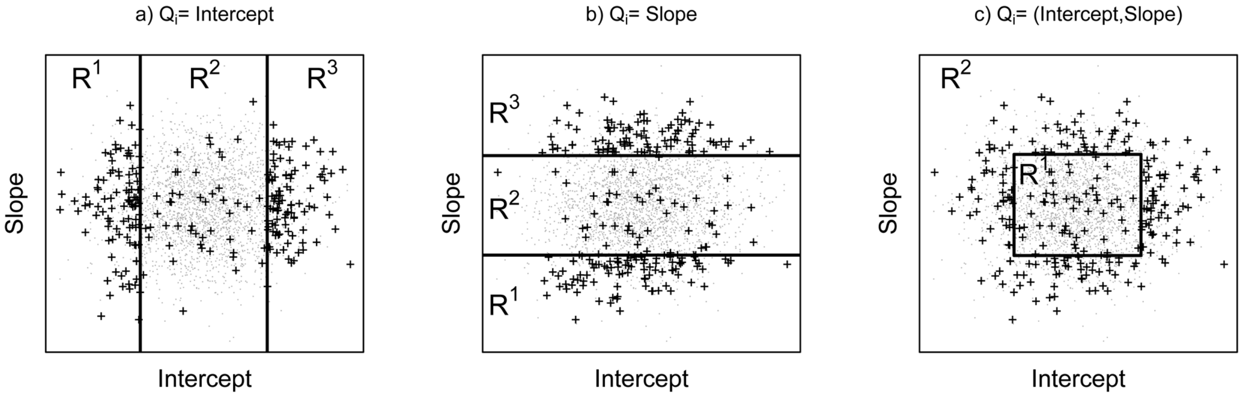


Figure 2: Sampling strata from three different two phase ODS designs. R^i indicates the i^{th} stratum $(i = 1, 2, 3)$. Panel A shows sampling based on three strata of the subject-specific intercepts. Panel B shows sampling based on three strata of the subject-specific slopes. Panel C shows sampling based on two strata of the joint distribution of subject-specific intercepts and slopes. In all panels the black crosses represent subjects that have been sampled, the grey dots denote unsampled subjects. The majority of individuals sampled were in the more extreme strata. The figure was presented by Schildcrout et al in [17], and copied directly in this report for illustration purpose

An alternative sampling scheme was introduced by Sun et al who proposed to use the best linear unbiased predictors (BLUP) from a linear mixed model with random intercept and slope to define sampling strata [23]. Specifically, the authors identified sampling strata by 1) fitting a linear mixed effects model of the outcome $\boldsymbol{Y_i}$ on the available time-variant and time-invariant inexpensive covariates, 2) sorting the estimated random intercepts and/or random slopes and introducing $K-1$ cut-points to define K sampling strata. Similarly to the design proposed by Schildcrout et al, the choice of sampling based on the random intercept and/or the random slope would be influenced by the inferential target. Simulation studies presented in [23] showed that under balanced and complete data (i.e. every subject has complete information on outcome and covariates measured at the same time points), sampling based on BLUPs or based on subject-specific intercepts/slopes resulted in similar efficiency. However, if individuals had missing observations, BLUP based sampling resulted in more efficient estimates as the random effects are expected to better characterise the individual outcome trajectory in the presence of missing data [23].

It is important to note that the designs presented in this section are not the only options; other low-dimensional summaries of $\boldsymbol{Y_i}$ could have been used. However, the sampling strata defined by Schildcrout et al, and by Sun et al allow for an easier representation of the likelihood function used to estimate the parameters of interest. Section 3.2 discusses the inferential procedure in more details.

3.2 Conducting Inference

The methods presented in Sections 2.1 and 2.2 can be extended to longitudinal outcomes [9]. In the following sections we briefly discuss some of the likelihood based approaches that can be used under two phase ODS sampling with longitudinal outcome.

3.2.1 Methods Using Phase Two Data Only

A simple method to estimate $\boldsymbol{\theta}$ is the *ascertainment corrected maximum likelihood* (ACML) procedure introduced by Schildcrout et al [16, 17]. The ACML extends the complete data likelihood presented in Section 2.1 to longitudinal data. Specifically, regardless of whether the outcome of interest is binary or continuous, the likelihood in Equation 5 can be written as:

$$\prod_{i\in V}\left[\frac{\pi(q_i)f(\boldsymbol{Y_i}|\boldsymbol{X_i},\boldsymbol{Z_i};\boldsymbol{\theta})}{P(R_i=1|\boldsymbol{X_i},\boldsymbol{Z_i};\boldsymbol{\theta})}\right]=\prod_{i\in V}\left[\frac{\pi(q_i)f(\boldsymbol{Y_i}|\boldsymbol{X_i},\boldsymbol{Z_i};\boldsymbol{\theta})}{\sum_{k=1}^{K}\pi(\mathcal{S}_k)\int_{\mathcal{S}_k}f(q_i|\boldsymbol{X_i};\boldsymbol{\theta})dq_i}\right] \tag{25}$$

where q_i is the observed value of the low dimensional summary statistic of $\boldsymbol{Y_i}$ used for sampling (e.g., the value of the subject specific slope), $\pi(q_i)$ is the subject i sampling probability based on q_i, and $\pi(\mathcal{S}_k)$ is the sampling probability for stratum $\mathcal{S}_k$. The parameters $\boldsymbol{\theta}$ can be estimated by maximising Equation 25 using a Newton-Rapson approach. Specifically, the algorithm aims to maximise the log-likelihood:

$$\sum_{i\in V}log[f(\boldsymbol{Y_i}|\boldsymbol{X_i},\boldsymbol{Z_i};\theta)]-log\left[\underbrace{\sum_{k=1}^{K}\pi(\mathcal{S}_k)\int_{\mathcal{S}_k}f(q_i|\boldsymbol{X_i},\boldsymbol{Z_i};\boldsymbol{\theta})dq_i}_{AC_i}\right] \tag{26}$$

with AC_i being the ascertainment correction piece. For continuous $\boldsymbol{Y_i}$, AC_i is given by:

$$AC_i = \sum_{k=1}^{K} \pi(\mathcal{S}_k) \left[F_{Q_i|\boldsymbol{X_i},\boldsymbol{Z_i}}(k_k) - F_{Q_i|\boldsymbol{X_i},\boldsymbol{Z_i}}(k_{k-1})\right] \tag{27}$$

where $F_{Q_i|\boldsymbol{X_i},\boldsymbol{Z_i}}(\cdot)$ is the cumulative distribution function of the summary used for sampling informative individuals. Under subject-specific intercept and/or slope sampling and under BLUP sampling, $Q_i|\boldsymbol{X_i}, \boldsymbol{Z_i}$ has a normal distribution for which the functional form is known. On the other hand, when $\boldsymbol{Y_i}$ is binary and sampling is based on whether subjects experience the outcome throughout the study, AC_i simplifies to:

$$AC_i = [\pi(0) - \pi\{(0, m_i)\}]L_{i0} + [\pi(m_i) - \pi\{(0, m_i)\}]L_{im_i} + \pi\{(0, m_i)\} \tag{28}$$

where $\pi(0)$, and $\pi(m_i)$ $\pi\{(0, m_i)\}$ represent the sampling probabilities for subjects who never experienced the outcome, always experienced the outcome, and exhibited response variation respectively. L_{i0} and L_{im_i} correspond to individual i's contribution to the likelihood if simple random sampling was done for $\sum_{j=1}^{m_i} Y_{ij} = 0$ and for $\sum_{j=1}^{m_i} Y_{ij} = m_i$ respectively.

