Group testing regression analysis with covariates and specimens subject to missingness

Aurore Delaigle[1] and Ruoxu Tan[1,2]
[1]School of Mathematics and Statistics, University of Melbourne, Parkville, VIC, 3010, Australia, aurored@unimelb.edu.au and ruoxut@outlook.com, [2]Department of Statistics and Actuarial Science, University of Hong Kong, Hong Kong.

Abstract: We develop parametric estimators of a conditional prevalence in the group testing context. Group testing is applied when a binary outcome variable, often a disease indicator, is assessed by testing a specimen for the presence of the disease. Instead of testing all individual specimens separately, these are pooled in groups and the grouped specimens are tested for the disease, which permits to significantly reduce the number of tests to be performed. Various techniques have been developed in the literature for estimating a conditional prevalence from group testing data, but most of them are not valid when the data are subject to missingness. We consider this problem in the case where the specimen and the covariates are subject to nonmonotone missingness. We propose parametric estimators of the conditional prevalence, establish identifiability conditions for a logistic missing not at random model, and introduce an ignorable missing at random model. In theory, our estimators could be applied with multiple covariates missing, but in practice, they face numerical challenges when more than one covariate is missing for given individuals. We illustrate the method on simulated data and on a dataset from the Demographics and Health Survey.

Keywords: cost saving, fast screening, prevalence estimation, pooling data.

1 Introduction

In large epidemiological studies, the group testing technique introduced by Dorfman[1] during WWII is often employed to screen more individuals in less time and use less resources. When a disease is detected through a specimen test, instead of testing each individual specimen, the specimens of groups of individuals are pooled together, and a single test is applied to the pooled specimens in each group, a technique that has been widely used during the covid-19 pandemic[2,3]. If a group tests negative, the individuals from the group are declared negative whereas a positive test result for a group indicates that at least one individual in the group may be positive. If we need to identify all infected individuals, the specimens from the positive groups are retested individually whereas prevalence estimation can be done without individual retesting[4]. This pooling strategy can significantly reduce the number of tests that need to be performed when prevalence is low[5]; when prevalence is high, too many groups

test positive for it to be useful. The ability to apply group testing can play an important role during pandemics. For example, at the start of 2022 in Australia, the PCR covid-19 testing system collapsed in many parts of the country, with many testing centers unable to process the tests in a timely manner due to the sudden massive increase of the number of people needing to be tested combined with a too high prevalence for using group testing.

Since a large part of the statistics literature on group testing is concerned with disease detection or identification, throughout we will use disease terminology. However, group testing is employed in a variety of other applications, for example to detect contaminants in food or water, to test batches of objects at once, to preserve the confidentiality of participants in a study[6], as well as for DNA screening or communication and security networks[7].

In infectious disease studies, an important quantity is the prevalence of the disease conditional on an explanatory random vector $\mathbf{X}$. There is a rich literature estimating that prevalence from group testing data; a variety of methods have been developed, including parametric[8–10] and nonparametric[11–17] techniques. Motivated by the fact that data from infectious disease studies often have missing values, Delaigle et al.[18] suggested estimators valid when the data on a subvector of $\mathbf{X}$ are missing at random (MAR) (for each individual, the same subvector is either observed or entirely missing), but all specimens are observed; under a MAR assumption, Delaigle and Tan[19] studied the case where only specimens are missing. Those works describe missingness by a univariate indicator (the same variable or vector is subject to missingness for all individuals), and only $\mathbf{X}$ or the specimen is subject to missingness, but not both; here we consider the case where both are subject to missingness.

We develop maximum likelihood estimators (MLEs) of the conditional prevalence that can in principle be applied under any type of missing assumption on $\mathbf{X}$ and the specimen, including MAR and missing not at random (MNAR), and monotone or nonmonotone missingness. Reflecting the applications of group testing where there is often no particular ordering among the missing patterns of individuals, we focus on nonmonotone missingness, where specifying a coherent and tractable missing model is often considered challenging. We establish identifiability of the popular multinomial logistic MNAR model in our context, and propose a class of ignorable MAR models. Our estimators are consistent under general missingness scenarios and with several covariates missing, but as in the non grouped case, they

face numerical difficulties when more than one covariate is missing for the same individual.

This article is organised as follows. We introduce our main models and data in Section 2. In Section 3, we develop MLEs of the conditional prevalence. In Section 4, we study the multinomial logistic MNAR model and introduce a class of ignorable MAR models. We illustrate our procedures on simulated data in Section 5 and on real data in Section 6. We discuss extensions in Section 7. The supplementary file contains technical details.

2 Model and data

We are interested in the conditional prevalence of a disease or other phenomenon,

$$p(\mathbf{x}) = P(D = 1|\mathbf{X} = \mathbf{x}) = E(D|\mathbf{X} = \mathbf{x}), \tag{2.1}$$

where $\mathbf{X} = (X_1, \ldots, X_\mathfrak{d}) \in \mathbb{R}^\mathfrak{d}$ is a random vector of explanatory variables such as age, weight or time spent at a risky activity, and D is a binary variable indicating presence ($D = 1$) or absence ($D = 0$) of the phenomenon of interest. In many situations, D cannot be directly observed and is measured imperfectly by another binary variable Y. For example, in the case of a disease, D is typically assessed through a specimen (e.g. blood, urine or swab) test whose outcome $Y = \mathbb{1}\{\text{specimen tests positive}\}$ is often error-prone, i.e. is not always equal to D. Throughout, for simplicity we will use disease terminology (e.g. test, individuals, specimen) but our methods can be applied to any phenomenon generating the data introduced below.

In large population screenings, it is often not possible to test all individuals of interest, for example because of time constrains or limited resources. Group testing is an approach that consists in pooling the individuals randomly into groups, and testing the pooled specimens of each group. Using $_{i,j}$ to refer to the ith individual from the jth group (omitting the index when referring to generic individuals), instead of observing the individual $D_{i,j}$'s or $Y_{i,j}$'s, we observe a group version that will be defined formally below. The $(\mathbf{X}_{i,j}, D_{i,j})$'s are generated by (2.1) and are assumed to be independent and identically distributed (i.i.d). Throughout, for all j we denote the size of the jth group by n_j and the sample size by N.

In practice, $\mathbf{X}$ and the specimen can be subject to missingness. For example, not all covariates are always reported and some patients may be less likely to provide their specimen depending on their age or overall health condition. When a specimen for an individual is

missing, none of the group test results contain information about the disease status D of that individual. Reflecting this, we let $R^D = \mathbb{1}\{\text{specimen is available}\}$ and $\mathbf{R}^{\mathbf{X}} = (R^{X_1}, \ldots, R^{X_\mathfrak{d}})$, with $R^{X_k} = \mathbb{1}\{X_k \text{ is observed}\}$, for $k = 1, \ldots, \mathfrak{d}$. Throughout, we use $\mathbf{T}^o$ to denote the observed (i.e. non missing) components of a random vector $\mathbf{T}$.

When we pool the specimens and some of them are missing, their missing status has an impact on estimation procedures. We need to distinguish between the case where the missing status is known before the groups are created and that where it is known after. In both cases, the test for the disease can only be applied to the pooled non missing specimens, but the way in which we create the groups differs between the two cases, so that the distribution of the group version of the test result Y differs too.

If we know the missing status before grouping, we have the possibility to group only the individuals whose specimen is observed. There, given the N missing specimen statuses $R_1^D, \ldots, R_N^D$, we pool uniformly at random the $N' = \sum_{i=1}^N R_i^D$ individuals with non missing specimen into the first J' groups of respective and potentially different sizes $n_1, \ldots, n_{J'}$, where we select J' and the n_j's as in the standard group testing setting with sample size N'. The remaining $N - N'$ individuals are not pooled, i.e. are assigned to a group of size $n_j = 1$, for $j = J' + 1, \ldots, J' + N - N'$. Recalling the notation $_{i,j}$ from above, we define the true group status of the individuals with non missing specimen from group j as

$$\widetilde{D}_j^* = \begin{cases} \max_{i=1,\ldots,n_j} D_{i,j}, & \text{for } j = 1, \ldots, J', \\ -1, & \text{for } j = J' + 1, \ldots, J' + N - N'. \end{cases} \tag{2.2}$$

Here, when $j > J'$, since no specimen is observed in group j, $\widetilde{D}_j^* = -1$ does not reflect the disease status of the group but rather indicates that it is missing.

On the other hand, if the groups are defined before collecting the specimens and cannot be redefined after collection (e.g. because of cost or experimental constraints), then we create, uniformly at random, J groups of fixed and potentially different sizes $n_1, \ldots, n_J$ using all N individuals regardless of their missing status. Thus, each group is susceptible to contain individuals with missing specimen. For $j = 1, \ldots, J$, let $I_j = \{i = 1, \ldots, n_j : R_{i,j}^D = 1\}$ denote the set of indices of individuals from group j with non missing specimen. Then only the specimens of individuals in I_j can be grouped to be tested. Letting $|I_j| = \sum_{i=1}^{n_j} R_{i,j}^D$, we

define the true group status for the individuals from group j with non missing specimen by

$$D_j^* = \begin{cases} \max_{i \in I_j} D_{i,j} & \text{if } |I_j| > 0\,, \\ -1 & \text{if } |I_j| = 0\,. \end{cases} \tag{2.3}$$

As in (2.2), we use the value -1 in (2.3) to code the fact that the disease status is missing.

As mentioned above, often we can only observe an imperfect version of each group status, which we refer to as the result of a test, following the disease terminology. That is, instead of observing $\widetilde{D}_j^*$ (resp., D_j^*), we observe the result $\widetilde{Y}_j^*$ (resp., Y_j^*) of a test performed on group j. In practice, tests are prone to two types of errors: false positive ($\widetilde{Y}_j^* = 1$ when $\widetilde{D}_j^* = 0$ or $Y_j^* = 1$ when $D_j^* = 0$) and false negative ($\widetilde{Y}_j^* = 0$ when $\widetilde{D}_j^* = 1$ or $Y_j^* = 0$ when $D_j^* = 1$). When no specimen is available in group j, i.e. when $\widetilde{D}_j^* = -1$ (resp., $D_j^* = -1$), we define $\widetilde{Y}_j^* = -1$ (resp., $Y_j^* = -1$) to indicate that no test is performed for group j; in that case there is no test error and we have $P(\widetilde{Y}_j^* = -1 | \widetilde{D}_j^* = -1) = P(Y_j^* = -1 | D_j^* = -1) = 1$.

