The International Journal of Biostatistics

Volume 4, Issue 1

2008

Article 15

Systematic Missing-At-Random (SMAR) Design and Analysis for Translational Research Studies

Ilana Belitskaya-Levy*

Yongzhao Shao[†]

Judith D. Goldberg[‡]

*New York University School of Medicine, beliti01@nyumc.org

[†]Iowa State University, yshao@iastate.edu

[‡]New York University School of Medicine, jd.goldberg@nyumc.org

Systematic Missing-At-Random (SMAR) Design and Analysis for Translational Research Studies*

Ilana Belitskaya-Levy, Yongzhao Shao, and Judith D. Goldberg

Abstract

Translational research studies often involve a central study (e.g. clinical trial, cohort of patients, etc.) and multiple investigators who are each interested in addressing different research questions using the same patient population. However, it is often impossible for the investigators to include all patients in all of the ancillary translational research substudies that are part of the main study. This arises due to time and budgetary constraints and other logistical considerations. In this paper, we propose a prospective Systematic Missing-At-Random study design (SMAR) with planned partially missing covariates collected using a nested random sampling scheme that allows an integrated statistical analysis across all domains of data. We propose an algorithm for data analysis that incorporates the features of the design. We show that the SMAR design is computationally and statistically efficient as well as cost effective using simulation studies and a published data example. An extension to a two-stage prospective-retrospective design is discussed.

KEYWORDS: study design, translational research, prospective design, monotone missing pattern, nested random sampling, EM algorithm, outcome-dependent sampling, clinical trial design

*This research was supported in part by the General Clinical Research Center Grant MO1 00096RR00096-43 from NIH/NCRR, the National Cancer Institute Grant P30 CA16087-23 and the Department of Defense Grant W81XWH-04-2-0905.

1 Introduction

Studies of complex human diseases often involve a central study (e.g. clinical trial, cohort of patients, etc) and a translational research component with multiple investigators who are each interested in addressing different research questions using the same patient population. For example, in an ongoing multi-center clinical trial of patients with Locally Advanced Breast Cancer (LABC) funded by the Department of Defense, a cohort of patients, identified at several international sites, treated under a standard protocol, is followed for response to treatment, progression, and survival (Formenti and Schneider, 2004). The number of patients required for the evaluation of primary endpoints is relatively large. In order to develop a comprehensive model for the understanding of LABC, genetic, molecular and immunological aspects of the disease are under study by several different research groups, in addition to studies of social, cultural, and other clinical and demographic factors. However, many of the modern technologies required for gene expression, genomic mutation, immune profile and molecular/biochemical analyses, are expensive in both time and budgetary resources. These constraints limit the number of patients that can be assessed with these technologies. The usual approach to such trials has been to provide each research group with a small number of patient samples for their specific substudies in a relatively ad hoc manner. This conventional approach makes it difficult to develop a comprehensive, integrated analysis of the disease and treatment across all of the domains of data.

1.1 The prospective SMAR design

In this paper, we describe an approach to study design, Systematic Missing-At-Random (SMAR), that incorporates the resource limitations and allows for a completely integrated analysis of data across all domains. The SMAR design uses the *entire study group* for the measurement of primary endpoints, important covariates and ‘inexpensive variables’ and uses *nested random subsamples* to measure the more ‘expensive’ classes of variables. The more expensive the variable is to measure, the smaller the nested random subsample for which it is measured. In the LABC study currently underway, the multiple laboratory investigators have all agreed to participate and to receive samples allocated according to the proposed systematic design. Such a centralized design is feasible and easily manageable using modern IT technologies.

The following are some of the features of the SMAR design:

1. The nested design produces data with a *monotone missing pattern*, as described in Little and Rubin (2002), that allows efficient statistical computation for the integrated analysis of data with partially observed covariates.
2. SMAR data are collected prospectively so that all the parameters in the model are identifiable. Prospective study designs (including controlled and uncontrolled clinical trial designs) are often used in cancer studies for the evaluation of efficacy and safety of treatment. For example, in the LABC study, the prospective clinical trial design is used to estimate the rate of response to therapy and long-term survival on therapy, which would not have been possible under a retrospective design.
3. SMAR uses the entire study group for the analyses of primary endpoints and covariates of interest yielding the required power for primary analyses.
4. SMAR produces a common random subsample of predetermined minimal size N_0 with complete observations on all variables, that can be used to obtain initial parameter estimates and for other purposes.
5. The proposed SMAR design is versatile in that it can be applied to clinical trials, cohort studies, and many other study designs, including retrospective ones such as case-control studies. This is further discussed in Section 2.1

In the later part of this paper, we demonstrate that the proposed systematic design yields estimators that are generally more statistically efficient and more cost effective than the estimators resulting from designs with completely observed covariates for a relatively small number of subjects or from the missing-at-random designs where each investigator independently selects a subgroup of patients to conduct individual research and measures the covariates independently of others.

1.2 Two-stage prospective-retrospective SMAR design

In the LABC study described above, a prospective design is necessary to ensure that the primary endpoint (i.e. treatment response rate) is identifiable. However, in the case of rare outcomes (e.g. low response rates), such prospective sampling is inefficient. To increase efficiency, we propose a second-stage

retrospective outcome-dependent sampling procedure that involves randomly sampling patients with incomplete covariates to obtain more covariate data for these patients. For these subjects sampled retrospectively at the second stage, we can measure the previously incompletely observed covariates using the original, stored patient samples (e.g. stored blood samples). These additional measurements can be used to improve accuracy of estimation of the key model parameters (e.g. odds ratios) in terms of decreased variances and shorter confidence intervals. We will demonstrate that the efficiency and cost-effectiveness of the proposed two-stage design can be greater than that of a one-stage SMAR design with prospective sampling only (e.g. in case of low response rates). The two-stage SMAR design is further discussed in Section 3.1.

1.3 Comparison to two-stage stratified sampling designs

The proposed extension of the SMAR design, the two-stage SMAR design, is different from the traditional two-stage designs, introduced originally by Neyman (1938) as a technique for stratification, and currently used in complex survey sampling (Skinner et al., 1989), in validation sampling for measurement error problems (Carroll et al., 1995) and in stratified case-control sampling in epidemiology (Breslow, 1996). These two-stage stratified sampling designs involve taking a large simple random sample from the population of interest and collecting data on the outcome variable and some of the covariates of interest (e.g. disease and exposure). In the second stage, one takes a subsample of individuals in the study, usually stratified by disease and/or exposure status, and gathers information about some extra covariates for this subsample only. Therefore, all the data gathered on these covariates are collected retrospectively, based on the outcome and other covariates. In contrast, the SMAR design ensures that at least some data are collected prospectively for all variables. This prospective design enables identification of all parameters in the model, such as rate of response to treatment, survival rates, etc., which is not possible with the traditional two-stage designs that are retrospective by nature. Analytical approaches for two-stage stratified sampling designs are well summarized in Chatterjee et al. (2003).

1.4 Comparison to case-control and case-cohort designs

The proposed SMAR design also differs from nested case-control and case-cohort studies which arise from existing cohort studies when novel methodological and analytic advances allow for new measurements to be made on previously collected and stored samples. In nested case-control and case-cohort studies, patient samples are collected at baseline and stored for all patients; however, the data that are analyzed are based only on a subsample of the entire cohort. In contrast, the SMAR design allows us to combine data from the nested random subsamples with data from the entire cohort in an integrated analysis across all domains of the data. Additionally, the proposed SMAR design ensures that at least *some* data are collected prospectively for *all* variables (i.e. stage-one data). Then in the second stage, *more* data may be collected retrospectively, based on the outcome, on the variables with planned missing values, using the original, stored patient samples. The stage-one prospective collection of data ensures that all parameters in the model are identifiable, while the additional data collected in the second stage are used to calibrate the model and to improve the precision of the parameter estimates. Therefore, the proposed SMAR design has advantages over the two-stage stratified sampling designs as well as nested case-control and case-cohort studies.