Alternatively to the ACML, the parameters of interest can be estimated by considering the joint distribution of the outcome and the expensive covariate, and adding information from the distribution of the expensive covariates $dG(\boldsymbol{X}|\boldsymbol{Z})$. Zelnick et al showed that under balanced and complete data with no inexpensive covariates, the marginal distribution of the $\boldsymbol{X}$ does not provide any additional information about the parameter $\boldsymbol{\theta}$. However, for not balanced data, more efficient estimates could be obtained if $dG(\boldsymbol{X})$ was included in the maximisation procedure [29].

3.2.2 Methods Using Phase One and Two Data

Similarly to the scenario with a cross-sectional outcome, we would expect efficiency improvement if phase one data measured for all subjects were included in the estimation procedure. A possible way to estimate $\boldsymbol{\theta}$ is to consider only the distributions of the outcome $\boldsymbol{Y}$ conditioned on whether subject i was sampled or not for phase two:

$$\prod_{i \in V} f(\boldsymbol{Y_i}|\boldsymbol{X_i}, \boldsymbol{Z_i}, R_i = 1; \boldsymbol{\theta}) \prod_{j \in \bar{V}} f(\boldsymbol{Y_j}|\boldsymbol{Z_j}, R_j = 0; \boldsymbol{\theta}, \boldsymbol{\gamma}) \tag{29}$$

where $\boldsymbol{\gamma}$ represents the vector of parameters indexing the marginal distribution of the expensive covariate. To maximise Equation 29, the estimation of $\boldsymbol{\gamma}$ is required. While $\boldsymbol{\gamma}$ can be theoretically estimated, its validity depends on the parametric model hypothesised for the covariates. Similarly to what we discussed for the pseudoscore estimator and for the MELE, we can substitute $\boldsymbol{\gamma}$ with a consistent estimator [29].

Alternatively, one can include data on unsampled subjects and consider covariates information explicitly by analysing the full conditional likelihood:

$$\prod_{i \in V} \left[\frac{\pi(\boldsymbol{Y_i}, \boldsymbol{Z_i}) f(\boldsymbol{Y_i}|\boldsymbol{X_i}, \boldsymbol{Z_i}; \boldsymbol{\theta}) dG(\boldsymbol{X_i}|\boldsymbol{Z_i}; \gamma)}{P(R_i = 1|\boldsymbol{Z_i}; \boldsymbol{\theta}, \gamma)}\right] \prod_{j \in \bar{V}} \left[\frac{[1 - \pi(\boldsymbol{Y_j}, \boldsymbol{Z_j})] f(\boldsymbol{Y_j}|\boldsymbol{Z_j}; \boldsymbol{\theta}, \gamma) dG(\boldsymbol{X_j}|\boldsymbol{Z_j}; \gamma)}{1 - P(R_j = 1|\boldsymbol{Z_j}; \boldsymbol{\theta}, \gamma)}\right] \tag{30}$$

Simulation studies presented in [29] showed that when including phase one covariates for all subjects, full information for the variance components was recovered; however, only a minimal gain in efficiency was observed for parameters related to the expensive covariates. For these parameters, the greater efficiency gain was due to the choice of design rather than the inclusion of information on unsampled subjects.

3.3 Multiple Imputation in Two Phase ODS with Continuous Outcome

The two phase ODS sampling schemes described in Section 3.1 depend on a low dimensional summary of $\boldsymbol{Y}$, and possibly on observed inexpensive covariates $\boldsymbol{Z}$. Thus, by knowing $\boldsymbol{Y}$ and $\boldsymbol{Z}$, we have the information necessary to understand why expensive covariates $\boldsymbol{X}$ are not measured in the unsampled subjects:

$$f(\boldsymbol{X_i}|\boldsymbol{Z_i}, \mathbf{Y_i}, R_i = 0) = f(\boldsymbol{X_i}|\boldsymbol{Z_i}, \mathbf{Y_i}) = f(\boldsymbol{X_i}|\boldsymbol{Z_i}, \mathbf{Y_i}, R_i = 1) \tag{31}$$

Since the missing data mechanism in these designs is ignorable, we can multiply impute $\boldsymbol{X}$ for the unsampled subjects using the available data without accounting for having a biased sample [20]. First, data on all individuals are used to construct a conditional exposure model for the unsampled individuals $f(\boldsymbol{X_i}|\mathbf{Y_i}, \boldsymbol{Z_i}, R_i = 0)$; second, missing observations are filled-in by sampling from the constructed conditional exposure model. The process of filling-in missing data is repeated independently M times to create M datasets whose values are identical to each other with the exception of the imputed values. In each dataset the target model is fit using standard maximum likelihood and inference is made after pooling together the resulting M vectors of parameters using Rubin's rule [14]. For instance, for a single parameter θ in the vector $\boldsymbol{\theta}$, the estimated value and the corresponding variance are computed as:

$$\begin{aligned}\hat{\theta} &= \frac{1}{M}\sum_{m=1}^{M}\hat{\theta}^{(m)}\\ \widehat{Var}(\hat{\theta}) &= \frac{1}{M}\sum_{m=1}^{M}\widehat{Var}\left(\hat{\theta}^{(m)}\right) + \left(1+\frac{1}{M}\right)\left[\frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{\theta}^{(m)}-\hat{\theta}\right)^2\right]\end{aligned} \tag{32}$$

where $\hat{\theta}^{(m)}$ is the parameter estimated in the m^{th} dataset.

Consequently, as an alternative to ACML, Schildcrout et al proposed to estimate the parameters $\boldsymbol{\theta}$ using multiple imputation [20]. Focusing on a binary expensive covariate X, the authors introduced two different approaches to estimate the conditional exposure model $f(X_i|\mathbf{Y_i}, \boldsymbol{Z_i}, R_i = 0)$, and fill-in the missing data. The first method, called the *indirect approach*, combined the ascertainment corrected likelihood in Equation 25 with an exposure model $P(X_i|\boldsymbol{Z_i}, R_i = 1)$. The second method, called the *direct approach*, implemented a logistic regression model using data on outcome and inexpensive covariates. Both approaches started from re-expressing $f(X_i|\mathbf{Y_i}, \boldsymbol{Z_i}, R_i = 0)$ using Bayes' theorem:

$$\begin{aligned}\frac{P(X_i = 1|\boldsymbol{Z_i}, \mathbf{Y_i}, R_i = 0)}{P(X_i = 0|\boldsymbol{Z_i}, \mathbf{Y_i}, R_i = 0)} &= \frac{f(\mathbf{Y_i}|X_i = 1, \boldsymbol{Z_i}, R_i = 1; \boldsymbol{\theta})}{f(\mathbf{Y_i}|X_i = 0, \boldsymbol{Z_i}, R_i = 1; \boldsymbol{\theta})} \times \frac{P(X_i = 1|\boldsymbol{Z_i}, R_i = 1)}{P(X_i = 0|\boldsymbol{Z_i}, R_i = 1)}\\ &= \underbrace{\frac{f(\mathbf{Y_i}|X_i = 1, \boldsymbol{Z_i}, R_i = 1; \boldsymbol{\theta})}{f(\mathbf{Y_i}|X_i = 0, \boldsymbol{Z_i}, R_i = 1; \boldsymbol{\theta})}}_{(a)} \times \underbrace{\left[\frac{P(R_i = 1|X_i = 1, \boldsymbol{Z_i})}{P(R_i = 1|X_i = 0, \boldsymbol{Z_i})} \times \frac{P(X_i = 1|\boldsymbol{Z_i})}{P(X_i = 0|\boldsymbol{Z_i})}\right]}_{(b)}\end{aligned} \tag{33}$$