Quite naturally, we assume throughout that the specificity $\mathrm{sp} = P(\widetilde{Y}_j^* = 0 | \widetilde{D}_j^* = 0) = P(Y_j^* = 0 | D_j^* = 0)$ and sensitivity $\mathrm{se} = P(\widetilde{Y}_j^* = 1 | \widetilde{D}_j^* = 1) = P(Y_j^* = 1 | D_j^* = 1)$ of the test are larger than 0.5. Following common practice in the group testing literature[8], we also assume that sp and se do not depend on the group sizes, which is often reasonable when the groups are not too large. Likewise, in the settings at (2.2) and (2.3), respectively, we assume that the test result depends only on the true status, i.e., for $y = 0, 1$,

$$P(\widetilde{Y}_j^* = y | \widetilde{D}_j^*, \mathbf{X}_{i,j}, \mathbf{R}_{i,j}^{\mathbf{X}}, R_{i,j}^D, i = 1, \ldots, n_j) = P(\widetilde{Y}_j^* = y | \widetilde{D}_j^*)\,, \tag{2.4}$$

$$P(Y_j^* = y | D_j^*, \mathbf{X}_{i,j}, \mathbf{R}_{i,j}^{\mathbf{X}}, R_{i,j}^D, i = 1, \ldots, n_j) = P(Y_j^* = y | D_j^*)\,. \tag{2.5}$$

3 Maximum likelihood estimators

In Sections 3.1 and 3.2, we develop MLEs of p at (2.1) for the settings at (2.2) and (2.3), respectively. We use parametric models $f_{\mathbf{X}}(\cdot;\theta)$, $p(\cdot;\gamma)$ and $f_{\mathbf{R}^{\mathbf{X}},R^D|\mathbf{X},D}(\cdot|\cdot;\phi)$, where throughout we let $f_{\mathbf{Z}}$ denote joint density or probability mass function of a generic random vector $\mathbf{Z}$. The choice of a specific missing model for $f_{\mathbf{R}^{\mathbf{X}},R^D|\mathbf{X},D}$ will be discussed in Section 4.

3.1 Regression model and estimation for the setting at (2.2)

In this section, under (2.4), we propose a MLE of p constructed from data obtained from the setting at (2.2). There, for $i = 1, \ldots, N$, the incompletely observed $(D_i, \mathbf{X}_i, \mathbf{R}_i^{\mathbf{X}}, R_i^D)$'s are i.i.d.; we observe the $(\mathbf{X}_i^o, \mathbf{R}_i^{\mathbf{X}}, R_i^D)$'s, and the group test results $\widetilde{Y}_j^*$, $j = 1, \ldots, J' + N - N'$ obtained conditionally on the R_i^D's. For $i = 1, \ldots, N$, we let $\mathbf{W}_i = (\mathbf{X}_i, \mathbf{R}_i^{\mathbf{X}}, R_i^D)$. When needed below, we reindex the variables with the double index notation $_{i,j}$ introduced earlier.

The full likelihood of $\widetilde{Y}_1^*, \ldots, \widetilde{Y}_{J'+N-N'}^*, \mathbf{W}_1, \ldots, \mathbf{W}_N$, equal to

$$f_{\widetilde{Y}_1^*,\ldots\widetilde{Y}_{J'+N-N'}^*|\mathbf{W}_1,\ldots,\mathbf{W}_N}(\widetilde{Y}_1^*, \ldots \widetilde{Y}_{J'+N-N'}^*|\mathbf{W}_1, \ldots, \mathbf{W}_N; \gamma, \phi) \times \prod_{i=1}^{N} \left\{ f_{\mathbf{R}^{\mathbf{X}},R^D|\mathbf{X}}(\mathbf{R}_i^{\mathbf{X}}, R_i^D|\mathbf{X}_i; \gamma, \phi) f_{\mathbf{X}}(\mathbf{X}_i; \theta) \right\}, \tag{3.1}$$

cannot be computed since it depends on missing data. Following the strategy used by Little and Rubin[20] in non-grouped missing data problems, we define our observed-data likelihood function by integrating (3.1) with respect to the missing components $\mathbf{X}_{i,j}^m$ of the $\mathbf{X}_{i,j}$'s. In Appendix A, we prove that the observed-data log likelihood can be written as

$$\begin{aligned}\log \mathcal{L}_1(\theta, \gamma, \phi) = \sum_{j=1}^{J'} \log \Bigg[(-1)^{\widetilde{Y}_j^*} (\mathrm{sp} + \mathrm{se} - 1) \prod_{i=1}^{n_j} \int f_0(\mathbf{W}_{i,j}; \theta, \gamma, \phi)\, d\mathbf{X}_{i,j}^m \\ + \mathrm{se}^{\widetilde{Y}_j^*} (1 - \mathrm{se})^{1-\widetilde{Y}_j^*} \prod_{i=1}^{n_j} \int \sum_{d=0}^{1} f_d(\mathbf{W}_{i,j}; \theta, \gamma, \phi)\, d\mathbf{X}_{i,j}^m \Bigg] \\ + \sum_{j=J'+1}^{J'+N-N'} \log \left\{ \int \sum_{d=0}^{1} f_d(\mathbf{W}_{1,j}; \theta, \gamma, \phi)\, d\mathbf{X}_{1,j}^m \right\},\end{aligned} \tag{3.2}$$

with $f_d(\mathbf{W}_{i,j};\theta,\gamma,\phi) = f_{\mathbf{R}^{\mathbf{X}},R^D|\mathbf{X},D}(\mathbf{R}_{i,j}^{\mathbf{X}}, R_{i,j}^D|\mathbf{X}_{i,j}, d; \phi) f_{\mathbf{X}}(\mathbf{X}_{i,j}; \theta)\{1-p(\mathbf{X}_{i,j};\gamma)\}^{1-d}\{p(\mathbf{X}_{i,j};\gamma)\}^d$.

Let $(\widehat{\theta}_1, \widehat{\gamma}_1, \widehat{\phi}_1)$ denote the argmax of (3.2). We define our estimator of $p(x)$ by $\widehat{p}_{\mathrm{MLE},1}(x) = p(x; \widehat{\gamma}_1)$. If some of the components of $\mathbf{X}$, say $\mathbf{X}_{\mathrm{F}}$, are fully observed, we can reduce the dimension of the parameters to estimate via (3.2) as long as some of the components, θ_{F} say, of θ are identifiable from $f_{\mathbf{X}_{\mathrm{F}}}$; let θ_{F^C} denote the remaining components of θ. There, we can estimate θ_{F} by $\widehat{\theta}_{\mathrm{F}}$ obtained by maximising the marginal likelihood of $\mathbf{X}_{\mathrm{F}}$. Then, we can plug $\widehat{\theta}_{\mathrm{F}}$ into (3.2) and maximise (3.2) only wrt to θ_{F^C}, γ and ϕ.

Remark 1. *This estimator is consistent under general nonmonotone missingness and standard regularity conditions for MLE[21], as long as identifiability is ensured and the group sizes*

n_j are bounded. In practice, computing it requires modelling the missing mechanism by a coherent, identifiable model that needs to be estimated, unless missingness is ignorable. We consider both situations in Section 4, where we establish identifiability of a MNAR logistic model, and introduce a class of ignorable MAR models under which the estimator simplifies considerably. As usual with parametric techniques, the practical success of our estimator is limited by the number of parameters to estimate. In the non ignorable case, the dimensions of θ, γ and ϕ can increase quickly with the number of missing covariates, making the number of parameters to estimate prohibitively large if the specimen and more than one covariate are subject to missingness, unless we can reduce the number of parameters of the model. In the ignorable case, we only need to estimate θ and γ and the computations are more tractable. However, for each individual, (3.2) involves computing a numerical integral of dimension equal to the number of covariates missing for that individual, which can be difficult to do if more than one covariate is missing for the same individual, especially in the non ignorable case where the integrals take a more complex form.

3.2 Regression model and estimation for the setting at (2.3)

In this section, we propose a MLE of p in the case where we observe the $(\mathbf{X}^o_{i,j}, Y^*_j, \mathbf{R}^{\mathbf{X}}_{i,j}, R^D_{i,j})$'s obtained from the setting at (2.3). Letting $\mathbf{W}_{i,j} = (\mathbf{X}_{i,j}, \mathbf{R}^{\mathbf{X}}_{i,j}, R^D_{i,j})$, in this case, the non observable full likelihood of the $(Y^*_j, \mathbf{W}_{i,j})$'s can be written as

$$\prod_{j=1}^{J} \left\{ f_{Y^*_j|\mathbf{W}_{1,j},\ldots,\mathbf{W}_{n_j,j}}(Y^*_j|\mathbf{W}_{1,j},\ldots,\mathbf{W}_{n_j,j};\gamma,\phi) \prod_{i=1}^{n_j} \left\{ f_{\mathbf{R}^{\mathbf{X}},R^D|\mathbf{X}}(\mathbf{R}^{\mathbf{X}}_{i,j}, R^D_{i,j}|\mathbf{X}_{i,j};\gamma,\phi) f_{\mathbf{X}}(\mathbf{X}_{i,j};\theta) \right\} \right.$$

Proceeding as in Section 3.1, we estimate (θ,γ,ϕ) by $(\widehat{\theta}_2, \widehat{\gamma}_2, \widehat{\phi}_2)$ that maximises the observed-data log likelihood function of the $(\mathbf{X}^o_{i,j}, Y^*_j, \mathbf{R}^{\mathbf{X}}_{i,j}, R^D_{i,j})$'s, obtained by taking the log of the integrated above full likelihood with respect to the missing components $\mathbf{X}^m_{i,j}$ of the $\mathbf{X}_{i,j}$'s, which, taking f_d as in Section 3.1, is equal to (see Appendix B)

$$\log \mathcal{L}(\theta,\gamma,\phi) = \sum_{j=1}^{J} \mathbb{1}(Y^*_j = -1) \sum_{i=1}^{n_j} \log \int \sum_{d=0}^{1} f_d(\mathbf{W}_{i,j};\theta,\gamma,\phi)\, d\mathbf{X}^m_{i,j}$$
$$+ \sum_{j=1}^{J} \mathbb{1}(Y^*_j \neq -1) \log \left\{ \mathrm{se}^{Y^*_j} (1-\mathrm{se})^{1-Y^*_j} \prod_{i=1}^{n_j} \int \sum_{d=0}^{1} f_d(\mathbf{W}_{i,j};\theta,\gamma,\phi)\, d\mathbf{X}^m_{i,j} \right.$$

$$+(-1)^{Y_j^*}(\mathrm{sp}+\mathrm{se}-1)\prod_{i\in I_j}\int f_0(\mathbf{W}_{i,j};\theta,\gamma,\phi)\,d\mathbf{X}_{i,j}^m\times\prod_{i\notin I_j}\int\sum_{d=0}^{1}f_d(\mathbf{W}_{i,j};\theta,\gamma,\phi)\,d\mathbf{X}_{i,j}^m\Big\}. \quad (3.3)$$

We define our estimator of $p(x)$ by $\widehat{p}_{\mathrm{MLE},2}(x)=p(x;\widehat{\gamma}_2)$. As in Section 3.1, to reduce the number of parameters to estimate by maximising (3.3), if some of the components, $\mathbf{X}_{\mathrm{F}}$, of $\mathbf{X}$ do not have missing values, and some components θ_{F} of θ are identifiable from $f_{\mathbf{X}_{\mathrm{F}}}$, we can estimate θ_{F} by maximising the marginal likelihood of $\mathbf{X}_{\mathrm{F}}$, plug the estimator in (3.3) and then maximise (3.3) only wrt to γ, ϕ and the remaining components of θ.