The paper is organized as follows. The SMAR design is described in Section 2.1. The details of the EM algorithm used for the analysis of SMAR data are outlined in Section 2.2. Extensions to two-stage prospective-retrospective designs that involve additional sampling stratified by the outcome are discussed in Section 3.1. We illustrate statistical efficiency and cost effectiveness of SMAR using simulations and real data examples presented in Section 3.2. In example 1, we generate datasets with 8 binary predictors and heavy missingness on 5 of the predictors. Comparison between one-stage and two-stage SMAR designs is carried out in example 2. Example 3 involves a generalized linear model with gamma link. In example 4, we carry out a cost-benefit analysis of SMAR. The case of correlated covariates is presented in Example 5. Finally, we present a real data example. Discussion and final conclusions are presented in Section 4.

2 The SMAR design and analysis

2.1 The SMAR design

Nested random subsampling

Suppose we have a study population with N subjects. Denote the outcome variable (the primary endpoint) by Y which is assumed to be observable by the end of the trial for all subjects. Let $\mathbf{Z} = (Z_1, \dots, Z_m)$ be the covariates (e.g. baseline characteristics, clinical and demographic variables) that are also observed completely. Let $\mathbf{X} = (X_1, \dots, X_p)$ be the incompletely observed covariates (e.g. the more expensive variables such as those in genetic studies). Moreover, let X_j be observed only for n_j subjects for $j = 1, \dots, p$. Without loss of generality, we assume that $n_1 \geq \dots \geq n_p = N_0$ where N_0 is a predetermined minimal sample size for all covariates regarded as potentially important but constrained by time and other logistic factors.

The design is implemented in the following way. We first draw a simple random subsample (SRS) of size n_1 from the original study sample of N subjects and measure the covariate X_1 for this subsample of size n_1 ; then a second SRS of size n_2 , where $n_2 \leq n_1$, is drawn from the subsample of n_1 subjects to study X_2 , etc, until a SRS of size n_p from the $(p - 1)$ th selected subsample is drawn to measure the covariate X_p . This is a nested design with a monotone missing pattern as described in Little and Rubin (2002).

Sample size and cost considerations

In general, the more expensive the variable or the class of variables is, the smaller the nested random subsample on which it is measured. Additionally, the SMAR design allows the allocation of more resources to study the more ‘important’ variables. The actual sample sizes, $n_1 \geq \dots \geq n_p$, are determined for each variable or class of variables separately, taking into account the cost of collecting that particular variable and prior data.

Application to other study designs

The proposed SMAR design can be used with both cohort studies as well as randomized clinical trials of treatment/intervention. In the context of randomized clinical trials, the random sampling scheme for selection of subjects for ancillary studies can be set up using treatment group as a stratification variable. In the statistical modeling, the treatment group is used as an indicator explanatory variable that is assumed to be completely observed.

Similarly, for multi-center trials, an indicator variable can be used to identify the study center. This indicator variable can be used as a stratification variable in the random sampling scheme and as an explanatory variable for the analysis of data.

Incorporation of censored or longitudinal outcomes

The proposed SMAR design can be used for studies with censored or longitudinal outcomes. For the censored outcomes, the assumption is that the censoring status, just like the outcome variable, is completely observed by design. For the longitudinal outcomes, the random sampling scheme is set up at the design stage for all subjects. Thus if, according to the random sampling scheme, a certain measure is set to be collected for a subject, then all repeated measures will be collected for that subject as well; similarly, if a particular measure is set to be missing for a subject, then all repeated measures will be missing for this subject as well.

2.2 EM-based likelihood analysis

Preliminary likelihood analysis

Let $\mathbf{Z} = (Z_1, \dots, Z_m)$ be the covariates that are observed completely, and $\mathbf{X} = (X_1, \dots, X_p)$ be the covariates that are incompletely observed. We assume that some parametric model with parameter β is proper for the conditional distribution of $Y|(\mathbf{Z}, \mathbf{X})$, say $f(y|\mathbf{z}, \mathbf{x}, \beta)$, which is specified completely except for the value of the parameter β in $\mathbb{R}^d$. We denote the observed part of $\mathbf{X}$ as $\mathbf{X}_{obs}$ and the missing part of $\mathbf{X}$ as $\mathbf{X}_{mis}$, that is $\mathbf{X} = (\mathbf{X}_{obs}, \mathbf{X}_{mis})$. We assume that the covariates $\mathbf{X} = (X_1, \dots, X_p)$ are random variables that come from a discrete distribution with levels $1, \dots, L_j$ and probabilities $\pi_j = (\pi_{j1}, \dots, \pi_{jL_j})$, $j = 1, \dots, p$. The probability vectors, $\pi = (\pi_1, \dots, \pi_p)$, are regarded as nuisance parameters. Let R_j be the indicator of missing status of X_j , $j = 1, \dots, p$. The distribution of R_j 's is determined by design, i.e. R_j 's are independent of $\{Y, \mathbf{Z}, \mathbf{X}\}$. That is, the data are missing-completely-at-random (MCAR) (Little and Rubin, 2002).

Without loss of generality, we will assume that the covariates with missing values (i.e. $\mathbf{X} = (X_1, \dots, X_p)$) are binary. Categorical covariates with more than two levels can then be fitted using binary ("dummy") variables.

Initial estimates of β

To obtain initial estimates, $\hat{\beta}_0$, of β , we maximize the complete-case data log-likelihood function of β based on N_0 subjects for whom all the covariates are measured:

$$L(\beta|N_0) = \sum_{i=1}^{N_0} \log f_{\beta}(y_i, \mathbf{z}_i, \mathbf{x}_i). \quad (2.1)$$

These preliminary estimates $\hat{\beta}_0$ are used as initial values in the EM algorithm (see Section 2.2) to maximize the full data likelihood. The observed data from all N participating subjects, including the subjects with incomplete observations on covariates, is used to calibrate the model and increase the precision of the estimates of β .

Initial estimates of π

As we mentioned earlier, our main interest is the estimation of β . The probabilities, π , of the covariate distributions are treated as nuisance parameters. The monotone missing pattern of the SMAR data makes it possible to obtain explicit initial estimates, $\hat{\pi}_0$, of the probabilities π prior to EM iterations via generalization of results presented in Little and Rubin (2002, chapter 13.2). The closed form expression for the initial maximum likelihood estimates $\hat{\pi}_0$ of π can be found in the Appendix (see Section 5.1).

The EM algorithm

To obtain the observed data maximum likelihood estimates of β based on all N subjects, we use the EM algorithm by the method of weights described in Ibrahim (1990). Specifically, we denote $\theta = (\beta, \pi)$ as the augmented parameter and write the ‘complete data’ log-likelihood for θ as:

$$L(\theta; N) = \sum_{i=1}^N \log f_{\theta}(y_i, \mathbf{z}_i, \mathbf{x}_i) = \sum_{i=1}^N [\log f_{\beta}(y_i|\mathbf{z}_i, \mathbf{x}_i) + \log p_{\pi}(\mathbf{x}_i)]. \quad (2.2)$$

Since not all of the $\mathbf{x}_i$ ’s are completely observed, we derive the expected complete data likelihood conditioning on the observed data, i.e. $E(L(\theta; N)|Y, \mathbf{Z}, \mathbf{X}_{obs}, \theta^{(s)})$, which is

$$\sum_{i=1}^N \sum_{\mathbf{x}_{mis,i}} p(\mathbf{x}_{mis,i} | y_i, \mathbf{z}_i, \mathbf{x}_{obs,i}, \boldsymbol{\theta}^{(s)}) \log f_{\boldsymbol{\theta}}(y_i, \mathbf{z}_i, \mathbf{x}_i), \quad (2.3)$$

where $\boldsymbol{\theta}^{(s)}$ denotes the current estimate of $\boldsymbol{\theta}$ in the EM iterations started at $\boldsymbol{\theta}_0 = (\hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{\pi}}_0)$; the weights, $p(\mathbf{x}_{mis,i} | y_i, \mathbf{z}_i, \mathbf{x}_{obs,i}, \boldsymbol{\theta}^{(s)})$, are the conditional distributions of the missing covariates given the observed data and the current estimate of $\boldsymbol{\theta}$; and the inner sum is taken over all possible values (patterns) of $\mathbf{x}_{mis,i}$. Maximizing the expected log-likelihood in (2.3) as a function of $\boldsymbol{\theta}$ to obtain $\boldsymbol{\theta}^{(s+1)}$ and then iterating the procedure is the usual EM algorithm (Dempster et al., 1977).