The indirect approach calculated $f(X_i|\mathbf{Y_i}, \boldsymbol{Z_i}, R_i = 0)$ by separately computing terms (a) and (b) in Equation

33, and estimated the parameters of interest in the following steps:

Algorithm 5: Indirect Approach for Missing Data Imputation

```
1 On sampled subjects maximise the ascertainment corrected likelihood in Equation 25 to obtain $\hat{\boldsymbol{\theta}}$ and $\widehat{Cov}\left(\hat{\boldsymbol{\theta}}\right)$
2 Draw $\boldsymbol{\theta}^{(m)}$ $(m = 1, ..., M)$ from the approximate posterior distribution of $\hat{\boldsymbol{\theta}}$ given by the normalised likelihood function and calculate:
  2a)
      $f(\boldsymbol{Y_i}|X_i = 1, \boldsymbol{Z_i}, R_i = 1; \boldsymbol{\theta}^{(m)}) \left[f(\boldsymbol{Y_i}|X_i = 0, \boldsymbol{Z_i}, R_i = 1; \boldsymbol{\theta}^{(m)})\right]^{-1}$
  2b)
      $log\left[P(R_i = 1|X_i = 1, \boldsymbol{Z_i}; \boldsymbol{\theta}^{(m)})\right] \left\{log\left[P(R_i = 1|X_i = 0, \boldsymbol{Z_i}; \boldsymbol{\theta}^{(m)})\right]\right\}^{-1}$
3 On sampled subjects, fit an offsetted logistic regression of $X_i$ on $\boldsymbol{Z_i}$ with the quantity estimated in 2b) as the offset. Obtain the parameters $\hat{\boldsymbol{\gamma}}$ from the model, and their estimated uncertainty $\widehat{Cov}(\hat{\boldsymbol{\gamma}})$. For $m = 1, ..., M$ draw $\boldsymbol{\gamma}^{(m)}$ from a normal distribution with mean $\hat{\boldsymbol{\gamma}}$ and variance $\widehat{Cov}(\hat{\boldsymbol{\gamma}})$. Calculate:
  3a)
      $P(X_i = 1|\boldsymbol{Z_i}, R_i = 1; \boldsymbol{\gamma^{(m)}}) \left[P(X_i = 0|\boldsymbol{Z_i}, R_i = 1; \boldsymbol{\gamma^{(m)}})\right]^{-1}$
4 For unsampled subjects, multiply 2a) and 3a) to calculate the conditional exposure odds. Draw imputed values for the missing expensive covariate
5 Conduct standard maximum likelihood analysis on the response model using the imputed dataset
6 Repeat steps (2) to (5) $M$ times and combine the results using Rubin's rule
```

On the other hand, the direct approach estimated $f(X_i|\boldsymbol{Y_i}, \boldsymbol{Z_i}, R_i = 0)$ by re-expressing Equation 33 in terms of a logistic regression model, and estimated the parameters of interest in the following steps:

Algorithm 6: Direct Approach for Missing Data Imputation

```
1 On sampled subjects, fit the logistic regression of $X_i$ on a combination of observed outcome and inexpensive covariates
2 Obtain the parameters $\hat{\boldsymbol{\gamma}}$ from the model, and their estimated uncertainty $\widehat{Cov}(\hat{\boldsymbol{\gamma}})$. For $m = 1, ..., M$ draw $\boldsymbol{\gamma}^{(m)}$ from a normal distribution with mean $\hat{\boldsymbol{\gamma}}$ and variance $\widehat{Cov}(\hat{\boldsymbol{\gamma}})$
3 For each subject, use $\boldsymbol{\gamma}^{(m)}$ to calculate the predicted probability $\hat{p}_i^{(m)}$ from the logistic model in (1)
4 For unsampled subjects, impute the missing expensive covariate by sampling from a Bernoulli distribution with probability $\hat{p}_i^{(m)}$
5 Conduct standard maximum likelihood analysis on the response model using the imputed dataset
6 Repeat steps (2) to (5) $M$ times and combine the results using Rubin's rule
```

Both the indirect and the direct approach require careful consideration of the imputation model. However, while the direct approach specifies the full conditional exposure model $f(X_i|\boldsymbol{Z_i}, \boldsymbol{Y_i})$, the indirect approach is based on specification of the marginal exposure model $f(X_i|\boldsymbol{Z_i})$. Thus, finding the correct imputation model under the direct approach might be more involved as the time-variant outcome $\boldsymbol{Y_i}$ is included in the imputation of a time-invariant covariate X_i. However, the direct approach has advantages worth considering when deciding which method to use. First, the direct approach can be used for all two phase ODS designs without accounting for the sampling scheme since the method does not use the ascertainment corrected likelihood to estimate the parameters. Second, the direct approach can be more robust to outcome model misspecification since the imputation and the outcome models are separated. Simulation studies in [18, 20] showed that direct and indirect imputation performed similarly. When compared to methods that use complete data from subjects sampled in phase two only, multiple imputation resulted in substantial efficiency improvement of the coefficients associated with inexpensive covariates. For the coefficients associated with expensive covariates, both approaches had less impact; indeed, for those variables, the majority of efficiency gain came from the two phase ODS design rather than the imputation itself.

3.3.1 Weighted Imputation: An Example of the Direct Approach

To give an example of how the direct approach can be used in practice, we present a simulation study where we assume that outcome and inexpensive covariates are available on all individuals but resource constraints permit to collect the expensive covariate in 20% of the subjects. Specifically, suppose we are interested in understanding whether a single nucleotide polymorphisms (SNP) is associated with $FEV\%$. In this scenario, we can consider the outcome Y_{ij} to be $FEV\%$ for subject i at time j, the expensive covariate x_i to be the SNP binary variable for subject i, and the inexpensive covariates t_{ij} and c_i to be the visit time and a baseline continuous covariate for subject i. The model of interest can be written as:

$$Y_{ij} = \theta_0 + \theta_t t_{ij} + \theta_x x_i + \theta_{xt} x_i t_{ij} + \theta_c c_i + b_{0i} + b_{1i} t_{ij} + \epsilon_{ij} \tag{34}$$

where $i = 1, ..., N$ indicates the subject, $j = 1, ..., m_i$ indicates the number of observations per subject, $(b_{0i}, b_{1i}) \sim N(\mathbf{0}, \boldsymbol{D})$, and $\epsilon_{ij} \sim N(0, \sigma_e^2)$. Under the model in Equation 34, the conditional exposure model $f(x_i|\boldsymbol{y_i}, \boldsymbol{t_i}, c_i)$ from Equation 33 is [20]:

$$\underbrace{\boldsymbol{Y_i}^T \boldsymbol{V_i}^{-1}(\boldsymbol{\mu_{1,i}} - \boldsymbol{\mu_{0,i}})}_{(a)} - \underbrace{\frac{1}{2}\left(\boldsymbol{\mu_{1,i}}^T \boldsymbol{V_i}^{-1}\boldsymbol{\mu_{1,i}} - \boldsymbol{\mu_{0,i}}^T \boldsymbol{V_i}^{-1}\boldsymbol{\mu_{0,i}}\right)}_{(b)} + \underbrace{log\left[\frac{P(x_i = 1|\boldsymbol{t_i}, c_i)}{P(x_i = 0|\boldsymbol{t_i}, c_i)}\right]}_{(c)} \tag{35}$$