Remark 2. *The discussion in Remark 1 applies to this estimator. In particular, it is consistent under standard conditions and computing it requires either an explicit model, which is easier to provide in the MNAR case than in the MAR case, or an ignorable MAR model; both situations will be discussed in Section 4. Here too, good practical performance relies on the number of parameters in the model being not too large and the numerical integrals to be tractable, which makes the estimator difficult to compute if more than one covariate is missing for the same individual.*

4 Model for missing data mechanism

Our estimators for p require parametric models $f_{\mathbf{X}}(\mathbf{x};\theta)$ and $p(\mathbf{x};\gamma)$, and unless missingness is ignorable, $f_{\mathbf{R}^{\mathbf{X}},R^D|\mathbf{X},D}(\mathbf{r}^{\mathbf{X}},r^D|\mathbf{x},d;\phi)$. Specifying models for $f_{\mathbf{X}}(\mathbf{x};\theta)$ and $p(\mathbf{x};\gamma)$ is easy, e.g. the multivariate normal distribution for $f_{\mathbf{X}}(\mathbf{x};\theta)$ and the logistic regression model for $p(\mathbf{x};\gamma)$. In contrast, as in the standard non grouped case[22], specifying a coherent, identifiable missing data model for $f_{\mathbf{R}^{\mathbf{X}},R^D|\mathbf{X},D}$ can be challenging. In Section 4.1, we show that, as in the non grouped case[22], the multinomial logistic MNAR model is coherent and identifiable. We discuss MAR models in Section 4.2, where we introduce a class of ignorable missing models.

4.1 Logistic MNAR model

In this section, we show that as in the non grouped case[22], the multinomial logistic MNAR model is coherent and identifiable from our data. Of course, this popular model does not necessarily always reflect the true missing pattern of the data, but it has the advantage

of being a genuine MNAR mechanism that is computationally tractable, at least in the case where D and one covariate are missing; see Remarks 1 and 2. We will examine its practical performance as a working model in Section 5. In principle our estimators could be computed with other identifiable MNAR models, but as in the standard non grouped setting, the difficulty is to formulate explicitly such models. Another possibility is to assume an ignorable MAR model; see Section 4.2.

Let $\mathbf{R} = (\mathbf{R}^{\mathbf{X}}, R^D)$ denote the joint missing indicator and $\mathbf{1} = (1, \ldots, 1)$ be the pattern with no missing component. The multinomial logistic model assumes that

$$\begin{aligned} f_{\mathbf{R}^{\mathbf{X}}, R^D|\mathbf{X},D}(\mathbf{r}^{\mathbf{X}}, r^D|\mathbf{x}, d; \phi) &= \exp\{g_{\mathbf{r}}(\mathbf{x}, d; \phi)\}/\mathcal{D}(\mathbf{x}, d; \phi)\,, \\ f_{\mathbf{R}^{\mathbf{X}}, R^D|\mathbf{X},D}(\mathbf{1}|\mathbf{x}, d; \phi) &= 1/\mathcal{D}(\mathbf{x}, d; \phi)\,, \end{aligned} \tag{4.1}$$

with ϕ an unknown parameter vector, $g_{\mathbf{r}}$ a function associated with the pattern $\mathbf{r} = (\mathbf{r}^{\mathbf{X}}, r^D)$, and $\mathcal{D}(x, d; \phi) = 1 + \sum_{\mathbf{r}\neq\mathbf{1}} \exp\{g_{\mathbf{r}}(\mathbf{x}, d; \phi)\}$. Throughout we use $\sum_{\mathbf{r}\neq\mathbf{1}}$ to denote the sum over all possible missing data patterns different from $\mathbf{1}$.

To establish identifiability of the model, we assume that the complete pattern $\mathbf{1}$ arises with strictly positive probability. Under this assumption, in the non-grouped case, Little[23] showed that for general models, if the components $T_1, \ldots, T_d$ of a random vector $\mathbf{T} = (T_1, \ldots, T_d)$ are subject to missingness, then the joint distribution of $(\mathbf{T}, \mathbf{R}^{\mathbf{T}})$ is identified by that of $(\mathbf{T}^o, \mathbf{R}^{\mathbf{T}})$ under the so-called complete case missing value (CCMV) restriction; in the particular case of model (4.1), Tchetgen Tchetgen et al.[22] showed that the CCMV restriction is equivalent to imposing that for all $\mathbf{r} \neq \mathbf{1}$,

$$g_{\mathbf{r}}(\mathbf{x}, d; \phi) = g_{\mathbf{r}}(\mathbf{x}^o_{\mathbf{r}}, d^o_{\mathbf{r}}; \phi)\,, \tag{4.2}$$

where $(\mathbf{x}^o_{\mathbf{r}}, d^o_{\mathbf{r}})$ denotes the observed part of $(\mathbf{x}, d)$ for pattern $\mathbf{r}$. Note that although, for a given missing data pattern $\mathbf{r}$, the $g_{\mathbf{r}}$'s depend only on observed data, $\mathcal{D}$ at (4.1) depends on unobserved data, which makes the mechanism MNAR.

The next lemma establishes that (4.2) ensures identifiability also in our two group testing settings with imperfect tests defined in Section 2; see Appendix C for a proof.

Lemma 4.1. *Under model (4.1) and the identifiability condition (4.2):*
(i) in the setting at (2.2), recalling that $\widetilde{Y}_j = -1$ for $j = J'+1, \ldots, J'+N-N'$, the distribution of $(\mathbf{X}, D, \mathbf{R}^{\mathbf{X}}, R^D)$ is identifiable from those of $(\mathbf{X}^o_{1,J'+1}, \widetilde{Y}^*_{J'+1}, \mathbf{R}^{\mathbf{X}}_{1,J'+1}, R^D_{1,J'+1})$ and

$(\mathbf{X}^o_{1,j}, \ldots, \mathbf{X}^o_{n_j,j}, \widetilde{Y}^*_j, \mathbf{R}^{\mathbf{X}}_{1,j}, \ldots, \mathbf{R}^{\mathbf{X}}_{n_j,j}, R^D_{1,j}, \ldots, R^D_{n_j,j})$, *where* j *denotes any integer in* $\{1, \ldots, J'\}$; *(ii) in the setting at* (2.3), *the distribution of* $(\mathbf{X}, D, \mathbf{R}^{\mathbf{X}}, R^D)$ *is identifiable from that of* $(\mathbf{X}^o_{1,j}, \ldots, \mathbf{X}^o_{n_j,j}, Y^*_j, \mathbf{R}^{\mathbf{X}}_{1,j}, \ldots, \mathbf{R}^{\mathbf{X}}_{n_j,j}, R^D_{1,j}, \ldots, R^D_{n_j,j})$, *where* j *denotes any integer in* $\{1, \ldots, J\}$.

We deduce from this lemma that with model (4.1), our target $p(\mathbf{x}) = E(D|\mathbf{X} = \mathbf{x})$ is identified under condition (4.2) as long as the distributions in the lemma are identifiable from our data for at least one j, as well as for $J' + 1$ in case (i). The former is usually satisfied since the group sizes n_j are bounded, so that there exists at least one group j for which the number of groups of size n_j tends to infinity as $N \to \infty$. Likewise, in case (i), since $P(R^D = 0) > 0$, the number of individuals from which we can consistently estimate the distribution of group $J' + 1$ tends to infinity as $N \to \infty$.

Remark 3. *These results can be applied to the case where only* $\mathbf{X}$ *is subject to missingness. There,* $\widetilde{Y}^*_j = Y^*_j$, *the test result measuring the usual group testing disease status* $D^*_j = \max_{i=1,\ldots,n_j} D_{i,j}$, *there is no* R^D *and the estimators from Sections 3.1 and 3.2 are equal.*

4.2 MAR models

Instead of assuming a logistic MNAR model, another possibility is to assume that the data are MAR. In its most general form, a MAR model assumes that $f_{\mathbf{R}^{\mathbf{X}},R^D|\mathbf{X},D} = f_{\mathbf{R}^{\mathbf{X}},R^D|\mathbf{X}^o,D^o}$, where $D^o = D$ if $R^D = 1$ and $D^o = \emptyset$ otherwise, and the realisations of $\mathbf{R}^{\mathbf{X}}$ and R^D can differ between individuals. Under this assumption, in the standard non grouped i.i.d. case ($n_j = 1$) without test errors (sp = se = 1), it is well known that missingness is ignorable for MLEs[20]; there, computing MLEs with MAR models is much simpler than with MNAR ones: it requires neither the estimation nor the explicit specification of the MAR model. MAR is often assumed even in cases where MNAR may seem more natural, and it can give better practical performance; see e.g. Example 1.13 in Little and Rubin[20].

By contrast, in our setting, missingness is non ignorable under this general MAR assumption: $f_{\mathbf{R}^{\mathbf{X}},R^D|\mathbf{X}^o,D^o}$ cannot be factored out of the likelihood equations in Sections 3.1 and 3.2, so that we need to specify (and estimate) an explicit MAR model to apply our method. However, specifying a coherent MAR and computationally tractable model for nonmontone missing data is often considered challenging[24]. For example, Robins and Gill[25] and Sun and

Tchetgen Tchetgen[24] noted that if $\mathbf{0} = (0, \ldots, 0)$ is one of the possible patterns, the only way to satisfy the assumption that the data on $(\mathbf{X}, D)$ are MAR under the multinomial logistic model at (4.1) is to reduce it to a MCAR model. Those authors proposed MAR models that do not systematically reduce to MCAR, but they are computationally intractable, at least when combined with MLEs[24]. This makes general MAR models unattractive in our setting.