The E-step of the algorithm reduces to the computation of weights using the Bayes rule. Specifically, for the i th observation, we have

$$p(\mathbf{x}_{mis,i} | y_i, \mathbf{z}_i, \mathbf{x}_{obs,i}, \boldsymbol{\theta}^{(s)}) = \frac{f(y_i | \mathbf{z}_i, \mathbf{x}_i, \boldsymbol{\beta}^{(s)}) p(\mathbf{x}_i | \hat{\boldsymbol{\pi}}^{(s)})}{\sum_{\mathbf{x}_{mis,i}} f(y_i | \mathbf{z}_i, \mathbf{x}_i, \boldsymbol{\beta}^{(s)}) p(\mathbf{x}_i | \hat{\boldsymbol{\pi}}^{(s)})}. \quad (2.4)$$

The M-step of the EM algorithm maximizes the function in (2.3) iteratively over $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\pi})$, which is equivalent to doing complete case maximum likelihood with each incomplete observation replaced by a set of weighted “filled-in” observations. The maximization over $\boldsymbol{\beta}$ can be carried out using any standard optimization algorithm such as the Newton-Raphson step. The maximization over $\boldsymbol{\pi}$ yields the MLEs for $\boldsymbol{\pi}$ given in Appendix (see Section 5.2).

3 Extensions and examples

In this section, we discuss an extension of the proposed prospective SMAR design to a two-stage prospective-retrospective design and give examples.

3.1 A second stage retrospective outcome-dependent sampling

So far we have described a study design where all the data are collected prospectively. However, for rare outcomes (e.g. low response rates), such prospective sampling might be inefficient. To increase efficiency, we propose a second-stage retrospective outcome-dependent sampling procedure that involves randomly sampling patients with incomplete covariates to obtain more

covariate data for these patients. For these subjects sampled retrospectively, we can measure the previously missing covariates using the original, stored patient samples (e.g. stored blood samples). These additional measurements are used to improve accuracy of estimation of the key model parameters (e.g. odds ratios) in terms of decreased variances and shorter confidence intervals. We expect the efficiency and cost-effectiveness of the proposed two-stage design to be greater than that of a larger scale one-stage design with prospective sampling only. On the other hand, if only the second stage is conducted, then some of the key parameters cannot be identified (Chen, 2003). For example, in the LABC study described above, the prospective design is necessary for the primary endpoints (e.g. treatment response rate) to be identifiable.

Notice that the missing status indicators, $R_1, \dots, R_p$, in this two-stage design depend only on the observed variable Y (i.e. the outcome); therefore, the resulting two-stage data are missing-at-random (MAR) and the likelihood-based inference is valid (Rubin, 1976).

Since the second stage of the proposed two-stage design introduces outcome-dependent sampling, a proper likelihood formulation must be used in order to obtain valid and efficient estimates of the treatment effects. Prentice and Pyke (1979) showed that for logistic regression models applying the prospective model (2.2) to retrospective data leads to correct parameter and odds ratio estimates. For other models, such as survival and time series models, the prospective SMAR design can be used to provide consistent estimation of the parameters.

3.2 Examples

In this section, the SMAR design and algorithm are illustrated using five examples with simulated data and one published data. All examples involve generalized linear models.

Example 1: Comparison of SMAR and the analysis using completely observed data only

For this example, the outcome variable Y is generated as Bernoulli with parameter $p = P(Y = 1|\mathbf{X})$, where p comes from the following logit model:

$$\log \frac{p}{1-p} = 1 - X_1 + X_2 - X_3 + X_4 - X_5 + X_6 - X_7 + X_8, \quad (3.1)$$

and the eight covariates $X_1, \dots, X_8$ are independent Bernoulli (q) random variables with ‘success’ probabilities $q = (0.25, 0.75, \dots, 0.25, 0.75)$. The

total number of subjects is $N = 500$. Forty subjects have missing values for variables X_3 through X_8 . An additional 40 subjects have missing values for variables X_5 through X_8 . An additional 120 subjects have missing values for variables X_7 and X_8 . The remaining 300 subjects have no missing values. Thus, variables X_7 and X_8 have the most missing values (i.e. 200 each), variables X_5 and X_6 have missing values for 80 cases and variables X_3 and X_4 have missing values for 40 cases. Variables X_1 and X_2 have no missing values.

We compare the performance of the logistic model in terms of bias, standard error and mean squared error (MSE) of estimated coefficients for analysis using only the complete-case data (CC, $N = 300$) and the SMAR design. We also report the estimated coefficients and their standard errors obtained using the complete-case data for all $N = 500$ subjects (ALL), which could only be available for a study with unlimited resources.

The standard errors of the SMAR estimates can be computed empirically using 1000 simulations. For each simulation, SMAR data with missing values are generated using model 3.1 and the coefficients are estimated using complete-case data ($N = 300$), SMAR data, and all the data ($N = 500$). Empirical standard errors are computed as standard deviations of coefficients over 1000 simulations.

Asymptotic variances of the SMAR estimates can be obtained by finding the observed information matrix using the method of Louis (1982). In this example, the observed information matrix of β based on the observed data can be written as:

$$I(\beta) = \tilde{\mathbf{X}}^T \mathbf{W} \hat{\mathbf{P}} (\mathbf{I} - \hat{\mathbf{P}}) \tilde{\mathbf{X}} - \tilde{\mathbf{X}}^T \mathbf{W} (\tilde{\mathbf{Y}} - \hat{\mathbf{P}})^2 (\mathbf{I} - \mathbf{W}) \tilde{\mathbf{X}}$$

where $\tilde{\mathbf{X}}$ is the augmented $\mathbf{X}$ matrix, consisting of the complete observations and the weighted filled-in observations, $\tilde{\mathbf{Y}}$ is the augmented $\mathbf{Y}$ vector, $\mathbf{W}$ is a diagonal matrix of weights, and $\hat{\mathbf{P}}$ is a diagonal matrix of estimated ‘success’ probabilities. The estimated asymptotic covariance matrix of $\hat{\beta}$ is the inverse of the above information matrix evaluated at $\hat{\beta}$.

The average estimated coefficients and their standard errors computed empirically based on 1000 simulations ($\hat{\text{sd}}_e(\hat{\beta})$) and asymptotically ($\hat{\text{sd}}_a(\hat{\beta})$) are shown in Table 1. From this table, observe that SMAR yields smaller biases, standard errors and MSE of estimated coefficients than those obtained using complete-case data, for every coefficient in the model. In addition, SMAR estimates are comparable to those obtained using the complete-case data for all $N = 500$ subjects (ALL).

Table 1: Data with eight predictors (Example 1): estimated coefficients, standard errors and MSE for complete-case data (CC, $N = 300$), SMAR data (SMAR) and complete-case data on all subjects (ALL, $N = 500$).