where $\boldsymbol{\mu_{1,i}} = E[\boldsymbol{y_i}|x_i = x, \boldsymbol{t_i}, c_i]$, and $\boldsymbol{V_i} = Var(\boldsymbol{y_i}|x_i, \boldsymbol{t_i}, c_i)$. By expanding terms (a), (b), (c) in Equation 35, the imputation model can be re-written as follows:

$$\begin{aligned} log\left[\frac{P(x_i = 1|\boldsymbol{Y_i}, \boldsymbol{t_i}, c_i)}{P(x_i = 0|\boldsymbol{Y_i}, \boldsymbol{t_i}, c_i)}\right] &= \gamma_0 + \gamma_1 \sum_{j=1}^{m_i}\sum_{i=1}^{m_i} \nu_{ijk} Y_{ij} + \gamma_2 \sum_{j=1}^{m_i}\sum_{i=1}^{m_i} \nu_{ijk} Y_{ij} t_{ik} \\ &+ \gamma_3 \sum_{j=1}^{m_i}\sum_{i=1}^{m_i} \nu_{ijk} t_{ij} + \gamma_4 \sum_{j=1}^{m_i}\sum_{i=1}^{m_i} \nu_{ijk} t_{ij} t_{ik} + \gamma_5 c_i \\ &+ \gamma_6 \sum_{j=1}^{m_i}\sum_{i=1}^{m_i} \nu_{ijk} t_{ij} c_i + \gamma_7 \sum_{j=1}^{m_i}\sum_{i=1}^{m_i} \nu_{ijk} c_i + \gamma_8 \sum_{j=1}^{m_i}\sum_{i=1}^{m_i} \nu_{ijk} \end{aligned} \tag{36}$$

where ν_{ijk} is the $(j,k)^{th}$ element of $\boldsymbol{V_i}^{-1}$. In a balanced and complete design, ν_{ijk} and t_{ij} are constant across all subjects; thus, the imputation model in Equation 36 reduces to a logistic regression with covariates Y_{ij}. However, if the data are not balanced and complete, ν_{ijk} needs to be estimated in order to perform the direct approach to multiple imputation. Particularly, we proposed to estimate ν_{ijk} iteratively in the following steps:

1. Estimate an initial set of weights ν_{ijk} from a model that does not include the expensive covariate
2. Use multiple imputation to impute the missing expensive covariate
3. Fit the linear mixed effects model of interest and estimate a new set of ν_{ijk}

4. Repeat (2) and (3) M times and burn the first m iterations

Going back to our example, let us assume that 2,000 subjects have available data on outcome and inexpensive covariates; however, SNP can be measured on 400 individuals only. To understand the performance of direct imputation, we run a simulation study under two different scenarios and present summaries obtained after running each simulation 1,000 times. Across all scenarios, we set $(\theta_0, \theta_t, \theta_x, \theta_{xt}, \theta_c) = (75, -0.5, -1, -0.5, -2)$, $(b_{0i}, b_{1i}) \sim N(\mathbf{0}, \boldsymbol{D})$ with $\boldsymbol{D} = diag(81, 1.56)$ and $\epsilon_{ij} \sim N(0, 12.25)$. We examine three sampling schemes discussed in Section 3.1: random sampling, sampling based on extreme subject-specific intercept (i.e., intercept sampling), and sampling based on extreme subject-specific slope (i.e., slope sampling). We consider balanced and complete data as well as balanced and incomplete data (i.e. different individuals have an unequal number of observations). Particularly, for the balanced and complete data case we fit as our imputation model a logistic regression with each Y_{ij} as covariates; whereas, for the balanced and incomplete scenario we iteratively estimate the weights ν_{ijk}. The results are presented in Table 3.

Design	θ_0	θ_t	θ_x	θ_{xt}	θ_c
Balanced and Complete Data					
Random Sampling	75.00 (0.38)	-1.00 (0.05)	-0.49 (1.00)	-0.49 (0.14)	-2.00 (0.22)
Intercept Sampling	75.00 (0.29)	-1.00 (0.05)	-0.51 (0.66)	-0.49 (0.14)	-2.01 (0.21)
Slope Sampling	75.00 (0.38)	-1.00 (0.04)	-0.49 (1.02)	-0.50 (0.09)	-2.00 (0.23)
Balanced and Incomplete Data					
Random Sampling	75.00 (0.54)	-1.00 (0.06)	-0.50 (1.01)	-0.49 (0.14)	-2.00 (0.22)
Intercept Sampling	75.00 (0.42)	-1.00 (0.06)	-0.50 (0.66)	-0.49 (0.15)	-2.01 (0.21)
Slope Sampling	75.00 (0.53)	-1.00 (0.04)	-0.48 (1.02)	-0.49 (0.10)	-2.01 (0.21)

Table 3: Mean estimates and standard errors obtained after 1,000 simulation runs. Across all scenarios $(\theta_0, \theta_t, \theta_x, \theta_{xt}, \theta_c) = (75, -0.5, -1, -0.5, -2)$, $(b_{0i}, b_{1i}) \sim N(\mathbf{0}, \boldsymbol{D})$ with $\boldsymbol{D} = diag(81, 1.56)$ and $\epsilon_{ij} \sim N(0, 12.25)$. In the balanced and complete data scenario all subjects are measured at times 0, 1, ..., 5. In the balanced and incomplete design subjects are measured either 4, 5 or 6 times.

The coefficients' estimate in Table 3 are unbiased regardless of the type of sampling and data. By comparing the standard errors we can observe how intercept sampling leads to higher precision in the estimation of the coefficients of time-invariant covariates; whereas, slope sampling leads to higher precision in the estimation of the coefficients associated with time-variant covariates. For instance, when we compare two phase ODS with simple random sampling under balanced and complete data, intercept sampling increased efficiency for the coefficients associated with SNP by 52%, while slope sampling increased the efficiency of the coefficient associated with the SNP by time interaction by approximately 55%.

3.4 A Data Augmentation Approach to Estimate the Parameters

The direct and the indirect approaches to multiple imputation cannot be easily generalised to account for more complex sampling schemes and/or models. On one hand, since the indirect approach requires the specification of the ascertainment corrected log-likelihood, Algorithm 5 needs to be tweaked for every two phase ODS design; on the other hand, if the data are not balanced and complete, the imputation model in the direct approach is highly dependent on the number of inexpensive covariates. Thus, we developed a data augmentation algorithm [15] that is simpler to extend to models with more than one inexpensive covariate, and can be used for any design since it does not rely on the ascertainment corrected log-likelihood. Similarly to the multiple imputation approach, we focused on scenarios where the expensive covariate is binary; however, extensions to continuous and categorical expensive covariates should be feasible.