However, if we are willing to make the stronger assumption that $f_{\mathbf{R}^{\mathbf{X}}, R^D|\mathbf{X},D} = f_{\mathbf{R}^{\mathbf{X}}, R^D|\mathbf{X}^o}$, then missingness becomes ignorable in our case too. To see this, in the setting from Section 3.1 (resp., Section 3.2), let $K = J' + N - N'$ and $Y_j^\dagger = \widetilde{Y}_j^*$ (resp., $K = J$ and $Y_j^\dagger = Y_j^*$). In Appendix D, we show that if we assume that $f_{\mathbf{R}^{\mathbf{X}}, R^D|\mathbf{X},D} = f_{\mathbf{R}^{\mathbf{X}}, R^D|\mathbf{X}^o}$, the likelihoods at (3.2) and (3.3) simplify into $\ell(\theta, \gamma) + C(\phi)$, where $C(\phi) = \sum_{i=1}^N \log f_{\mathbf{R}^{\mathbf{X}}, R^D|\mathbf{X}^o}(\mathbf{R}_i^{\mathbf{X}}, R_i^D|\mathbf{X}_i^o; \phi)$ does not depend on θ or γ, and where

$$\ell(\theta,\gamma) = \sum_{j=1}^{K} \mathbb{1}(Y_j^\dagger \neq -1) \log \Big[(1-\mathrm{se})^{1-Y_j^\dagger}\, \mathrm{se}^{Y_j^\dagger} \prod_{i=1}^{n_j} f_{\mathbf{X}^o}(\mathbf{X}_{i,j}^o; \theta) + (-1)^{Y_j^\dagger} (\mathrm{sp} + \mathrm{se} - 1) \times \prod_{i=1}^{n_j} \int f_{\mathbf{X}}(\mathbf{X}_{i,j}; \theta) \{1 - p(\mathbf{X}_{i,j}; \gamma)\}^{R_{i,j}^D}\, d\mathbf{X}_{i,j}^m \Big] + \sum_{j=1}^{K} \sum_{i=1}^{n_j} \mathbb{1}(Y_j^\dagger = -1) \log f_{\mathbf{X}^o}(\mathbf{X}_{i,j}^o; \theta)\,.$$

Therefore, to estimate θ and γ it suffices to maximise $\ell(\theta, \gamma)$ without specifying or estimating the missing model. In practice, in the same way as MAR models are often used in the non grouped i.i.d. case, the simplified MAR assumption $f_{\mathbf{R}^{\mathbf{X}}, R^D|\mathbf{X},D} = f_{\mathbf{R}^{\mathbf{X}}, R^D|\mathbf{X}^o}$ could be used as a working model for computing estimators more simply than under the logistic MNAR model. Indeed, under ignorable MAR, these are easier to compute as they involve estimating less parameters; see Section 5 for practical investigation.

Remark 4. *In the case where only* $\mathbf{X}$ *is subject to missingness, the estimators from Section 3.1 and 3.2 are equal, with the adaptations discussed in Remark 3. In the general nonmonotone MAR case, missingness of* $\mathbf{X}$ *is non ignorable and we run into the difficulties discussed above, but as above, missingness is ignorable if we assume that* $f_{\mathbf{R}^{\mathbf{X}}|\mathbf{X},D} = f_{\mathbf{R}^{\mathbf{X}}|\mathbf{X}^o}$. *Delaigle et al.[18] considered the simpler case where* $\mathrm{sp} = \mathrm{se} = 1$ *and a single covariate or the same subvector* $\mathbf{V}$ *is subject to missingness for all individuals, and* $f_{\mathbf{R}^V|\mathbf{X},D} = f_{\mathbf{R}^V|\mathbf{T},D}$, *with* $\mathbf{T}$ *the components of* $\mathbf{X}$ *not subject to missingness; there they propose semi- and nonparametric estimtors of* p. *Our estimators can also be applied in the case where the specimens are*

Table 1: Simulation results for all estimators: median (interquartile range) ISE$\times 10^4$ in cases (i) and (ii), and ISE$\times 10^3$ in cases (iii) to (v), computed from 200 samples.

	Grouping	N	$\widehat{p}^{\text{LOG}}_{\text{MLE},1}$	$\widehat{p}^{\text{MISP}}_{\text{MLE},1}$	$\widehat{p}^{\text{IGN}}_{\text{MLE},1}$	$\widehat{p}^{\text{LOG}}_{\text{MLE},2}$	$\widehat{p}^{\text{MISP}}_{\text{MLE},2}$	$\widehat{p}^{\text{IGN}}_{\text{MLE},2}$	$\widehat{p}_{\text{UMLE}}$
(i)	(A)	3000	8.80 (17.55)	9.44 (19.46)	18.16 (32.08)	7.50 (18.89)	8.36 (19.66)	15.53 (26.97)	4.11 (8.65)
		6000	5.91 (11.82)	6.46 (12.01)	14.24 (19.53)	5.86 (14.81)	5.98 (14.98)	12.17 (20.65)	1.86 (3.20)
		12000	2.50 (5.67)	2.65 (6.14)	8.51 (11.45)	2.45 (4.24)	2.75 (4.59)	8.38 (10.42)	1.05 (2.38)
	(B)	2500	12.02 (21.08)	12.09 (20.89)	18.10 (30.43)	9.22 (17.25)	9.39 (17.51)	15.02 (28.45)	5.09 (9.11)
		5000	4.90 (9.40)	4.77 (10.47)	10.71 (18.25)	4.85 (9.12)	4.96 (10.22)	10.01 (14.36)	2.56 (4.36)
		10000	2.39 (5.94)	2.56 (6.21)	9.30 (12.75)	2.56 (4.95)	2.85 (4.97)	8.97 (9.76)	1.05 (2.25)
(ii)	(A)	3000	24.35 (53.29)	24.20 (54.17)	24.32 (49.79)	21.84 (38.47)	21.52 (37.88)	22.66 (36.77)	7.04 (13.45)
		6000	11.93 (24.66)	11.99 (23.72)	14.75 (25.51)	11.17 (19.66)	11.60 (18.59)	11.61 (20.56)	3.49 (5.41)
		12000	6.28 (9.67)	6.03 (9.89)	8.09 (11.04)	5.29 (7.52)	5.34 (7.29)	6.16 (8.55)	1.46 (2.44)
	(B)	2500	23.52 (53.36)	22.32 (51.06)	26.59 (45.42)	18.24 (34.92)	17.65 (33.36)	18.56 (29.20)	9.22 (11.00)
		5000	10.86 (20.38)	10.83 (18.39)	11.58 (17.70)	10.21 (16.33)	9.48 (15.43)	10.30 (16.41)	3.73 (6.56)
		10000	6.42 (9.17)	6.11 (8.78)	6.54 (11.07)	4.66 (6.08)	4.25 (6.10)	5.40 (7.49)	2.03 (2.79)
(iii)	(A)	3000	2.55 (3.10)	2.45 (3.09)	2.79 (4.33)	2.19 (2.11)	2.12 (2.29)	2.29 (3.09)	1.03 (1.68)
		6000	1.41 (1.89)	1.38 (1.68)	1.44 (2.02)	1.29 (1.80)	1.27 (1.66)	1.09 (1.44)	0.44 (0.61)
		12000	0.53 (0.82)	0.56 (0.86)	0.62 (1.05)	0.43 (0.71)	0.45 (0.68)	0.53 (0.74)	0.18 (0.23)
	(B)	2500	2.57 (3.18)	2.36 (2.64)	2.44 (3.54)	2.17 (2.26)	2.15 (2.09)	2.42 (2.90)	1.31 (1.77)
		5000	1.26 (1.99)	1.27 (1.93)	1.27 (1.83)	0.98 (1.51)	0.96 (1.44)	0.93 (1.60)	0.49 (0.86)
		10000	0.54 (0.85)	0.54 (0.86)	0.65 (0.91)	0.44 (0.82)	0.43 (0.81)	0.55 (0.84)	0.23 (0.32)
(iv)	(A)	3000	24.74 (40.31)	24.78 (41.44)	27.36 (45.70)	18.25 (25.30)	17.61 (23.75)	18.86 (28.45)	1.84 (2.97)
		6000	10.82 (19.23)	10.17 (20.43)	15.50 (24.21)	6.17 (10.28)	6.34 (10.32)	8.22 (13.62)	0.89 (1.25)
		12000	6.22 (7.35)	6.20 (8.35)	8.81 (11.00)	3.76 (5.54)	3.86 (5.66)	5.81 (8.67)	0.57 (0.67)
	(B)	2500	26.57 (36.03)	24.89 (36.09)	30.72 (41.89)	16.68 (20.60)	16.58 (22.12)	21.47 (30.99)	2.33 (3.38)
		5000	7.93 (12.12)	7.82 (12.98)	10.70 (18.32)	5.91 (8.05)	5.82 (7.52)	8.42 (12.17)	1.18 (1.48)
		10000	4.32 (6.81)	4.58 (6.85)	6.93 (8.57)	3.32 (4.92)	3.27 (5.02)	5.30 (7.89)	0.66 (0.71)
(v)	(A)	3000	23.40 (36.37)	70.82 (97.66)	33.68 (44.39)	22.72 (19.61)	45.88 (63.62)	20.96 (25.55)	1.99 (2.64)
		6000	12.46 (18.44)	49.92 (75.14)	17.13 (23.69)	15.70 (11.73)	27.75 (43.57)	11.76 (13.39)	1.06 (1.17)
		12000	5.60 (9.14)	31.10 (33.29)	8.45 (12.00)	11.59 (5.81)	21.23 (35.66)	7.39 (7.82)	0.52 (0.74)
	(B)	2500	21.31 (35.88)	57.90 (90.61)	31.75 (52.73)	19.65 (19.00)	35.57 (40.54)	18.84 (24.03)	2.74 (3.21)
		5000	10.02 (16.01)	42.02 (70.70)	15.85 (20.09)	13.51 (9.50)	22.34 (26.74)	10.02 (14.69)	1.10 (1.67)
		10000	4.87 (6.37)	21.65 (48.55)	7.59 (11.09)	10.62 (4.48)	15.83 (20.54)	6.58 (8.72)	0.64 (0.86)

MAR and $\mathbf{X}$ *is not subject to missingness. That setting is simpler (*$\mathbf{X}^o = \mathbf{X}$*) as even though missingness is non ignorable, the likelihoods do not involve numerical integration; there too, it is possible to construct a nonparametric estimator*[19].