β	CC	SMAR	ALL	β	CC	SMAR	ALL
	$\hat{\beta}$	$\hat{\beta}$	$\hat{\beta}$		$\hat{\beta}$	$\hat{\beta}$	$\hat{\beta}$
	$\hat{\text{sd}}_e(\hat{\beta})$	$\hat{\text{sd}}_e(\hat{\beta})$	$\hat{\text{sd}}_e(\hat{\beta})$		$\hat{\text{sd}}_e(\hat{\beta})$	$\hat{\text{sd}}_e(\hat{\beta})$	$\hat{\text{sd}}_e(\hat{\beta})$
	$\hat{\text{sd}}_a(\hat{\beta})$	$\hat{\text{sd}}_a(\hat{\beta})$	$\hat{\text{sd}}_a(\hat{\beta})$		$\hat{\text{sd}}_a(\hat{\beta})$	$\hat{\text{sd}}_a(\hat{\beta})$	$\hat{\text{sd}}_a(\hat{\beta})$
	$\hat{\text{mse}}(\hat{\beta})$	$\hat{\text{mse}}(\hat{\beta})$	$\hat{\text{mse}}(\hat{\beta})$		$\hat{\text{mse}}(\hat{\beta})$	$\hat{\text{mse}}(\hat{\beta})$	$\hat{\text{mse}}(\hat{\beta})$
$\beta_0 = 1$	1.02 0.55 0.53 0.29	0.99 0.48 0.48 0.23	1.01 0.41 0.40 0.17	$\beta_5 = -1$	-1.04 0.33 0.33 0.11	-1.02 0.29 0.29 0.08	-1.03 0.25 0.25 0.06
$\beta_1 = -1$	-1.01 0.35 0.33 0.12	-0.99 0.27 0.29 0.07	-1.00 0.26 0.25 0.07	$\beta_6 = 1$	1.03 0.34 0.33 0.11	1.01 0.28 0.30 0.08	1.02 0.26 0.25 0.07
$\beta_2 = 1$	1.04 0.35 0.33 0.12	1.01 0.27 0.29 0.07	1.01 0.26 0.25 0.07	$\beta_7 = -1$	-1.02 0.34 0.33 0.11	-0.99 0.34 0.30 0.11	-1.01 0.25 0.25 0.06
$\beta_3 = -1$	-1.04 0.34 0.33 0.12	-1.02 0.28 0.29 0.08	-1.02 0.26 0.25 0.06	$\beta_8 = 1$	1.03 0.34 0.33 0.12	1.00 0.34 0.30 0.12	1.02 0.26 0.25 0.07
$\beta_4 = 1$	1.04 0.34 0.33 0.11	1.01 0.27 0.29 0.07	1.02 0.25 0.25 0.06				

Example 2: Comparison between one-stage and two-stage SMAR designs

For this example, the outcome variable Y is generated as Bernoulli with parameter $p = P(Y = 1|\mathbf{X})$, where p comes from the following model:

$$\log \frac{p}{1-p} = -1.7 + 0.3X_1 - 0.1X_2 - 0.3X_3, \quad (3.2)$$

and the covariates $X_j, j = 1, 2, 3$ are *iid* Bernoulli(0.5) random variables.

The choice of the model in (3.2) ensures that a positive ‘outcome’ (i.e. $Y = 1$) is relatively rare with $p = P(Y = 1|\mathbf{X})$ ranging from 0.05 to 0.15. Thus, we expect a large improvement in terms of estimated MSEs of the coefficients between the one-stage and two-stage SMAR designs, since the second-stage outcome-based sampling can help to significantly lessen the dis-

proportion between the numbers of ‘cases’ (i.e. $Y = 1$) and ‘controls’ (i.e. $Y = 0$) for whom complete data are available.

The one-stage SMAR data were generated based on the above model with 20%, 30%, 40%, 50%, 60% and 70% of subjects missing X_3 values. The total number of subjects was $N = 500$.

For the two-stage SMAR design, we first generated stage 1 data with 40%, 50%, 60%, 70%, 80% and 90% of subjects with missing X_3 values. Thus, the stage 1 data of the two-stage SMAR design had an extra 100 subjects with missing X_3 values compared to the one-stage SMAR data. In the second stage, 100 additional subjects were sampled and complete data were obtained for them. These additional subjects were sampled “retrospectively” based on the values of the outcome variable to ensure that in the resulting two-stage SMAR data, the complete data are available for more subjects with the rare outcome (i.e. $Y = 1$).

We compared the performance of the logistic regression model in terms of MSEs of the estimated coefficients for various proportions of subjects missing X_3 values based on (1) complete-case data (CC); (2) one-stage SMAR design; (3) two-stage SMAR design and (4) complete-case data on all $N = 500$ subjects (ALL).

The estimated MSEs of $\beta_0, \beta_1, \beta_2, \beta_3$ based on 1000 simulations are shown in Figure 1. We can see how the estimates change as the missing data fractions are varied. We observe that one-stage and two-stage SMAR yield smaller estimated MSEs for the coefficients than the complete-case analysis. In fact, the MSEs obtained using the SMAR designs are as small or almost as small as those obtained using complete-case data on all $N = 500$ subjects (ALL), especially for coefficients for which the corresponding variables had no missing values (e.g. β_1 and β_2). Additionally, the two-stage SMAR design yields more accurate estimates than the one-stage SMAR design for β_3 , which is the coefficient of the incompletely observed covariate.

Example 3: GLM with gamma link

In this example, we consider a GLM with gamma errors. The outcome variable Y was generated from the gamma distribution with the shape parameter $\alpha = 1$ and the scale parameter $\lambda = \frac{1}{E(Y|\mathbf{X})}$ generated from the following model:

$$\lambda|\mathbf{X} = 0.8 - 0.5X_1 + 0.1X_2 - 0.2X_3, \quad (3.3)$$

where the three covariates X_j ’s are *iid* Bernoulli (0.25) random variables. The total number of subjects is $N = 200$. Variables X_1 and X_2 have no

Figure 1: Comparison of one-stage SMAR design, two-stage SMAR design, complete-case analysis (CC) and the complete-case analysis on all 500 subjects (ALL): MSE of coefficients for various proportions of missing values.

missing values. Variable X_3 has missing values for 150 randomly selected subjects. Thus, the data follows a monotone missing pattern.

Table 2 summarizes the results in terms of bias, standard error and MSE of estimated coefficients based on 1000 simulations.

The standard errors of the SMAR estimates were obtained empirically $\hat{\text{sd}}_e(\hat{\beta})$ and asymptotically. Using the method of Louis (1982), the observed information matrix for this example can be written as:

$$I(\beta) = \tilde{\mathbf{X}}^T \mathbf{W} (\tilde{\mathbf{X}}^T \beta)^{-2} \tilde{\mathbf{X}} + \tilde{\mathbf{X}}^T \mathbf{W} \left(\tilde{\mathbf{Y}} - (\tilde{\mathbf{X}}^T \beta)^{-1} \right)^2 (\mathbf{I} - \mathbf{W}) \tilde{\mathbf{X}}$$

The estimated asymptotic covariance matrix of $\hat{\beta}$ is the inverse of the above information matrix evaluated at $\hat{\beta}$.

Table 2: GLM with gamma errors (Example 3): estimated coefficients, standard errors and MSE for complete-case data (CC, $N = 50$), SMAR data and complete-case data on all subjects (ALL, $N = 200$).