Under a multivariate Gaussian model where we assume that homoscedasticity holds, we have already seen that the log-transformed of Equation 33 can be written as:

$$\underbrace{\boldsymbol{Y}_i^t\boldsymbol{V}_i^{-1}(\boldsymbol{\mu}_{1,i}-\boldsymbol{\mu}_{0,i})-\frac{1}{2}\left(\boldsymbol{\mu}_{1,i}^t\boldsymbol{V}_i^{-1}\boldsymbol{\mu}_{1,i}-\boldsymbol{\mu}_{0,i}^t\boldsymbol{V}_i^{-1}\boldsymbol{\mu}_{0,i}\right)}_{(a)}+\underbrace{log\left[\frac{pr(X_i=1|\boldsymbol{Z_i})}{pr(X_i=0|\boldsymbol{Z_i})}\right]}_{(b)} \tag{37}$$

In many scenarios, such as those where the expensive covariate is genotype, it is reasonable to assume that X is independent of time. Thus, expression (b) in Equation 37 becomes a logistic regression model of the expensive covariate on time-invariant inexpensive covariates, and the conditional exposure odds model is simply an offsetted logistic regression with expression (a) being the offset. Given Equation 37, the steps for performing data augmentation are delineated in the algorithm below.

Algorithm 7: Data Augmentation for Two Phase ODS Design

```
1  On sampled subjects fit the linear mixed effects model:
       $\mathbf{Y}_i = \boldsymbol{X_i\theta_1} + \boldsymbol{Z_i\theta_2} + \boldsymbol{W_ib_i} + \boldsymbol{\epsilon_i}$
   where $[\boldsymbol{X_i}, \boldsymbol{Z_i}]$ is the matrix of expensive and inexpensive covariates, and $\boldsymbol{W_i}$ is the matrix of random effects.
   Let $\boldsymbol{\alpha}$ be the vector of the variances and covariances of the random effects, and the variance of the error.
   Obtain initial estimates of $\hat{\boldsymbol{\theta}}^{(0)} = \left(\hat{\boldsymbol{\theta_1}}^{(0)}, \hat{\boldsymbol{\theta_2}}^{(0)}\right)$, $\hat{\boldsymbol{\alpha}}^{(0)}$, and $\widehat{Cov}\left(\hat{\boldsymbol{\theta}}\right)$
2  At the $t^{th}$ iteration:
3  Sample
       $\boldsymbol{\theta}^{(t-1)} \sim N\left(\hat{\boldsymbol{\theta}}^{(t-1)}, Cov\left(\hat{\boldsymbol{\theta}}^{(t-1)}\right)\right)$
   Calculate $\texttt{offset}_i^{(t-1)}$ for all subjects using expression (a) in Equation 37
4  For sampled subject, fit:
       $log\left(\frac{pr(X=1|\mathbf{Y_i}, \boldsymbol{Z_i})}{pr(X=0|\mathbf{Y_i}, \boldsymbol{Z_i})}\right) = \boldsymbol{Z_i}\gamma + \texttt{offset}_i^{(t-1)}$
   to obtain $\hat{\boldsymbol{\gamma}}^{(t)}$ and $\widehat{Cov}(\hat{\boldsymbol{\gamma}}^{(t)})$
5  Sample $\boldsymbol{\gamma}^{(t)} \sim N(\hat{\boldsymbol{\gamma}}^{(t)}, \widehat{Cov}(\hat{\boldsymbol{\gamma}}^{(t)}))$
6  Combine $\boldsymbol{\gamma}^{(t)}$ with $\boldsymbol{Z_i}$, and $\texttt{offset}_i^{(t-1)}$ to estimate $\hat{p}_{X_i}^{(t)} = \widehat{pr}^{(t)}(X_i = 1|\mathbf{Y_i}, \boldsymbol{Z_i})$
7  For all unsampled subjects, sample $X_i^{(t)} \sim Bern\left(\hat{p}_{X_i}^{(t)}\right)$
8  Using all subjects, fit
       $\mathbf{Y}_i = \boldsymbol{X_i\theta_1} + \boldsymbol{Z_i\theta_2} + \boldsymbol{W_ib_i} + \boldsymbol{\epsilon_i}$
   to obtain estimates of $\hat{\boldsymbol{\theta}}^{(t)}$, $\boldsymbol{\alpha}^{(t)}$, and $\widehat{Cov}\left(\hat{\boldsymbol{\theta}}^{(t)}\right)$
9  Repeat steps (3) - (8) $M$ times until the distribution of $\boldsymbol{\theta}^{(t)}$ converges, and keep the last $m$ iterations ($m < M$).
10 Combine the $m$ estimates using Rubin's rule.
```

3.5 Extending the Methods to Multivariate Outcomes

The methods presented in this report can be extended to multivariate longitudinal data. In this section we focus on a bivariate outcome and compare ACML with data augmentation. We assume that our interest lies in a time-variant expensive covariate and we generated data for $N = 2,000$ subjects from the bivariate linear mixed model:

$$Y_{1ij} = \theta_{1int} + \theta_{1x}x_i + \theta_{1c}c_i + \theta_{1t}t_{ij} + \theta_{1xt}x_it_{ij} + \theta_{1ct}c_it_{ij} + b_{01i} + b_{11i}t_{ij} + \epsilon_{1ij}$$

$$Y_{2ij} = \theta_{2int} + \theta_{2x}x_i + \theta_{2c}c_i + \theta_{2t}t_{ij} + \theta_{2xt}x_it_{ij} + \theta_{2ct}c_it_{ij} + b_{02i} + b_{12i}t_{ij} + \epsilon_{2ij}$$

where $i = 1, ..., N$ denotes a subject, $j = 1, ..., n_i$ is the number of observations per subject, X_i is an expensive binary covariate such that $P(x_i = 1) = 0.3$, t_{ij} denotes the time variable, and c_i is a continuous confounder generated from a normal random variable with mean $-0.15 - 0.05x_i$ and variance 1. For each subject i, we generated equally spaced t_{ij} values from 0 to 5, selected dropout time randomly, and removed all subsequent time points. This resulted in subjects having either 2, 3, 4, 5 or 6 observations. We set $(\theta_{1int}, \theta_{1x}, \theta_{1c}, \theta_{1t}, \theta_{1xt}) = (75, 0.5, -2.5, -1.5, -0.25, -0.10)$, $(\theta_{2int}, \theta_{2x}, \theta_{2c}, \theta_{2t}, \theta_{2xt}) = (65, -0.6, -2, -1, -0.15, -0.15)$. The random effects $(b_{01i}, b_{11i}, b_{02i}, b_{12i})$ were generated from a multivariate normal distribution:

$$\begin{bmatrix} b_{01i} \\ b_{11i} \\ b_{02i} \\ b_{12i} \end{bmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 20.25 & 0.25 & 7.50 & 0.125 \\ & 1 & 0.75 & 0.250 \\ & & 9 & 0.375 \\ & & & 0.250 \end{pmatrix}\right)$$

The error components, independent of the random effects, were normally distributed with mean 0 and variance $\boldsymbol{\Sigma}_i = (\sigma_1^2\boldsymbol{I}, \sigma_2^2\boldsymbol{I})$ with $\sigma_1^2 = 2.25$ and $\sigma_2^2 = 1$.