5 Simulation study

We applied two versions of our estimators $\widehat{p}_{\text{MLE},1}$ and $\widehat{p}_{\text{MLE},2}$ from Section 3 to simulated data with $\mathbf{X}$ and specimen both MNAR: $\widehat{p}^{\text{LOG}}_{\text{MLE},1}$ and $\widehat{p}^{\text{LOG}}_{\text{MLE},2}$, which assume that the data

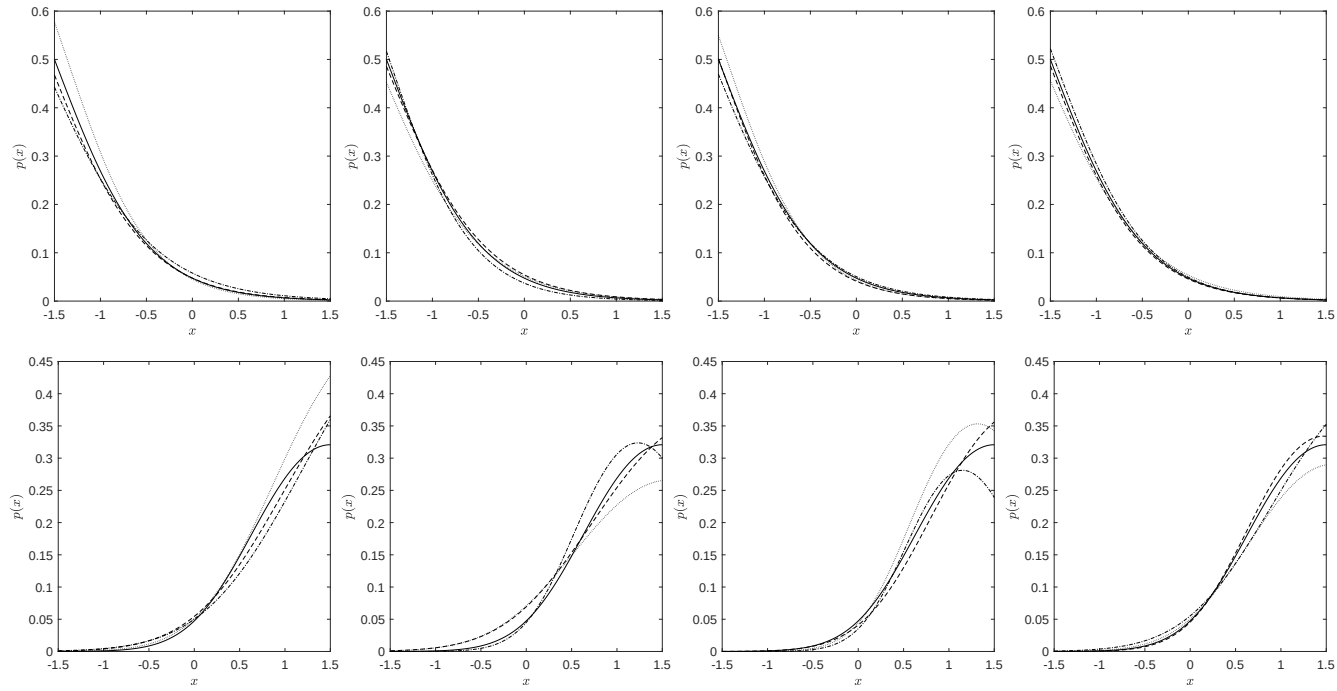

Figure 1: True curve (—), first (- - -), second (- · - · -) and third (· · ·) quartile estimated curves of, from left to right: $\widehat{p}^{\mathrm{LOG}}_{\mathrm{MLE},1}$ with $N = 3000$, $\widehat{p}^{\mathrm{LOG}}_{\mathrm{MLE},1}$, $\widehat{p}^{\mathrm{MISP}}_{\mathrm{MLE},1}$ and $\widehat{p}^{\mathrm{LOG}}_{\mathrm{MLE},2}$ with $N = 12000$ for model (i) with grouping (A) (first row), and $\widehat{p}^{\mathrm{LOG}}_{\mathrm{MLE},1}$, $\widehat{p}^{\mathrm{IGN}}_{\mathrm{MLE},2}$, $\widehat{p}^{\mathrm{LOG}}_{\mathrm{MLE},2}$ and $\widehat{p}_{\mathrm{UMLE}}$ for model (ii) with grouping (B) and $N = 5000$ (second row).

are MNAR with the logistic model from Section 4.1, and $\widehat{p}^{\mathrm{IGN}}_{\mathrm{MLE},1}$ and $\widehat{p}^{\mathrm{IGN}}_{\mathrm{MLE},2}$, which wrongly make the ignorable MAR assumption from Section 4.2, but are easier to compute because they require estimating less parameters. As noted in Remarks 1 and 2 (see also Section 7), the computations are challenging when several covariates are missing *for a given individual*. Therefore, as in the non grouped setting of Tchetgen Tchetgen et al.[22], we only consider examples where at most one covariate is missing for each individual, although in one of our examples, two covariates are subject to missingness. Performing J tests on N pooled individuals usually incurs a loss of information compared to performing N individual tests; to illustrate this, we compared our estimators with the estimator $\widehat{p}_{\mathrm{UMLE}}$ obtained from N non-grouped individuals, which is computed by taking $\widehat{p}^{\mathrm{LOG}}_{\mathrm{MLE},2}$ with N groups of size $n_j = 1$.

To generate the $(\mathbf{X}^o_{i,j}, \widetilde{Y}^*_j, \mathbf{R}^{\mathbf{X}}_{i,j}, R^D_{i,j})$'s and the $(\mathbf{X}^o_{i,j}, Y^*_j, \mathbf{R}^{\mathbf{X}}_{i,j}, R^D_{i,j})$'s, we generated the individual $(\mathbf{X}^o_i, D_i, \mathbf{R}^{\mathbf{X}}_i, R^D_i)$'s, then grouped them to obtain the $(\mathbf{X}^o_{i,j}, \mathbf{R}^{\mathbf{X}}_{i,j}, R^D_{i,j})$'s and the $\widetilde{Y}^*_j$'s and the Y^*_j's following (2.2)–(2.5). We chose $\mathrm{sp} = 0.99$, $\mathrm{se} = 0.85$, and took $\mathbf{X}$ to be of dimension one or two. In the one dimensional case, we took $\mathbf{X} = X \sim N(0, 0.75^2)$ and $D|X \sim$

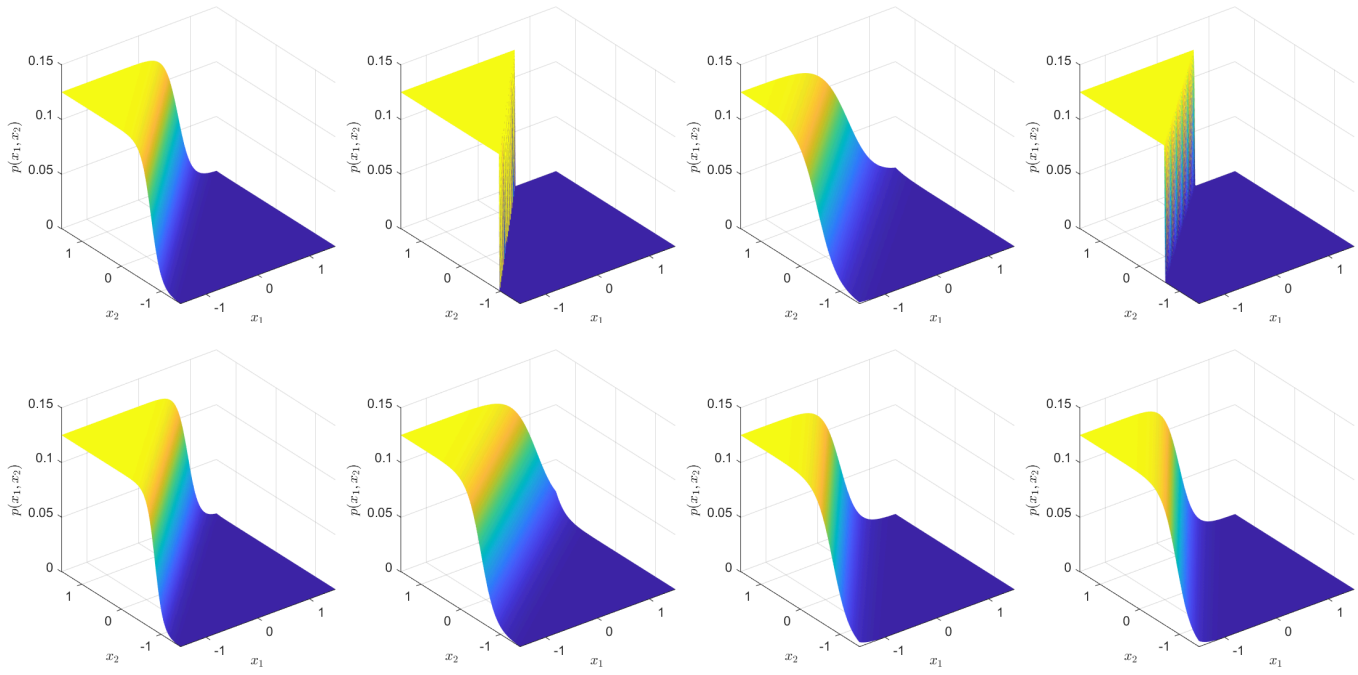

Figure 2: True p for model (iii) (top left) and, from column 2 to 4, $\widehat{p}^{\mathrm{LOG}}_{\mathrm{MLE},1}$, $\widehat{p}^{\mathrm{LOG}}_{\mathrm{MLE},2}$ and $\widehat{p}^{\mathrm{MISP}}_{\mathrm{MLE},2}$ corresponding to the median ISEs with grouping (A), and $N = 3000$ (first row) or $N = 6000$ (second row), and (bottom left) $\widehat{p}^{\mathrm{IGN}}_{\mathrm{MLE},2}$ when $N = 6000$.

Be$\{p(X)\}$, where (i) $p(x) = 1/\{1+\exp(2x+3)\}$ or (ii) $p(x) = 1/\{1+\exp(x^2-3x+3)\}$. Following (4.1), for $\mathbf{r} = (r^X, r^D) = (0,0)$, $(1,0)$ and $(0,1)$ we generated the (R_i^X, R_i^D)'s according to the model $f_{R^X,R^D|X,D}(r^X, r^D|X, D) = \exp\{g_{\mathbf{r},1}(X, D)\}/\mathcal{D}_1(X, D)$, $f_{R^X,R^D|X,D}(1,1|X, D) = 1/\mathcal{D}_1(X, D)$, where $g_{00,1}(X, D) = -3$, $g_{10,1}(X, D) = X - 1.8$, $g_{01,1}(X, D) = D - 1.5$ and $\mathcal{D}_1(X, D) = 1 + \exp\{g_{00,1}(X, D)\} + \exp\{g_{10,1}(X, D)\} + \exp\{g_{01,1}(X, D)\}$.