Coefficient	Estimate	CC	SMAR	ALL
$\beta_0 = 0.8$	$\hat{\beta}$	0.92	0.83	0.83
	$\hat{\text{sd}}_e(\hat{\beta})$	0.30	0.15	0.13
	$\hat{\text{sd}}_a(\hat{\beta})$	0.26	0.11	0.12
	$\hat{\text{mse}}(\hat{\beta})$	0.10	0.02	0.02
$\beta_1 = -0.5$	$\hat{\beta}$	-0.57	-0.52	-0.52
	$\hat{\text{sd}}_e(\hat{\beta})$	0.27	0.12	0.11
	$\hat{\text{sd}}_a(\hat{\beta})$	0.22	0.10	0.10
	$\hat{\text{mse}}(\hat{\beta})$	0.08	0.01	0.01
$\beta_2 = 0.1$	$\hat{\beta}$	0.09	0.10	0.10
	$\hat{\text{sd}}_e(\hat{\beta})$	0.08	0.03	0.03
	$\hat{\text{sd}}_a(\hat{\beta})$	0.06	0.03	0.03
	$\hat{\text{mse}}(\hat{\beta})$	0.01	0.001	0.001
$\beta_3 = -0.2$	$\hat{\beta}$	-0.23	-0.20	-0.21
	$\hat{\text{sd}}_e(\hat{\beta})$	0.14	0.11	0.06
	$\hat{\text{sd}}_a(\hat{\beta})$	0.13	0.05	0.06
	$\hat{\text{mse}}(\hat{\beta})$	0.02	0.01	0.004

In terms of biases, standard errors and MSEs of the estimated coefficients, SMAR by far outperforms the analysis using the complete-case data only. The SMAR estimates were almost as good as the ones that could be obtained if no values were missing (ALL).

Example 4: Cost-benefit analysis

In this example, we compare the SMAR design and a complete-case design in terms of cost and efficiency. Suppose that there are six covariates of interest, $X_1, \dots, X_6$, to be measured. Furthermore, suppose that the cost of measuring each of the first four covariates is one unit per subject while the cost of measuring each of the last two covariates is ten units per subjects. Thus, the total cost of measuring all six covariates is 24 units per subject. Suppose that our resources allow us to spend 3000 units for the study. We have two choices: we can obtain complete observations for 125 subjects (complete-case design); or we can use the SMAR design. Consider the following SMAR design:

1. Measure the covariates $X_1, \dots, X_4$ for $N = 250$ subjects.

2. Take a SRS of size $N_0 = 100$ and measure the covariates X_5 and X_6 for this SRS only.

The total cost of this SMAR design is 3000 units which equals to the cost of the complete-case design with 125 subjects.

The outcome variable Y is generated as Bernoulli random variable with parameter $p = P(Y = 1|\mathbf{X})$ where p comes from the following model:

$$\log \frac{p}{1-p} = 1 - X_1 + X_2 - X_3 + X_4 - \beta_5 X_5 + \beta_6 X_6, \quad (3.4)$$

for a range of values of β_5 and β_6 , varying from 0.01 to 2 in magnitude, and the covariates X_j 's are independent Bernoulli random variables with parameter $q = (0.25, 0.75, 0.25, 0.75, 0.25, 0.75)$.

To compare the prediction accuracy of SMAR to that of the complete-case design, we generated test sets of size $N_{test} = 250$ from the same model. We used Mean Absolute Error of p ($MAE(\hat{p})$) defined as $\sum_{N_{test}} |p - \hat{p}|/N_{test}$,

to estimate the prediction accuracy of the logistic regression model based on (1) complete-case analysis of SMAR data with $N_0 = 100$ (CC_{100}), (2) complete-case design with $N = 125$ (CC_{125}) (3) the SMAR data, $N = 250$ and (4) complete-case data on all $N = 250$ subjects (ALL).

The averages and standard deviations of $MAE(\hat{p})$ over 1000 simulations are shown in Table 3 for a range of values of coefficients β_5 and β_6 varying from 0.01 to 2 in magnitude.

Table 3: Cost-benefit analysis (Example 4): averages and standard deviations of MAE over 1000 simulations for complete-case analysis of SMAR data (CC_{100}), complete-case design (CC_{125}), SMAR data and complete-case data on all subjects (ALL, $N = 250$).

β_5	β_6	CC_{100}	CC_{125}	SMAR	ALL
0.01	-0.01	0.101 (0.027)	0.089 (0.026)	0.072 (0.022)	0.062 (0.018)
0.1	-0.1	0.100 (0.028)	0.090 (0.026)	0.072 (0.022)	0.061 (0.017)
1	-1	0.096 (0.027)	0.085 (0.024)	0.073 (0.023)	0.058 (0.016)
1.5	-1.5	0.091 (0.025)	0.079 (0.023)	0.072 (0.021)	0.056 (0.016)
2	-2	0.085 (0.024)	0.076 (0.022)	0.073 (0.021)	0.052 (0.014)

The SMAR design yields smaller $MAE(\hat{p})$ than CC and NM, and the difference in $MAE(\hat{p})$ gets larger when the coefficients β_5 and β_6 decrease in magnitude.

We conclude that the SMAR design with $N_0 = 100$ is more statistically efficient than the non-missing design. Is there an optimal SMAR design for

a given cost? We compared six different SMAR designs of 3000 units in cost with N_0 (i.e. the number of subjects with no missing values) varying from 70 to 120 and complete-case design of the same cost. Listed in Table 4 are means and standard deviations of $MAE(\hat{p})$ over 1000 simulations for $\beta_5 = 1$ and $\beta_6 = -1$ in model (3.4). Notice that as N_0 varies from 70 to 120 in the SMAR design, the total number of subjects varies from 400 to 150 in order to keep the total cost of the design at 3000 units.

We observe that the SMAR design has smaller prediction errors compared to the complete-case design of the same cost (CC_{125}) over the entire range of values for N_0 (i.e. $N_0 = 70$ to 120). As N_0 gets closer to 125, the two designs become similar and so do their prediction accuracies. As N_0 gets smaller, the prediction accuracy of the SMAR design decreases.

Table 4: Cost-benefit analysis (Example 4): averages and standard deviations of MAE over 1000 simulations for for complete-case analysis of SMAR data (CC_{N_0}), complete-case design (CC_{125}), SMAR data (SMAR) and complete-case data on all subjects (ALL) generated using model 3.4 with $\beta_5 = 1$ and $\beta_6 = -1$.

N_0	N	CC_{N_0}	CC_{125}	SMAR	ALL
70	400	0.117 (0.030)	0.084 (0.024)	0.076 (0.024)	0.046 (0.012)
80	350	0.110 (0.031)	0.086 (0.024)	0.074 (0.024)	0.050 (0.015)
90	300	0.102 (0.028)	0.085 (0.025)	0.072 (0.022)	0.054 (0.016)
100	250	0.096 (0.027)	0.085 (0.024)	0.073 (0.023)	0.058 (0.016)
110	200	0.093 (0.028)	0.085 (0.026)	0.075 (0.023)	0.067 (0.019)
120	150	0.088 (0.025)	0.085 (0.026)	0.081 (0.024)	0.077 (0.022)

Example 5: Dependent predictors

In this example, the outcome variable Y is generated as Bernoulli with parameter $p = P(Y = 1|\mathbf{X})$, where p comes from the following logit model:

$$\log \frac{p}{1-p} = 1 - X_1 + 2X_2 - 2X_3, \quad (3.5)$$

the predictors X_1 and X_2 are *iid* Bernoulli (0.5) random variables, and X_3 is Bernoulli with parameter $(X_1 + X_2 + 1)/4$. Thus, the predictors are not independent. The total number of subjects is generated to be 500, of whom one hundred have missing values for the variable X_3 .

Just like in the previous examples, we compare the performance of the logistic model in terms of bias, standard error and MSE estimates of the

coefficients for complete-case data (CC, $N = 400$) and the SMAR design. We also report the estimated coefficients and their standard errors obtained using the complete-case data on all $N = 500$ subjects (ALL).