We assumed that, on average, x_i could be ascertained on 400 of the original 2,000 subjects, and we considered three designs: 1) random sampling, 2) subject-specific slope sampling and 3) BLUP slope sampling. For the random sampling design we took a simple random sample of 400 subjects. For the subject-specific and the BLUP slope sampling we randomly assigned 1,000 subjects to be selected based on $\boldsymbol{Y_1}$ and the remaining to be selected based on $\boldsymbol{Y_2}$. Specifically, for subject-specific slope sampling, we generated three strata $(\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3)$ for each outcome by 1) fitting subject-specific simple linear regression of Y_{1ij} (or Y_{2ij} if the subject was assigned to be selected based on the second outcome) on t_{ij}, and 2) searching for the cutpoints of the estimated subject-specific slopes such that 10% of the original cohort was in $\mathcal{S}_1$ and in $\mathcal{S}_3$ respectively, and the remaining 80% was in the $\mathcal{S}_2$. Similar procedure was carried out for BLUP sampling where we defined $(\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3)$ based on the BLUP estimates of the random slope. For each outcome we sampled on average 200 subjects with $(75, 50, 75)$ selected from each stratum. After sub-sampling from the original cohort, we conducted complete case analysis and data augmentation. First, we considered only those subjects whose x_i was ascertained and we implemented maximum likelihood and ACML for random sampling and subject-specific slope sampling respectively; afterwards, we included partial information on unsampled subjects and used data augmentation. We run our data augmentation algorithm 650 times and burned the first 50 iterations.

We conducted 1,000 replicates and reported the estimated regression coefficients with their associated standard errors, the coverage probability of the 95% confidence interval, the ratio between the average estimated standard error and the empirical standard error, and the relative efficiency computed as:

$$\frac{Var_{RS}(\boldsymbol{\theta})}{Var_{\text{Design}}(\boldsymbol{\theta})}$$

where $Var_{RS}(\boldsymbol{\theta})$ is the empirical variation in the parameter estimates across replications under complete case random sampling and $Var_{\text{Design}}(\boldsymbol{\theta})$ is the empirical variation in the parameter estimates across replicates under complete case subject-specific slope sampling, or under any of the data augmentation methods.

Table 4 and Figure 3 summarise the results. With the exception of "Random Sampling + Data Augmentation", the bias of the estimated coefficients was less than 5%, the coverage of the confidence intervals was close to the nominal 0.95, and the standard errors estimator accurately reflected the true variation. The bias and the low coverage observed in the "Random Sampling + Data Augmentation" disappeared when increasing the sample size and the number of subjects sampled. Similarly to what observed in Schildcrout et al [20], efficiency gains were high for the inexpensive covariates available on all subjects, and moderate for the imputed expensive covariate whose efficiency gain came from the sampling design rather than the data augmentation.

			$\mathbf{Y_1}$						$\mathbf{Y_2}$			
	θ_{1int}	θ_{1x}	θ_{1c}	θ_{1t}	θ_{1xt}	θ_{1ct}	θ_{2int}	θ_{2x}	θ_{2c}	θ_{2t}	θ_{2xt}	θ_{2ct}
Random Sampling Complete Case												
Estimate	80.000	0.479	-2.497	-1.501	-0.245	-0.099	65.001	-0.607	-1.997	-1.000	-0.148	-0.150
SE	0.280	0.508	0.233	0.068	0.122	0.056	0.187	0.339	0.155	0.036	0.065	0.030
Coverage	0.954	0.937	0.945	0.955	0.952	0.958	0.947	0.952	0.953	0.937	0.940	0.954
SE Ratio	1.000	0.966	0.977	0.995	0.991	1.029	0.989	1.009	1.021	0.965	0.965	1.007
Subject-Specific Slope Sampling Complete Case												
Estimate	80.003	0.494	-2.480	-1.499	-0.253	-0.102	65.006	-0.602	-1.995	-1.000	-0.151	-0.152
SE	0.280	0.507	0.229	0.051	0.092	0.042	0.185	0.335	0.151	0.028	0.051	0.024
Coverage	0.943	0.951	0.955	0.945	0.957	0.946	0.942	0.956	0.955	0.955	0.942	0.950
SE Ratio	0.985	1.019	1.041	0.985	0.994	0.989	0.978	0.990	1.006	1.015	0.965	1.000
Random Sampling + Data Augmentation												
Estimate	80.008	0.471	-2.497	-1.502	-0.243	-0.100	65.001	-0.598	-2.001	-1.001	-0.146	-0.150
SE	0.181	0.487	0.106	0.044	0.117	0.026	0.121	0.325	0.071	0.023	0.063	0.014
Coverage	0.937	0.925	0.957	0.938	0.935	0.957	0.945	0.947	0.960	0.931	0.928	0.960
SE Ratio	0.965	0.940	1.046	0.974	0.973	1.016	0.983	0.992	1.045	0.948	0.947	1.019
Subject-Specific Slope Sampling + Data Augmentation												
Estimate	80.003	0.491	-2.497	-1.500	-0.252	-0.101	64.998	-0.593	-2.001	-1.000	-0.151	-0.150
SE	0.182	0.493	0.106	0.037	0.090	0.026	0.120	0.325	0.071	0.020	0.049	0.014
Coverage	0.952	0.942	0.960	0.944	0.945	0.947	0.941	0.940	0.954	0.951	0.931	0.947
SE Ratio	1.004	1.007	1.038	0.979	0.986	0.999	0.980	0.974	1.025	0.983	0.964	1.005
BLUP Slope Sampling + Data Augmentation												
Estimate	80.003	0.492	-2.497	-1.501	-0.249	-0.100	64.994	-0.580	-1.999	-1.000	-0.151	-0.150
SE	0.181	0.490	0.106	0.037	0.089	0.026	0.119	0.319	0.071	0.020	0.048	0.014
Coverage	0.953	0.940	0.957	0.941	0.945	0.946	0.947	0.939	0.959	0.958	0.944	0.945
SE Ratio	0.993	0.954	1.038	0.971	0.983	1.004	0.970	0.967	1.039	1.000	0.988	1.005

Table 4: Estimated coefficients, standard errors, and coverage probability across 1,000 replicates. SE, standard error. SE ratio indicates the ratio between the average estimated standard error and the empirical standard error. BLUP, best linear unbiased predictor.

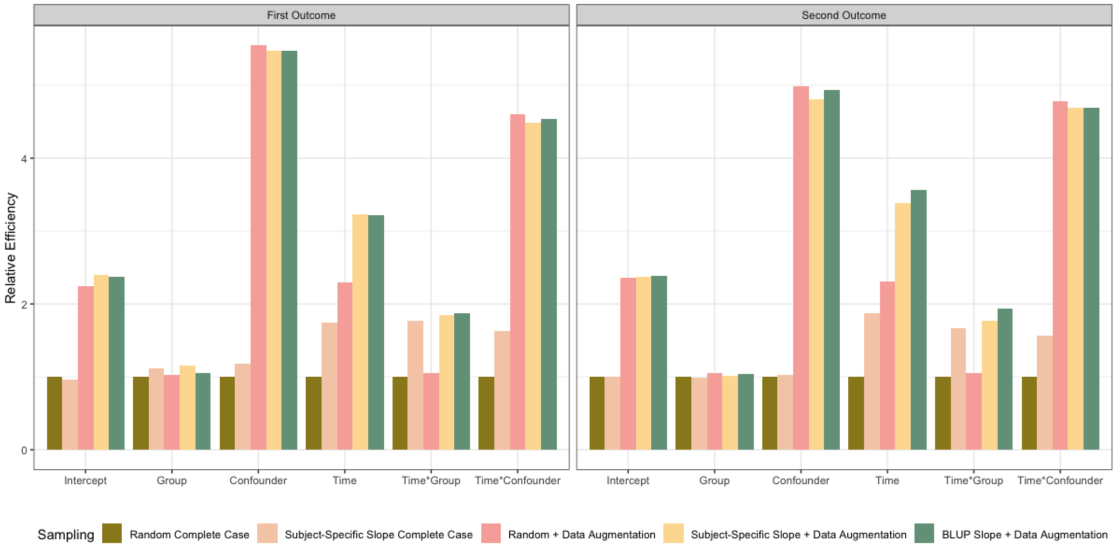


Figure 3: Relative efficiency of the parameters estimate under different sampling designs and inference methods. BLUP, best linear unbiased predictor.