We considered three 2-dimensional cases, where $\mathbf{X} = (X_1, X_2) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with $\boldsymbol{\mu} = (0,0)$ and $\boldsymbol{\Sigma} = (\sigma_{ij})_{i,j=1,2}$ with $\sigma_{11} = \sigma_{22} = 0.75^2$ and $\sigma_{12} = \sigma_{21} = 0.5^2$. For the first two, we took (iii) $p(\mathbf{x}) = 1/\{8 + \exp(8x_1 - 8x_2 + 8)\}$ and (iv) $p(\mathbf{x}) = 1/\{1 + \exp(x_1^2 + 0.5x_2 + 1)\}$, and as in our real data example, we assumed that X_1 was fully observed and X_2 and D were subject to missingness. There, using R^X to denote the missing indicator of X_2, for $\mathbf{r} = (r^X, r^D) = (0,0)$, $(1,0)$ and $(0,1)$ we generated the (R_i^X, R_i^D)'s according to the model $f_{R^X,R^D|X,D}(r^X, r^D|\mathbf{X}, D) = \exp\{g_{\mathbf{r},2}(\mathbf{X}, D)/\mathcal{D}_2(\mathbf{X}, D)\}$, $f_{R^X,R^D|X,D}(1,1|\mathbf{X}, D) = 1/\mathcal{D}_2(\mathbf{X}, D)$, where $g_{00,2}(\mathbf{X}, D) = 0.5X_1 - 3$, $g_{10,2}(\mathbf{X}, D) = X_1 + X_2 - 1.8$, $g_{01,2}(\mathbf{X}, D) = -0.5X_1 + D - 1.5$, $\mathcal{D}_2(\mathbf{X}, D) = 1 + \exp\{g_{00,2}(X_1)\} + \exp\{g_{10,2}(X_1, X_2)\} + \exp\{g_{01,2}(X_1, D)\}$.

In our third 2-dimensional case (model (v)), X_1, X_2 and D were all subject to miss-

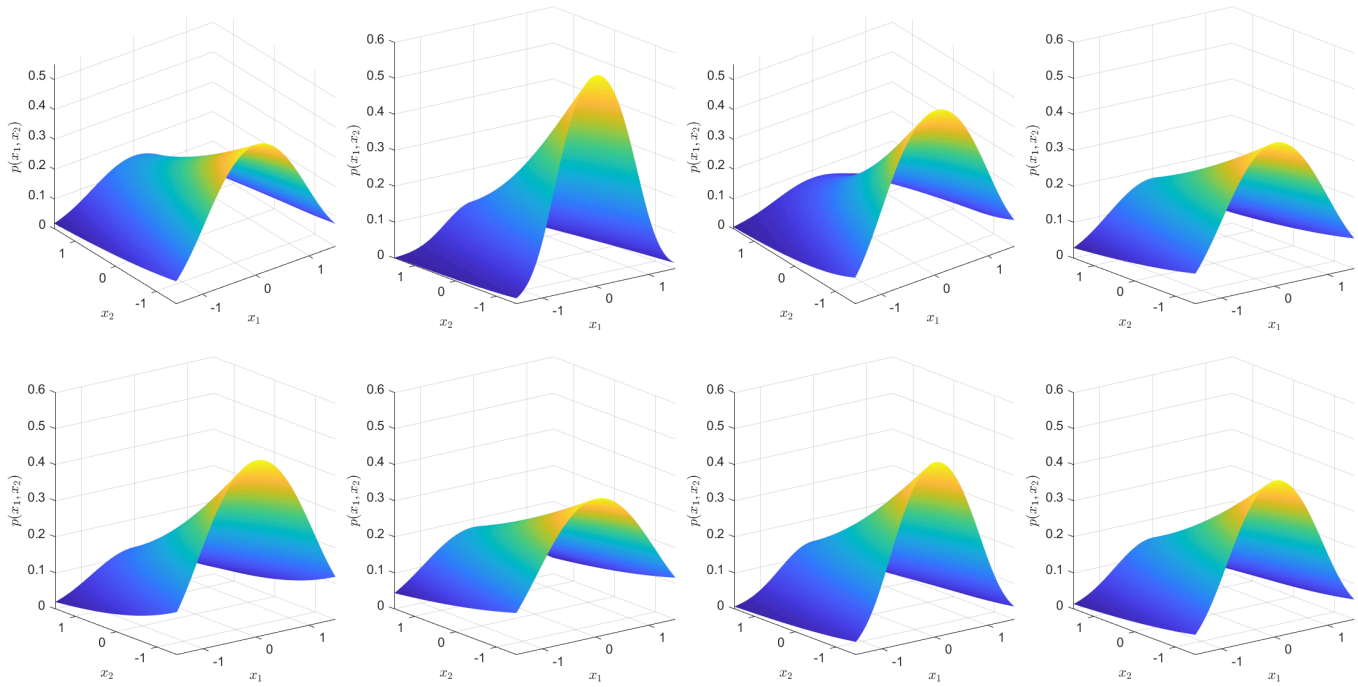

Figure 3: True p for model (iv) (top left) and, from column 2 to 4, $\widehat{p}^{\mathrm{LOG}}_{\mathrm{MLE},1}$ $\widehat{p}^{\mathrm{LOG}}_{\mathrm{MLE},2}$, $\widehat{p}_{\mathrm{UMLE}}$ corresponding to the median ISEs with grouping (B) and $N = 2500$ (first row) or $N = 5000$ (second row), and (bottom left) $\widehat{p}^{\mathrm{MISP}}_{\mathrm{MLE},2}$ with $N = 5000$.

ingness but at most one of them was missing per individual. We took $p(\mathbf{x}) = 1/\{1 + \exp(x_1^2 + 0.5x_2 + 1)\}$, and for $\mathbf{r} = (r^{X_1}, r^{X_2}, r^D) = (0,1,1), (1,0,1)$ and $(1,1,0)$, the $(R_i^{X_1}, R_i^{X_2}, R_i^D)$'s were generated according to the model $f_{R^{X_1},R^{X_2},R^D|\mathbf{X},D}(r_1^X, r_2^X, r^D|\mathbf{X}, D) = \exp\{g_{\mathbf{r},3}(\mathbf{X}, D)\}/\mathcal{D}_3(\mathbf{X}, D)$, $f_{R^{X_1},R^{X_2},R^D|\mathbf{X},D}(1,1,1|\mathbf{X}, D) = 1/\mathcal{D}_3(\mathbf{X}, D)$, where $g_{011,3}(\mathbf{X}, D) = 0.5X_2 + D - 2.5$, $g_{101,3}(\mathbf{X}, D) = X_1 + D - 1.5$, $g_{110,3}(\mathbf{X}, D) = -0.5X_1 + X_2 - 1$, $\mathcal{D}_3(\mathbf{X}, D) = 1 + \exp\{g_{011,3}(X_2, D)\} + \exp\{g_{101,3}(X_1, D)\} + \exp\{g_{110,3}(X_1, X_2)\}$.

On average, about 54% to 63% of the individuals are completely observed ($R^X = R^D = 1$) in the data generating models defined above. To assess the model sensitivity, we also computed $\widehat{p}^{\mathrm{MISP}}_{\mathrm{MLE},1}$ (resp., $\widehat{p}^{\mathrm{MISP}}_{\mathrm{MLE},2}$), our estimator $\widehat{p}_{\mathrm{MLE},1}$ (resp., $\widehat{p}_{\mathrm{MLE},2}$) with misspecified MNAR model obtained by pretending that $g_{10,1}(X, D) = \phi_1 X^2 + \phi_2$ in the univariate X case, or $g_{00,2}(\mathbf{X}, D; \phi) = \phi_1 X_1^2 + \phi_2$, $g_{10,2}(\mathbf{X}, D; \phi) = \phi_3 X_1 + \phi_4 X_2^2 + \phi_5$ and $g_{011,3}(\mathbf{X}, D; \phi) = \phi_1 X_2^2 + \phi_2 D + \phi_3$ in the bivariate case (we correctly specified the other functions). For all the estimators of p that we computed in models (iii) and (iv), since X_1 was fully observed, we estimated the mean and variance of X_1 by the empirical mean and variance of X_1 and only optimized the log likelihood with respect to the remaining parameters, as discussed in Sec-

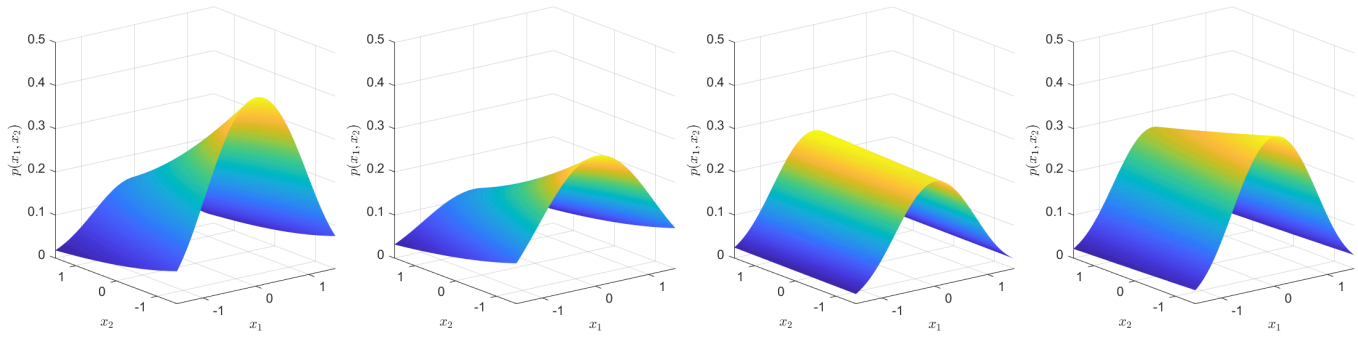

Figure 4: From left to right: true p for model (v) and $\widehat{p}^{\,\mathrm{LOG}}_{\mathrm{MLE},2}$, $\widehat{p}^{\,\mathrm{MISP}}_{\mathrm{MLE},2}$ and $\widehat{p}^{\,\mathrm{IGN}}_{\mathrm{MLE},2}$ corresponding to the median ISEs with grouping (A) and $N = 3000$.

tion 3. We computed all estimators using the fminunc function in MATLAB, where as initial values, we took the empirical mean and variance of $\mathbf{X}|\mathbf{R}^{\mathbf{X}} = 1$ and empirical (co)variance of $\mathbf{X}|\mathbf{R}^{\mathbf{X}} = 1$ for the parameters of $f_{\mathbf{X}}$ and set the other parameters to zero.