The average estimated coefficients and their standard errors computed empirically based on 1000 simulations ($\hat{\text{sd}}_e(\hat{\beta})$) and asymptotically ($\hat{\text{sd}}_a(\hat{\beta})$) are shown in Table 5. From this table, we observe that SMAR yields smaller standard errors and MSE of estimated coefficients than those obtained using complete-case data, for every coefficient in the model. In addition, SMAR estimates of coefficients that correspond to variables with no missing values (i.e. β_1 and β_2) are comparable to those obtained using the complete data for all $N = 500$ subjects (ALL).

Table 5: Dependent predictors (Example 5): estimated coefficients, standard errors and MSE for complete-case data (CC, $N = 400$), SMAR data and complete-case data on all subjects (ALL, $N = 500$).

Coefficient	Estimate	CC	SMAR	ALL
$\beta_0 = 1$	$\hat{\beta}$	1.040	1.031	1.032
	$\hat{\text{sd}}_e(\hat{\beta})$	0.290	0.276	0.255
	$\hat{\text{sd}}_a(\hat{\beta})$	0.294	0.283	0.262
	$\text{mse}(\hat{\beta})$	0.086	0.077	0.066
$\beta_1 = -1$	$\hat{\beta}$	-1.034	-1.023	-1.024
	$\hat{\text{sd}}_e(\hat{\beta})$	0.277	0.252	0.242
	$\hat{\text{sd}}_a(\hat{\beta})$	0.283	0.246	0.252
	$\text{mse}(\hat{\beta})$	0.078	0.064	0.059
$\beta_2 = 2$	$\hat{\beta}$	2.019	2.016	2.017
	$\hat{\text{sd}}_e(\hat{\beta})$	0.270	0.240	0.233
	$\hat{\text{sd}}_a(\hat{\beta})$	0.261	0.246	0.233
	$\text{mse}(\hat{\beta})$	0.073	0.058	0.054
$\beta_3 = -2$	$\hat{\beta}$	-2.046	-2.038	-2.041
	$\hat{\text{sd}}_e(\hat{\beta})$	0.287	0.285	0.249
	$\hat{\text{sd}}_a(\hat{\beta})$	0.286	0.275	0.255
	$\text{mse}(\hat{\beta})$	0.084	0.083	0.064

Example 6: Translaryngeal intubation data example from Ibrahim (1990)

We demonstrate our method on translaryngeal intubation (TLI) data, considered by Ibrahim (1990), that involves a study of 82 patients who experienced TLI for more than four days and were prospectively evaluated for laryngeal

complications. The purpose of the study was to identify a group of patients experiencing prolonged TLI (more than four days) and to prospectively evaluate the incidence and type of laryngeal complications they might suffer. To assess this, at the entry of the study, data were collected on the patients regarding 13 continuous and dichotomous baseline explanatory variables during the period of TLI. Of the 13 covariates, 4 are incomplete. Following Ibrahim (1990), we select three covariates with the largest fraction of missing data. The outcome variable, originally with four levels, is dichotomized to 0 or 1, with 0 and 1 representing no damage and damage of the larynx at baseline, respectively.

Our goal is to obtain estimates for the model based on the SMAR algorithm and to compare these estimates to the estimates obtained by Ibrahim (1990) and to those obtained by deleting the incomplete cases, in terms of their standard errors. The data set considered by Ibrahim (1990) is shown in the left panel of Table 6. Only two observations had to be altered to make these data fit the monotone missing pattern. The altered data are displayed in the right panel of Table 6 with the altered values in red.

The estimates and their standard errors for complete cases only, the SMAR method and the method of Ibrahim (1990) are given in Table 7. The standard error estimates were computed using the results of Louis (1982). We observe that the EM method of Ibrahim (1990) and the SMAR method yield smaller estimated standard errors than those obtained using the complete case analysis.

4 Discussion

In large studies of complex human diseases it is often not possible for the investigators to include all patients in all of the ancillary translational research substudies that are part of the main study. This arises due to time and budgetary constraints and other logistical considerations. Motivated by a cancer clinical trial, in this article, we proposed the use of a Systematic Missing-At-Random (SMAR) design with planned partially missing covariates collected using a nested random sampling scheme, that enables an integrated statistical analysis across all domains of data. Using simulation studies and a published data example, we demonstrated that the proposed SMAR design can achieve high statistical and computational efficiency as well as cost effectiveness.

Table 6: Translaryngeal Intubation Data (Example 6). Left panel: original data as in Ibrahim (1990). Right panel: same data with three values (shown in red) deleted to fit monotone missing pattern.

x_1	x_2	x_3	y	<i>Frequency</i>	x_2	x_1	x_3	y	<i>Frequency</i>
-	0	0	0	1	-	-	-	1	1
0	0	0	0	10	-	-	-	1	1
0	0	1	0	10	0	-	-	1	1
0	1	0	0	6	0	-	-	0	1
0	1	-	0	1	1	0	-	0	1
0	1	1	0	2	1	1	-	0	1
1	0	1	0	7	0	0	-	1	2
1	0	0	0	1	0	1	-	1	1
1	1	-	0	1	1	1	-	1	4
-	-	-	1	1	0	0	0	0	10
-	0	-	1	1	0	0	1	0	10
0	-	0	1	1	1	0	0	0	6
0	0	0	1	9	1	0	1	0	2
0	0	1	1	10	0	1	1	0	7
0	0	-	1	2	0	1	0	0	1
0	1	0	1	2	0	0	0	1	9
0	1	1	1	7	0	0	1	1	10
1	0	0	1	3	1	0	0	1	2
1	0	-	1	1	1	0	1	1	7
1	0	1	1	1	0	1	0	1	3
1	1	-	1	4	0	1	1	1	1
1	1	0	1	1	1	1	0	1	1

4.1 Feasibility of the SMAR design

The SMAR design with nested random sampling is feasible in practice. In the LABC study, mentioned earlier in this paper, the multiple laboratory investigators have all agreed to receive samples allocated according to the SMAR design. The prospective sampling in SMAR ensures that all the parameters in the model are identifiable and allows one to evaluate safety and efficacy of treatment in cancer studies. For example, in the LABC study, the prospective design is necessary to estimate the rate of response to therapy which is an important goal of the trial. This is an advantage of the SMAR design over case-control, case-cohort and other retrospective designs.

Table 7: Comparison of SMAR method, the method of Ibrahim (1990) and complete-case analysis of Translaryngeal Intubation Data (Example 6)

	Complete-case analysis $\hat{\beta}$ (SE($\hat{\beta}$))	SMAR $\hat{\beta}$ (SE($\hat{\beta}$))	Ibrahim (1990) $\hat{\beta}$ (SE($\hat{\beta}$))
Constant	-0.16 (0.41)	-0.07 (0.40)	-0.02 (0.39)
X_1	-0.40 (0.64)	0.07 (0.54)	0.04 (0.54)
X_2	0.35 (0.56)	0.40 (0.51)	0.40 (0.50)
X_3	0.11 (0.49)	-0.03 (0.47)	-0.04 (0.46)

4.2 Variable selection

Computational efficiency of the SMAR design allows us to handle data with many variables and heavy missingness. However, many translational research studies, such as microarray studies, generate data on thousands of variables. We can handle data from these substudies by combining the variables from each substudy into a smaller set of predictor variables. For each study, a set of relevant variables or a single predictor variable can be obtained using a variable selection procedure suitable for the data. For example, the method of ‘nearest shrunken centroids’ proposed by Tibshirani et al. (2003) can be used to derive a gene expression predictor based on microarray data.

4.3 Continuous predictors

In this paper, we demonstrated the SMAR design using data analytic approaches where the covariates are either discrete or categorical. If some covariates are continuous, there are at least three possible ways to handle them: (1) Design to obtain complete observations for continuous variables; (2) Use nearest neighbor methods (or binning) to discretize (i.e. bin) the continuous variables; or (3) When some incompletely observed covariates are continuous, e.g. normally distributed, the modeling approach of Gao and Hui (1997) can be applied and the algorithm and analyses proposed in the present paper can be used to estimate the model parameters.