3.6 Methods for Re-Using Data from an ODS Two Phase Sampling Scheme

Similarly to the case of a cross-sectional outcome, researchers might be interested in re-using data from a two phase ODS design in order to make inference about a secondary longitudinal outcome. The methods introduced in Section 3.5 can be extended to accommodate a secondary outcome by building a bivariate model where the first outcome is the one previously used for sampling, and the second outcome is the one we are currently interested in.

In this Section we perform a simulation study to better understand how we can re-use data from an ODS two phase sampling scheme. We generated data as in Section 3.5, and assumed that a previous two phase ODS design was carried out, and 400 individuals (out of 2,000) were sampled based on the subject-specific slopes estimated from the regression of $\boldsymbol{Y}_{1i}$ on t_{ij}. We conducted 1,000 replicates and reported the estimated regression coefficients with their associated standard errors, the coverage probability of the 95% confidence interval, and the ratio between the average estimated standard error and the empirical standard error.

Table 5 summarises the results obtained under three different analyses: the naive analysis that ignores the previous two phase ODS sampling scheme; the subject-specific slope sampling that accounts for the previous two phase ODS, but consider subjects with complete information on outcome and expensive covariates only; and the subject-specific slope sampling plus data augmentation that accounts for the previous two phase ODS and includes all individuals. By accounting for the sampling scheme, the estimated coefficients and standard errors are unbiased. While we are not gaining any efficiency in the analysis of the secondary outcome, we are still able to re-use data from a biased sampling scheme to make valid inference.

	$\mathbf{Y_1}$						$\mathbf{Y_2}$					
	θ_{1int}	θ_{1x}	θ_{1c}	θ_{1t}	θ_{1xt}	θ_{1ct}	θ_{1int}	θ_{1x}	θ_{1c}	θ_{1t}	θ_{1xt}	θ_{1ct}
Naive Analysis												
Estimate	-	-	-	-	-	-	65.023	-0.822	-2.087	-0.990	-0.223	-0.181
SE	-	-	-	-	-	-	0.194	0.369	0.160	0.040	0.077	0.033
Coverage	-	-	-	-	-	-	0.951	0.907	0.923	0.957	0.836	0.855
SE Ratio	-	-	-	-	-	-	1.001	1.023	1.027	1.085	0.983	1.007
Subject-Specific Slope Sampling Complete Case												
Estimate	80.004	0.499	-2.493	-1.501	-0.249	-0.101	64.990	-0.604	-1.999	-1.001	-0.148	-0.151
SE	0.280	0.535	0.231	0.043	0.083	0.036	0.184	0.352	0.152	0.035	0.066	0.029
Coverage	0.936	0.936	0.944	0.953	0.954	0.945	0.943	0.955	0.958	0.951	0.943	0.952
SE Ratio	0.939	0.954	1.021	0.993	1.006	0.990	0.975	1.014	1.004	0.997	0.972	1.005
Subject-Specific Slope Sampling + Data Augmentation												
Estimate	80.003	0.489	-2.498	-1.501	-0.248	-0.100	65.000	-0.595	-1.998	-1.001	-0.147	-0.150
SE	0.169	0.512	0.123	0.033	0.081	0.027	0.112	0.336	0.082	0.021	0.063	0.016
Coverage	0.938	0.928	0.946	0.947	0.952	0.946	0.949	0.945	0.958	0.948	0.942	0.943
SE Ratio	0.965	0.937	0.994	0.989	0.993	0.992	0.966	0.993	1.022	0.972	0.951	0.984

Table 5: Estimated coefficients, standard errors, and coverage probability across 1,000 replicates. SE, standard error. SE Ratio indicates the ratio between the average estimated standard error and the empirical standard error

4 Summary

The methods presented in this report are summarised in Table 6. For each method, we report the amount of data used, as well as the type of outcome supported, and the estimation. Overall, models that consider sampled and unsampled subjects (i.e., full likelihood) led to more efficient estimates than conditional likelihood methods. However, if the unsampled individuals did not contain any information on the expensive covariates of interest, conditional and full likelihood methods had similar efficiency [26].

Name	Data used	Outcome	Estimation type	Notes
Logistic regression [12]	Phase 2	Binary	Conditional likelihood	For cross-sectional data only. Can be used only when sampling for phase two does not depend on any inexpensive covariate
Complete data likelihood [6, 16, 17]	Phase 2	Categorical Continuous	Conditional likelihood	Ascertainment corrected likelihood is a complete data likelihood
Semiparametric empirical likelihood [31]	Phase 2	Categorical Continuous	Conditional likelihood	Assumes that strata are a function of the outcome only
Weighted pseudolikelihood [6]	Phase 2	Categorical Continuous	Conditional likelihood	Assumes each subject has non zero probability of being selected
Estimated pseudo-likelihood [6]	Phase 2 + phase 1 stratum membership	Categorical	Full likelihood	Requires continuous outcome and covariates data to be discretised
Semi-parametric likelihood [6]	Phase 2 + phase 1 stratum membership	Categorical	Full likelihood	Requires continuous outcome and covariates data to be discretised
Mean score estimator [13]	Phase 1 + 2	Categorical	Full likelihood	Requires categorical outcome and inexpensive covariates
Pseudoscore estimator [4, 3]	Phase 1 + 2	Categorical Continuous	Full likelihood	Allows for the inclusion of categorical and continuous inexpensive covariates measured in phase one
Estimated likelihood estimator [26]	Phase 1 + 2	Categorical Continuous	Full likelihood	Allows for the inclusion of categorical inexpensive covariates measured in phase one
Maximum likelihood [22, 7, 24]	Phase 1 + 2	Categorical Continuous	Full likelihood	Assumes the absence of inexpensive covariates. Extended to multivariate outcomes in [24]
Semiparametric maximum likelihood [25]	Phase 1 + 2	Categorical Continuous	Full likelihood	Allows for the inclusion of categorical and continuous inexpensive covariates measured in phase one. Can be extended to multivariate outcomes
Full conditional likelihood [23]	Phase 1 + 2	Categorical Continuous	Full likelihood	Likelihood used for BLUP based sampling
Multiple imputation [20]	Phase 1 + 2	Continuous	Full likelihood	Two different approaches are available: direct imputation and indirect imputation. The latter explicitly uses the likelihood in ACML. For binary expensive covariates only.
Data augmentation	Phase 1 + 2	Continuous	Full likelihood	For binary expensive covariates only. Can be extended to multivariate outcomes

Table 6: Summary of available methods for the estimation of parameters in two-phase design outcome dependent sampling studies

5 Conclusion and Future Work

When the research question of interest requires collection of costly variables that limit our sample size, it is important to select individuals that provide us with the greatest amount of information. Using available data, the two phase ODS design identifies informative individuals for whom we should devote our resources. This report introduced two phase ODS for both cross-sectional and longitudinal outcomes, summarised likelihood based methods that have been proposed in the literature, and showed how by sampling informative individuals we can increase the precision (and consequently the power) of a study.