For the setting at (2.3), we considered $J = 500$, 1000 and 2000 groups of sizes n_j chosen in two ways: (A) $J/2$ groups of size $n_j = 4$ and $J/2$ groups of size $n_j = 8$; (B) J groups of size $n_j = 5$. These correspond to samples of size $N = 3000$, 6000 and 12000 in case (A) and $N = 2500$, 5000 and 10000 in case (B); for $\widehat{p}^{\,\mathrm{LOG}}_{\mathrm{MLE},1}$ and $\widehat{p}^{\,\mathrm{MISP}}_{\mathrm{MLE},1}$ we took the same n_j's but replaced J by the random J' in each sample. For each model combination, we applied each method on 200 simulated samples. We summarize the results through the integrated square error, $\mathrm{ISE} = \int_{-1.5}^{1.5}\{\widehat{p}(x) - p(x)\}^2\,dx$ in the univariate case and $\mathrm{ISE} = \int_{-1.5}^{1.5}\int_{-1.5}^{1.5}\{\widehat{p}(x_1,x_1) - p(x_1,x_2)\}^2\,dx_1\,dx_2$ in the bivariate case, where $\widehat{p}$ denotes any estimator of p we computed; the probability that $\mathbf{X}$ lies in the range of integration is about 95% in the univariate case and 92% in the bivariate case.

Table 1 shows, for each estimator and all configurations, the median and interquartile range of the 200 ISEs multiplied by 10^4 in the univariate case, and by 10^3 in the bivariate case. Note that we cannot really compare the estimators from data grouped under the setting at (2.2) with those under the setting at (2.3), since they are computed from observations grouped differently. We learn from the table that, except for model (v) where two covariates are subject to missingness, the consistent $\widehat{p}^{\,\mathrm{LOG}}_{\mathrm{MLE},1}$ and $\widehat{p}^{\,\mathrm{LOG}}_{\mathrm{MLE},2}$ that assumes the right MNAR model outperformed $\widehat{p}^{\,\mathrm{IGN}}_{\mathrm{MLE},1}$ and $\widehat{p}^{\,\mathrm{IGN}}_{\mathrm{MLE},2}$ which wrongly assumes an ignorable MAR model, but the differences between them are often relatively small, suggesting that the ignorable MAR assumption can often be a reasonable working assumption. To support this further, in

Table 2: Estimators of p for the HIV infection dataset with groupings (C) and (D). The numbers shown are the median (interquartile range) ISD$\times 10^4$ computed from 200 samples.

		$\widehat{p}^{\text{LOG}}_{\text{MLE},1}$	$\widehat{p}^{\text{LOG}}_{\text{MLE},2}$	$\widehat{p}^{\text{IGN}}_{\text{MLE},1}$	$\widehat{p}^{\text{IGN}}_{\text{MLE},2}$
sp = se = 1	(C)	3.24 (4.03)	1.83 (2.42)	3.00 (3.53)	2.45 (2.89)
	(D)	0.94 (1.39)	0.66 (0.94)	1.24 (1.51)	0.75 (1.05)
sp = se = 0.99	(C)	3.36 (4.53)	2.03 (2.30)	3.27 (3.63)	2.45 (3.04)
	(D)	0.90 (1.29)	0.68 (0.95)	1.24 (1.57)	0.84 (1.01)

model (v), $\widehat{p}^{\text{IGN}}_{\text{MLE},2}$ performed better than $\widehat{p}^{\text{LOG}}_{\text{MLE},2}$. Likewise, while $\widehat{p}^{\text{LOG}}_{\text{MLE},1}$ and $\widehat{p}^{\text{LOG}}_{\text{MLE},2}$ performed very similarly to their versions $\widehat{p}^{\text{MISP}}_{\text{MLE},1}$ and $\widehat{p}^{\text{MISP}}_{\text{MLE},2}$ with misspecified MNAR mechanism in models (i) to (iv), the latter two performed much worse than the former two in model (v). Of course, $\widehat{p}_{\text{UMLE}}$ performed the best since this estimator was computed using the non-grouped observations. Since the bivariate case is more complex, estimators there require larger effective sample sizes to perform well; therefore, in that case grouping the data had more impact on the estimators than in the univariate case, especially for the smallest sample sizes we considered. Of course, in that case, the ISE is computed on a surface, so that small errors in the estimation of p accumulate along the surface, whence the large numbers in the table. Summary results for the estimators $\widehat{\gamma}$ of γ provided in Appendix E are similar to those for $\widehat{p}$, except in case (iii); there, p is less sensitive to the value of γ, so that $\widehat{\gamma}$ is a poor estimate of γ for small sample sizes, even though $\widehat{p}$ is not too far from p.

To illustrate the results visually, we show, for a few univariate settings, the true curve and three estimated curves corresponding in each setting to the samples that gave the first, second and third quartile values out of the 200 ISEs. For bivariate cases, we plot the true surfaces and the estimated surfaces corresponding in each setting to the samples giving the median ISE. The first row of Figure 1 illustrates the fact that $\widehat{p}^{\text{LOG}}_{\text{MLE},1}$ performed better with larger sample size, and it performed similarly to $\widehat{p}^{\text{MISP}}_{\text{MLE},1}$ and $\widehat{p}^{\text{LOG}}_{\text{MLE},2}$ for model (i). The second row of Figure 1 shows that $\widehat{p}^{\text{LOG}}_{\text{MLE},1}$, $\widehat{p}^{\text{LOG}}_{\text{MLE},2}$ and $\widehat{p}^{\text{IGN}}_{\text{MLE},2}$ performed reasonably well, and through the estimator $\widehat{p}_{\text{UMLE}}$, we can see the loss incurred by grouping the data. The first row of Figure 2 shows that for model (iii), the estimators were able to capture the main trend of the true surface, but the coefficients tended to be overestimated when the sample size was small. The second row illustrates a significant improved performance of all estimators as sample

size increases. Figure 3 illustrates similar properties for model (iv); there we also show the estimator $\widehat{p}_{\text{UMLE}}$ computed from non grouped data, and we see that while it performed much better than the estimators computed from grouped data for $N = 2500$, our estimators performed reasonably well for larger sample size, even with a misspecified missing model. Finally, Figure 4 shows some of the estimated surfaces for model (v) when $N = 5000$.

6 Real data application

We applied our methods from Section 3 to a dataset from the Demographics and Health Survey, which contains men's HIV and nutrition data from Zimbabwe collected from 2010 to 2011. We are interested in $p(x_1, x_2) = E(D|X_1 = x_1, X_2 = x_2)$ where D is the HIV infection status of an individual, X_1 is the log of the individual's age at the test and X_2 is the log of the individual's age at the first sexual intercourse (we took the log because X_1 and X_2 had skewed distributions with long right tails). X_1 and X_2 range from 2.71 to 3.99 and 2.40 to 3.91, respectively. The sample size is $N = 7480$; X_1 is fully observed, X_2 is missing for 2049 individuals (27.39% of the individuals), D is missing for 1435 individuals (19.18% of the individuals) and 4436 individuals (59.3% of the individuals) have fully observed (X_1, X_2, D). Since X_1 is fully observed, below we use R^X to denote the missing indicator of X_2 and model the missing data mechanism by the linear multinomial logistic model as in Section 5. In this sample, 811 individuals have positive HIV status.

As usual in real data analyses from the group testing literature, our goal was to compare our estimators based on grouped data with usual estimators based on non-grouped data, to illustrate the effect that using group testing would have on statistical analysis. As in that literature[4,10,26], in our datasets we had access to individual test results which we treated as perfect, i.e. $D_{i,j} \equiv Y_{i,j}$[27]. Then, like there, we grouped the individuals into groups of size (C) $n_j = 6$ and (D) $n_j = 3$, except for the last group which had a smaller size since $N/3 > 0$. For the setting at (2.3), this corresponds to $J = 1247$ groups in case (C) and $J = 2494$ groups in case (D). For the setting at (2.2), we grouped only the individuals with observed specimens, resulting in $J' = 1008$ groups for grouping (C) and $J' = 2015$ groups for grouping (D); there, individuals with missing specimens are seen as additional groups of size $n_j = 1$.

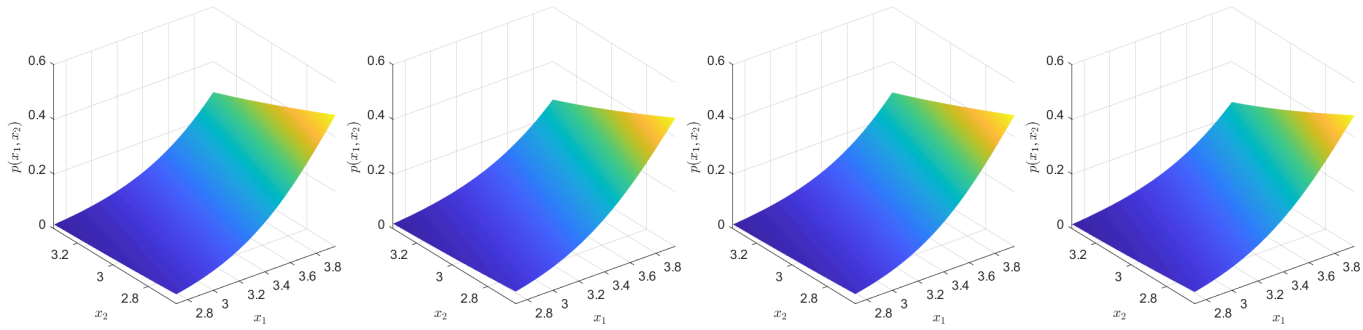

Figure 5: First row from left to right: $\widehat{p}^{\mathrm{LOG}}_{\mathrm{ideal}}$, $\widehat{p}^{\mathrm{LOG}}_{\mathrm{MLE},2}$ for grouping (D), $\widehat{p}^{\mathrm{IGN}}_{\mathrm{ideal}}$ and $\widehat{p}^{\mathrm{IGN}}_{\mathrm{MLE},2}$ for grouping (D). Here the estimated surfaces correspond to the median ISDs with sp = se = 1.