4.4 Comparison to other designs

The SMAR design proposed in this paper has two main advantages over the usual two-phase stratified sampling designs as well as the case-cohort and nested case-control designs. Firstly, motivated by a cancer clinical trial where the treatment response rate needs to be estimated, the SMAR design

ensures that at least some data are collected prospectively for all variables. This is done in the first stage and enables the identification of all parameters in the model, which is not possible with a retrospective design. Secondly, the nested random sampling scheme embedded in the SMAR design allows one to combine the data collected on the nested random subsamples with the data collected on the entire cohort (or clinical trial) in an integrated analysis across all domains of the data in a computationally effective way.

The proposed SMAR design can be used in cohort studies, multi-center studies as well as randomized clinical trials of treatment/intervention. An indicator variable can be used to identify the study center (for multi-center studies) and/or the treatment group (in clinical trials). Such indicator variables can be used as stratification variables in the random sampling scheme and as explanatory variables in data analysis.

The SMAR design is not limited to logistic models or GLMs, but can be used with a wide variety of statistical models, including survival and longitudinal data models. In the survival model, the censoring status, just like the outcome variable, is assumed to be completely observed by design. For the longitudinal data models, the random sampling scheme is set up at the design stage for all subjects. Thus if, according to the random sampling scheme, a certain measure is set to be collected for a subject, then all repeated measures will be collected for that subject as well; similarly, if a particular measure is set to be missing for a subject, then all repeated measures will be missing for this subject as well.

4.5 Extensions and future work

The SMAR design proposed in this paper is demonstrated using parametric models. The maximum likelihood estimates of the parameters, obtained using the EM algorithm, provide efficient estimation if the model is correctly specified. Misspecification of the parametric models can lead to inconsistent parameter estimates (Pepe and Fleming, 1991). Thus, appropriate goodness-of-fit tests and other model diagnostic methods should be applied when using parametric models for analysis. To avoid model misspecification, semi-parametric inference can be used (Tsiatis, 2006), including inverse probability weighted estimators (Lipsitz et al., 1999), conditional likelihood (Breslow and Cain, 1988), estimated likelihood (Pepe and Fleming, 1991), mean score estimators (Reilly and Pepe, 1995) and pseudoscore estimators (Chatterjee et al., 2003). Examples of semi-parametric methods that result in efficient parameter estimates include inverse probability weighted estimating equations (Robins et al., 1994) and other methods (Breslow and Holubkov,

1997; Robins et al., 1995). Analyzing data from the proposed systematic missing-at-random design using semi-parametric inference and more robust analysis is worthy further investigation which we intend to perform in the future.

5 Appendix

5.1 Initial maximum likelihood estimation of cell probabilities

Suppose we have p categorical variables $X_j, j = 1, \dots, p$ with cells $l_j = 1, \dots, L_j$. The data consists of $n^{(1)}$ observations $\{x_{i1}, i = 1, \dots, n^{(1)}\}$ with x_{i1} recorded and $x_{i2}, \dots, x_{ip}$ missing; $n^{(2)}$ observations $\{(x_{i1}, x_{i2}), i = n^{(1)} + 1, \dots, n^{(1)} + n^{(2)}\}$ with x_{i1} and x_{i2} recorded and $x_{i3}, \dots, x_{ip}$ missing; $\dots$; $n^{(p-1)}$ observations $\{(x_{i1}, \dots, x_{ip-1}), i = n^{(1)} + \dots + n^{(p-2)} + 1, \dots, n^{(1)} + \dots + n^{(p-2)} + n^{(p-1)}\}$ with $x_{i1}, \dots, x_{ip-1}$ recorded and x_{ip} missing; and $n^{(p)}$ complete observations $\{(x_{i1}, \dots, x_{ip}), i = n^{(1)} + \dots + n^{(p-1)} + 1, \dots, n^{(1)} + \dots + n^{(p-1)} + n^{(p)}\}$ with all p covariate values, $x_{i1}, \dots, x_{ip}$, recorded.

Then, using the notation introduced in Section 2, $\sum_{j=1}^p n^{(j)} = n_1$, where n_1 is the size of the first nested SRS drawn from among all the enrolled subjects; $\sum_{j=2}^p n^{(j)} = n_2$ where n_2 is the size of the second nested SRS selected from the subsample of n_1 subjects; etc. $n^{(p)} = n_p = N_0$, where N_0 is the number of subjects with complete observations.

The $n^{(p)}$ completely classified observations can be displayed in an $(L_1 \times \dots \times L_p)$ contingency table, with $n_{l_1 \dots l_p}^{(p)}$ observations in the cell with $x_{i1} = l_1, \dots, x_{ip} = l_p$. The $n^{(p-1)}$ observations can be displayed in an $(L_1 \times \dots \times L_{p-1})$ marginal contingency table, with $n_{l_1 \dots l_{p-1}}^{(p-1)}$ observations in the cell with $x_{i1} = l_1, \dots, x_{ip-1} = l_{p-1}$, etc. Finally, the $n^{(1)}$ observations can be displayed in an $(L_1 \times 1)$ marginal contingency table, with $n_{l_1}^{(1)}$ observations in the cell with $x_{i1} = l_1$. The last two of these p contingency tables are shown in Table 8.

We will use the standard plus notation for summation over subscripts $l_1, \dots, l_p$. Note that $n^{(p)} = n_{++ \dots +}^{(p)}, \dots, n^{(2)} = n_{++}^{(2)}$ and $n^{(1)} = n_{+}^{(1)}$, and $N = \sum_{k=1}^p n^{(k)}$.

There are $L_1 \times \dots \times L_p$ parameters of interest:

$$\boldsymbol{\pi} = (\pi_{11 \dots 1}, \pi_{11 \dots 12}, \dots, \pi_{L_1 \dots L_p})$$

Table 8: Two of the p contingency tables

	Levels	X_2			Totals		Levels	
	1	1	$\dots$	L_2			1	
X_1	$\vdots$	$\{n_{l_1 l_2}^{(2)}\}$			$\{n_{l_1+}^{(2)}\}$	X_1	$\vdots$	$\{n_{l_1}^{(1)}\}$
	L_1						L_1	
	Totals	$\{n_{+l_2}^{(2)}\}$			$n^{(2)}$		Totals	$n^{(1)}$

where

$$\sum_{l_1=1}^{L_1} \cdots \sum_{l_p=1}^{L_p} \pi_{l_1 \dots l_p} = \pi_{+ \dots +} = 1.$$

Following the notation of Little and Rubin (2002, Chapter 13.2), we write the likelihood of the data as follows:

$$L(\boldsymbol{\pi} | \{n_{l_1}^{(1)}, n_{l_1 l_2}^{(2)} \dots, n_{l_1 \dots l_p}^{(p)}\}) =$$

$$= \left(\prod_{l_1=1}^{L_1} \pi_{l_1+ \dots +}^{n_{l_1}^{(1)} + \dots + n_{l_1+ \dots +}^{(p)}} \right) \times \left(\prod_{l_1=1}^{L_1} \prod_{l_2=1}^{L_2} \pi_{l_2|l_1}^{n_{l_1 l_2}^{(2)} + \dots + n_{l_1 l_2+ \dots +}^{(p)}} \right) \times \cdots \times \left(\prod_{l_1=1}^{L_1} \cdots \prod_{l_p=1}^{L_p} \pi_{l_p|l_1, \dots, l_{p-1}}^{n_{l_1 l_2 \dots l_p}^{(p)}} \right),$$

where

$$\pi_{l_2|l_1} = P(X_2 = l_2 | X_1 = l_1) = \frac{\pi_{l_1 l_2+ \dots +}}{\pi_{l_1+ \dots +}},$$

$$\pi_{l_3|l_1, l_2} = P(X_3 = l_3 | X_1 = l_1, X_2 = l_2) = \frac{\pi_{l_1 l_2 l_3+ \dots +}}{\pi_{l_1 l_2+ \dots +}},$$

$$\dots$$

$$\text{and } \pi_{l_p|l_1, \dots, l_{p-1}} = P(X_p = l_p | X_1 = l_1, \dots, X_{p-1} = l_{p-1}) = \frac{\pi_{l_1 l_2 \dots l_p}}{\pi_{l_1 l_2 \dots l_{p-1}+}}.$$