In addition to the data augmentation approach described in this report, we plan to introduce an EM algorithm for inference under a two phase design that should be computationally faster than data augmentation and can easily deal with missing outcome variables. Another potential future direction is to use a two phase sampling for a prospective longitudinal study where the outcome would be considered an additional expensive variable, and sampling would be done based on an auxiliary variable related to the outcome of interest.

References

[1] N. E. Breslow, *Statistics in epidemiology: The case-control study*, J. Am. Stat. Assoc. **91** (1996), no. 433, 14–28.

[2] N. E. Breslow and N. Chatterjee, *Design and analysis of two-phase studies with binary outcome applied to Wilms tumour prognosis*, J. R. Stat. Soc. Ser. C Appl. Stat. **48** (1999), no. 4, 457–468.

[3] Nilanjan Chatterjee and Yi Hau Chen, *A semiparametric pseudo-score method for analysis of two-phase studies with continuous phase-I covariates*, Lifetime Data Anal. **13** (2007), no. 4, 607–622.

[4] Nilanjan Chatterjee, Yi Hau Chen, and Norman E. Breslow, *A pseudoscore estimator for regression problems with two-phase sampling*, J. Am. Stat. Assoc. **98** (2003), no. 461, 158–168.

[5] D Holt, TMF Smith, and PD Winter, *Regression Analysis of Data from Complex Survey*, J. R. Stat. Soc. Ser. A **143** (1980), no. 4, 474–487.

[6] J. F. Lawless, J. D. Kalbfleisch, and C. J. Wild, *Semiparametric methods for response-selective and missing data problems in regression*, J. R. Stat. Soc. Ser. B Stat. Methodol. **61** (1999), no. 2, 413–438.

[7] DY Lin, D Zeng, and ZZ Tang, *Quantitative trait analysis in sequencing studies under trait-dependent sampling*, Proc. Natl. Acad. Sci. U. S. A. **110** (2013), no. 30, 12247–12252.

[8] Thomas Lumley, *survey: analysis of complex survey samples*, 2020, R package version 4.0.

[9] J. M. Neuhaus, A. J. Scott, and C. J. Wild, *Family-specific approaches to the analysis of case-control family data*, Biometrics **62** (2006), no. 2, 488–494.

[10] John M. Neuhaus and Nicholas P. Jewell, *The Effect of Retrospective Sampling on Binary Regression Models for Clustered Data*, Biometrics **46** (1990), no. 4, 977.

[11] Y Pan, J Cai, MP Longnecker, and H Zhou, *Secondary outcome analysis for data from an outcome-dependent sampling design*, Stat. Med. **37** (2018), no. 15, 2321–2337.

[12] R. L. Prentice and R. Pyke, *Logistic disease incidence models and case-control studies*, Biometrika **66** (1979), no. 3, 403–411.

[13] Marie Reilly and Margaret Sullivan Pepe, *A mean score method for missing and auxiliary covariate data in regression models*, Biometrika **82** (1995), no. 2, 299–314.

[14] DB Rubin, *Inference and missing data*, Biometrika **63** (1976), no. 3, 581–590.

[15] JL Schafer, *Analysis of incomplete multivariate data*, CRC press, 2013.

[16] J Schildcrout and Patrick J Heagerty, *Outcome dependent sampling from existing cohorts with longitudinal binary response data: study planning and analysis*, Biometrics **67** (2011), no. 4, 1583–1593.

[17] Jonathan S Schildcrout, Shawn P Garbett, and Patrick J Heagerty, *Outcome vector dependent sampling with longitudinal continuous response data: stratified sampling based on summary statistics*, Biometrics **69** (2013), no. 2, 405–16.

[18] Jonathan S Schildcrout, Sebastien Haneuse, Ran Tao, Leila R Zelnick, Enrique F Schisterman, Shawn P Garbett, Nathaniel D Mercaldo, Paul J Rathouz, and Patrick J Heagerty, *Two-phase, generalized case-control designs for quantitative longitudinal outcomes*, Am. J. Epidemiol. **182** (2019), no. 2, 81–90.

[19] Jonathan S. Schildcrout and Patrick J. Heagerty, *On outcome-dependent sampling designs for longitudinal binary response data with time-varying covariates*, Biostatistics **9** (2008), no. 4, 735–749.

[20] Jonathan S Schildcrout, Paul J Rathouz, Leila R Zelnick, Shawn P Garbett, and Patrick J Heagerty, *Biased Sampling Design To Improve Research Eefficiency: Factors Influencing Pulmonary Function Over Time In Children With Asthma*, Ann. Appl. Stat. **9** (2015), no. 2, 731–753.

[21] Author A J Scott and C J Wild, *Fitting Regression Models to Case-Control Data by Maximum Likelihood*, Biometrika **84** (1997), no. 1, 57–71.

[22] Rui Song, Haibo Zhou, and Michael R. Kosorok, *A note on semiparametric efficient inference for two-stage outcome-dependent sampling with a continuous outcome*, Biometrika **96** (2009), no. 1, 221–228.

[23] Zhichao Sun, Bhramar Mukherjee, JP Estes, PS Vokonas, and SK Park, *Exposure Enriched Outcome Dependent Designs for Longitudinal Studies of Gene-Environment Interaction*, Stat. Med. **36** (2017), no. 18, 2947–2960.

[24] R Tao, D Zeng, N Franceshini, KE North, E Boerwinkle, and DY Lin, *Analysis of Sequence Data Under Multivariate Trait-Dependent Sampling*, J. Am. Stat. Assoc. **110** (2015), no. 510, 560–572.

[25] Ran Tao, Donglin Zeng, and Dan Yu Lin, *Efficient Semiparametric Inference Under Two-Phase Sampling, With Applications to Genetic Association Studies*, J. Am. Stat. Assoc. **112** (2017), no. 520, 1468–1476.

[26] Mark A. Weaver and Haibo Zhou, *An estimated likelihood method for continuous outcome regression models with outcome-dependent sampling*, J. Am. Stat. Assoc. **100** (2005), no. 470, 459–469.

[27] Emily J White, *A two stage design for the study of the relationship between a rare exposure and a rare disease*, Am. J. Epidemiol. **115** (1982), no. 1, 119–128.

[28] KW Whitworth, LS. Haug, DD. Baird, G Becher, JA Hoppin, R Skjaerven, C Thomsen, M Eggesbo, G Travlos, R Wilson, and MP Longnecker, *Perfluorinated compounds and subfecundity in pregnant women*, Epidemiology **23** (2012), no. 2, 257–263.

[29] Leila R Zelnick, Jonathan S Schildcrout, and Patrick J Heagerty, *Likelihood-based analysis of outcome-dependent sampling designs with longitudinal data*, Stat. Med. **37** (2018), no. 13, 2120–2133.

[30] Haibo Zhou, Jianwei Chen, Tiina Rissanen, Susan Korrick, Howard Hu, Jukka Salonen, and MP Longnecher, *An efficient sampling and inference procedure for studies with a continuous outcome*, Epidemiology **18** (2017), no. 4, 461–468.

[31] Haibo Zhou, Mark A. Weaver, J Qin, MP Longnecher, Wang, and MC, *A Semiparametric Empirical Likelihood Method for Data from an Outcome-Dependent Sampling Scheme with a Continuous Outcome*, Biometrics **58** (2002), no. 2, 413–421.