We chose different values of sp and se to generate $\widetilde{Y}^*$ and Y^* following (2.2) to (2.5): either sp = se = 1 or, to illustrate the impact of imperfect tests, sp = se = 0.99[28]. In each case, we grouped the data randomly 200 times and applied all estimators to each grouped sample. As often done in practice, we assumed that $(X_1, X_2) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and p was a logistic regression curve, i.e. $p(x_1, x_2; \gamma) = 1/\{1 + \exp(\gamma_1 x_1 + \gamma_2 x_2 + \gamma_3)\}$. We computed our estimators $\widehat{p}^{\mathrm{LOG}}_{\mathrm{MLE},1}$, $\widehat{p}^{\mathrm{LOG}}_{\mathrm{MLE},2}$, $\widehat{p}^{\mathrm{IGN}}_{\mathrm{MLE},1}$ and $\widehat{p}^{\mathrm{IGN}}_{\mathrm{MLE},2}$ as in Section 5. Since the true prevalence curve and the true missing data mechanism are unknown, we considered two targets: $\widehat{p}^{\mathrm{LOG}}_{\mathrm{ideal}} = \widehat{p}^{\mathrm{LOG}}_{\mathrm{MLE},2}$ and $\widehat{p}^{\mathrm{IGN}}_{\mathrm{ideal}} = \widehat{p}^{\mathrm{IGN}}_{\mathrm{MLE},2}$, both obtained from non grouped data with sp = se = n_j = 1, which gave $\widehat{p}^{\mathrm{LOG}}_{\mathrm{ideal}}(x_1, x_2) = 1/\{1 + \exp(-2.60x_1 + 1.18x_2 + 7.21)\}$ and $\widehat{p}^{\mathrm{IGN}}_{\mathrm{ideal}}(x_1, x_2) = 1/\{1 + \exp(-2.57x_1 + 1.19x_2 + 7.11)\}$, so that the targets are very close to each other, suggesting some robustness against missing model specification.

We evaluated the results of each estimator $\widehat{p}$ applied to grouped data by computing ISD $= \int_{a_2}^{b_2} \int_{a_1}^{b_1} \{\widehat{p}(x_1, x_2) - \widehat{p}_{\mathrm{ideal}}(x_1, x_2)\}^2 \, dx_1 \, dx_2$, where $\widehat{p}_{\mathrm{ideal}} = \widehat{p}^{\mathrm{LOG}}_{\mathrm{ideal}}$ for $\widehat{p} = \widehat{p}^{\mathrm{LOG}}_{\mathrm{MLE},1}$ and $\widehat{p}^{\mathrm{LOG}}_{\mathrm{MLE},2}$, and $\widehat{p}_{\mathrm{ideal}} = \widehat{p}^{\mathrm{IGN}}_{\mathrm{ideal}}$ for $\widehat{p} = \widehat{p}^{\mathrm{IGN}}_{\mathrm{MLE},1}$ and $\widehat{p}^{\mathrm{IGN}}_{\mathrm{MLE},2}$, $(a_1, b_1) = (2.71, 3.95)$ and $(a_2, b_2) = (2.64, 3.33)$ the $(2.5, 97.5)\%$ empirical quantiles of X_1 and $X_2 | R^X = 1$, respectively.

The results are summarised in Table 2, where we show the median and interquartile range of ISD$\times 10^4$ computed from the 200 grouped samples we randomly created. In Figure 5 we show $\widehat{p}^{\mathrm{LOG}}_{\mathrm{ideal}}$, $\widehat{p}^{\mathrm{IGN}}_{\mathrm{ideal}}$, and $\widehat{p}^{\mathrm{LOG}}_{\mathrm{MLE},2}$ and $\widehat{p}^{\mathrm{IGN}}_{\mathrm{MLE},2}$ corresponding to the median ISDs for grouping (D); see Figure 1 in Appendix F for additional plots. We see that the surfaces are close to their respective target, and close to each other, despite assuming a different missing model.

Our estimators computed from grouped data were able to capture the main trend of the target prevalence curves $\widehat{p}^{\mathrm{LOG}}_{\mathrm{ideal}}$ and $\widehat{p}^{\mathrm{IGN}}_{\mathrm{ideal}}$. In particular, they all show that HIV infection

is strongly associated to the age at the test, which can be expected since the older the individuals, the more chances they have had to be infected by HIV. We also see a negative association with age at first sexual intercourse, which corresponds to the fact that the younger the first sexual intercourse, the larger the number of years the person was exposed to HIV. This is in line with known facts about HIV in the literature[29,30].

7 Discussion

We have developed a MLE of the conditional prevalence p in a group testing setting where covariates $\mathbf{X}$ and the specimen measuring D are both subject to nonmonotone missingness. In the particular case of the multinomial logistic MNAR mechanism, we have established identifiability under the CCMV restriction. This restriction cannot be tested, but what could be done is to assess the sensitivity of the model to this assumption, using the strategy suggested by Tchetgen Tchetgen et al[22]. We have also demonstrated that under the assumption that missingness depends only on the observed part of $\mathbf{X}$, the missing model is ignorable and our estimators simplify considerably.

In theory, our methods are designed for general nonmonotone missing data. In practice, like their standard non grouped i.i.d. counterpart, they are difficult to compute if there are multiple missing covariates, especially in the MNAR case where the number of parameters to estimate can quickly become too large. We have introduced an ignorable MAR model which significantly reduces the number of parameters to estimate. However, even under this ignorable model, our estimators faces a challenge encountered in the MNAR case and in the case without grouped data: computing numerical integrals of dimension equal to the number of missing covariates. An interesting topic for future research would be to develop efficient procedures for computing these integrals numerically, such as importance sampling[31,32].

It would also be interesting to investigate ways of computing standard errors and confidence intervals. It seems very difficult to derive the explicit asymptotic variances of our estimators because of the integral over the nonmonotone missing data patterns in the likelihoods. Alternatively, we could investigate a bootstrap procedure. We leave this for future research as implementing it correctly would require to determine a consistent bootstrap

resampling approach that takes grouping, missingness, and test errors into account.

Acknowledgements

We thank Peter Braunsteins and Aihua Xia for useful discussions about our data generating process. Delaigle's research was supported by a discovery project of the Australian Research Council (ARC). Tan's research was supported by the ARC Center of Excellence for Mathematical and Statistical Frontiers and by the China Scholarship Council.

References

1. Dorfman R. The detection of defective members of large populations. *Ann Math Stat.* 1943;14(4):436–440.
2. Mallapaty S. The mathematical strategy that could transform coronavirus testing. *Nature.* 2020;583(7817):504–505.
3. Mutesa L, Ndishimye P, Butera Y, et al. A pooled testing strategy for identifying SARS-CoV-2 at low prevalence. *Nature.* 2021;589(7841):276–280.
4. Xie M. Regression analysis of group testing samples. *Stat Med.* 2001;20(13):1957–1969.
5. Bilder CR, Iwen PC, Abdalhamid B, Tebbs JM, McMahan CS. Tests in short supply? Try group testing. *Signif (Oxf).* 2020;17(3):15.
6. Gastwirth JL, Hammick PA. Estimation of the prevalence of a rare disease, preserving the anonymity of the subjects by group testing: Application to estimating the prevalence of AIDS antibodies in blood donors. *J Stat Plan Inference.* 1989;22(1):15–27.
7. Malinovsky Y, Albert PS. Revisiting nested group testing procedures: new results, comparisons, and robustness. *Am Stat.* 2019;73(2):117–125.
8. Vansteelandt S, Goetghebeur E, Verstraeten T. Regression models for disease prevalence with diagnostic tests on pools of serum samples. *Biometrics.* 2000;56(4):1126–1133.
9. Bilder CR, Tebbs JM. Bias, efficiency, and agreement for group-testing regression models. *J Stat Comput Simul.* 2009;79(1):67–80.
10. Chen P, Tebbs JM, Bilder CR. Group testing regression models with fixed and random effects. *Biometrics.* 2009;65(4):1270–1278.
11. Delaigle A, Meister A. Nonparametric regression analysis for group testing data. *J Am Stat Assoc.* 2011;106(494):640–650.
12. Delaigle A, Hall P. Nonparametric regression with homogeneous group testing data. *Ann Stat.* 2012;40(1):131–158.
13. Wang D, Zhou H, Kulasekera K. A semi-local likelihood regression estimator of the proportion based on group testing data. *J Nonparametr Stat.* 2013;25(1):209–221.

14. Delaigle A, Hall P, Wishart J. New approaches to nonparametric and semiparametric regression for univariate and multivariate group testing data. *Biometrika.* 2014;101(3):567–585.

15. Delaigle A, Hall P. Nonparametric methods for group testing data, taking dilution into account. *Biometrika.* 2015;102(4):871–887.

16. Delaigle A, Zhou WX. Nonparametric and parametric estimators of prevalence from group testing data with aggregated covariates. *J Am Stat Assoc.* 2015;110(512):1785–1796.

17. Lin J, Wang D. Single-index regression for pooled biomarker data. *J Nonparametr Stat.* 2018;30(4):813–833.

18. Delaigle A, Huang W, Lei S. Estimation of conditional prevalence from group testing data with missing covariates. *J Am Stat Assoc.* 2019;115:467-480.

19. Delaigle A, Tan, R. Group testing regression analysis with missing data and imperfect tests. *Statistica Sinica* 2022;doi:10.5705/ss.202021.0382.

20. Little RJ, Rubin DB. *Statistical analysis with missing data.* 2nd ed. John Wiley & Sons. 2002.

21. Van der Vaart AW. *Asymptotic statistics.* Cambridge University Press. 1998.

22. Tchetgen Tchetgen EJ, Wang L, Sun B. Discrete choice models for nonmonotone nonignorable missing data: Identification and inference. *Stat Sin.* 2018;28(4):2069–2088.

23. Little RJ. Pattern-mixture models for multivariate incomplete data. *J Am Stat Assoc.* 1993;88(421):125–134.

24. Sun B, Tchetgen Tchetgen EJ. On inverse probability weighting for nonmonotone missing at random data. *J Am Stat Assoc.* 2018;113(521):369–379.

25. Robins JM, Gill RD. Non-response models for the analysis of non-monotone ignorable missing data. *Stat Med.* 1997;16(1):39–56.

26. Zhang B, Bilder CR, Tebbs JM. Regression analysis for multiple-disease group testing data. *Stat Med.* 2013;32(28):4954–4966.

27. Maheu-Giroux M, Joseph L, Belisle P, Lancione S, Eaton JW, MEASURE D. Assessing the impact of imperfect immunoassays on HIV prevalence estimates from surveys conducted by the DHS Program. *DHS Methodological Reports No. 22.* 2017.

28. Galiwango RM, Musoke R, Lubyayi L, et al. Evaluation of current rapid HIV test algorithms in Rakai, Uganda. *J Virol Methods.* 2013;192(1-2):25–27.

29. Pettifor AE, Straten v. d A, Dunbar MS, Shiboski SC, Padian NS. Early age of first sex: a risk factor for HIV infection among women in Zimbabwe. *Aids.* 2004;18(10):1435–1442.

30. Bicego GT, Nkambule R, Peterson I, et al. Recent patterns in population-based HIV prevalence in Swaziland. *PLoS One.* 2013;8(10):e77101.

31. Sung YJ, Geyer CJ. Monte Carlo likelihood inference for missing data models. *Ann Stat.* 2007;35(3):990–1011.

32. Yang S, Kim, JK. Likelihood-based inference with missing data under missing-at-random. *Scand. J. Statist.*, 2016;43(2):436–454.