Maximizing each component separately, we obtain the maximum likelihood estimates,

$$\hat{\pi}_{l_1+\dots+} = (n_{l_1}^{(1)} + n_{l_1 l_2}^{(2)} + \dots + n_{l_1+\dots+}^{(p)})/N,$$

$$\hat{\pi}_{l_2|l_1} = (n_{l_1 l_2}^{(2)} + n_{l_1 l_2+}^{(3)} + \dots + n_{l_1 l_2+\dots+}^{(p)})/(n_{l_1+}^{(2)} + n_{l_1++}^{(3)} + \dots + n_{l_1+\dots+}^{(p)}),$$

...

$$\text{and } \hat{\pi}_{l_p|l_1,\dots,l_{p-1}} = n_{l_1 l_2 \dots l_p}^{(p)} / n_{l_1 l_2 \dots l_{p-1}+}^{(p)}.$$

Multiplying, we obtain the initial estimates of cell probabilities, π , as follows:

$$\begin{aligned} \hat{\pi}_{l_1 l_2 \dots l_p}^{(0)} &= \frac{1}{N} \left(1 + \frac{n_{l_1}^{(1)}}{n_{l_1+}^{(2)} + \dots + n_{l_1+\dots+}^{(p)}} \right) \left(1 + \frac{n_{l_1 l_2}^{(2)}}{n_{l_1 l_2+}^{(3)} + \dots + n_{l_1 l_2+\dots+}^{(p)}} \right) \times \dots \\ &\dots \times \left(1 + \frac{n_{l_1 l_2 \dots l_{p-2}}^{(p-2)}}{n_{l_1 l_2 \dots l_{p-2}+}^{(p-1)} + n_{l_1 l_2 \dots l_{p-2}++}^{(p)}} \right) \left(1 + \frac{n_{l_1 \dots l_{p-1}}^{(p-1)}}{n_{l_1 \dots l_{p-1}+}^{(p)}} \right) n_{l_1 \dots l_p}^{(p)} \\ &= \frac{1}{N} n_{l_1 \dots l_p}^{(p)} \prod_{k=1}^{p-1} \left(1 + \frac{n_{l_1 l_2 \dots l_k}^{(k)}}{n_{l_1 l_2 \dots l_k+}^{(k+1)} + n_{l_1 l_2 \dots l_k+\dots+}^{(p)}} \right) \\ &= \frac{1}{N} n_{l_1 \dots l_p}^{(p)} \prod_{k=1}^{p-1} \left(1 + \frac{n_{l_1 l_2 \dots l_k}^{(k)}}{\sum_{j=k+1}^p n_{l_1 l_2 \dots l_k+\dots+}^{(j)}} \right) \end{aligned}$$

5.2 M-step update for cell probabilities

In the M-step of the EM algorithm described in Section 2.2, the estimates of the cell probabilities, π , are updated by maximizing the full data likelihood (2.3) with each incomplete observation replaced by a set of weighted “filled-in” observations. Let N' and $\{w_i^{(s)}\}_{i=1}^{N'}$ denote the total number of observations and the current set of weights for the augmented data, respectively. Note that (1) $w_i^{(s)} = 1$ for each complete observation; (2) the weights corresponding to each set of “filled-in” observations sum to 1 and (3) $\sum_{i=1}^{N'} w_i^{(s)} = N$ where N is the number of observation in the original (missing) data. Maximizing (2.3) over π at iteration s of the EM algorithm yields the following MLEs for π :

$$\hat{\pi}_{l_1 l_2 \dots l_p}^{(s)} = \frac{1}{N} \sum_{i: (x_{i1}, \dots, x_{ip}) = (l_1, \dots, l_p)} w_i^{(s)},$$

for any covariate pattern $(l_1, \dots, l_p) : 1 \leq l_j \leq L_j, 1 \leq j \leq p$, where $L_1, \dots, L_p$ are the number of categories for the covariates defined in Section 5.1.

References

- Breslow, N. E. (1996). Statistics in epidemiology: The case-control study. *Journal of the American Statistical Association*, pages 14–28.
- Breslow, N. E. and Cain, K. C. (1988). Logistic regression for two-stage case-control data. *Biometrika*, 75(1):11–20.
- Breslow, N. E. and Holubkov, R. (1997). Maximum likelihood estimation of logistic regression parameters under two-phase, outcome-dependent sampling. *Journal of the Royal Statistical Society, Series B*, 59(2):447–461.
- Carroll, R. J., Ruppert, D., and Stefanski, L. A. (1995). *Measurement Error in Nonlinear Models*. London: Chapman and Hall.
- Chatterjee, N. C., Chen, Y.-H., and Breslow, N. E. (2003). A pseudoscore estimator for regression problems with two-phase sampling. *Journal of the American Statistical Association*, 98(461):158–168.
- Chen, H. Y. (2003). A note on the prospective analysis of outcome-dependent samples. *Journal of the Royal Statistical Society, Series B*, 65(2):575–584.
- Dempster, A. P., Laird, N., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–38.
- Formenti, S. and Schneider, R. (2004). Responses, resistance, and metastasis of locally advanced breast cancer in a multiethnic cohort. *DOD Grant W81XWH-04-2-0905*.
- Gao, S. J. and Hui, S. L. (1997). Logistic regression models with missing covariate values for complex survey data. *Statistics in Medicine*, 16:2419–2428.
- Ibrahim, J. G. (1990). Incomplete data in generalized linear models. *Journal of American Statistical Association*, 85:765–769.

- Lipsitz, S. R., Ibrahim, J. G., and Zhao, L. P. (1999). A weighted estimating equation for missing covariate data with properties similar to maximum likelihood. *Journal of the American Statistical Association*, 94:1147–60.
- Little, R. A. and Rubin, D. B. (2002). *Statistical analysis with missing data*. Wiley, New York.
- Louis, T. A. (1982). Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society, Series B*, 44(2):226–233.
- Neyman, J. (1938). Contribution to the theory of sampling human populations. *Journal of American Statistical Association*, 33:101–116.
- Pepe, M. S. and Fleming, T. R. (1991). A non-parametric method for dealing with mismeasured covariate data. *Journal of the American Statistical Association*, 86:108–113.
- Prentice, R. L. and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika*, 66(3):403–411.
- Reilly, M. and Pepe, M. S. (1995). A mean score method for missing and auxiliary covariate data in regression models. *Biometrika*, 82:299–314.
- Robins, J. M., Hsieh, F., and Newey, W. (1995). Semiparametric efficient estimation of a conditional density with missing or mismeasured covariates. *Journal of the Royal Statistical Society, Series B*, 57:409–424.
- Robins, J. M., Rolnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89:846–866.
- Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63:581–592.
- Skinner, C. J., Holt, D., and Smith, T. M. F. (1989). *Analysis of Complex Surveys*. Chichester, U.K.: Wiley.
- Tibshirani, R. J., Hastie, T., Narasimhan, B., and Chu, G. (2003). Class prediction by nearest shrunken centroids, with applications to dna microarrays. *Statist. Science*, 18:104–117.
- Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data*. Springer, New York.