Statistica Sinica **34** (2024), 201-228
doi:https://doi.org/10.5705/ss.202021.0382

GROUP TESTING REGRESSION ANALYSIS WITH MISSING DATA AND IMPERFECT TESTS

Aurore Delaigle and Ruoxu Tan*

University of Melbourne

Abstract: Estimating the prevalence of an infectious disease in a big population typically requires testing a specimen (e.g., blood, urine, or swab) for the disease. When the disease spreads quickly, time constraints and limited resources often restrict the number of tests that can be performed. In such cases, if the prevalence is not too high, the group testing procedure can be employed to save time, money, and resources. The procedure tests pooled specimens of groups of individuals, rather than testing each individual for the disease. This technique is also used in other contexts, for example, to detect abnormalities or contamination in animals, plants, food, or water. Although methods exist for estimating a prevalence conditional on the explanatory variables from the group testing data, they require the specimen to be available for all individuals, which is not always possible. Therefore, we construct new nonparametric estimators that are consistent when some of the specimens are missing. We demonstrate the numerical performance of our methods using simulations and a hepatitis B example.

Key words and phrases: Cost saving, disease monitoring, limited resources, pooling, time saving.

1. Introduction

Group testing refers to a technique originally introduced by Dorfman (1943) to reduce costs and accelerate the detection process of syphilis in soldiers during WWII. The method tests groups of individuals by testing pooled specimens of the individuals from each group. If a group tests negative, the individuals from the group are declared negative. If the goal is to detect infected individuals, all individuals from the positive groups are retested; if the goal is to estimate prevalence, these individuals may or may not be retested, depending on the context (see, e.g., Xie (2001)). This technique can significantly reduce the number of tests that need to be performed, especially when prevalence is low (Bilder et al. (2020)), such as during the Covid-19 pandemic (see, e.g., Mallapaty (2020); Mutesa et al. (2021)).

While it is often used in the context of disease infection, group testing is also used to detect transgenic plants, such as transgenic corn in fields. In this case, the leaf tissues of plants are pooled, and each pool of ground tissues is tested

*Corresponding author.

(see, e.g., Montesinos-López et al. (2016)). This approach is also used to detect a contaminant (e.g., in food or water) when batches are tested at once, and to preserve the confidentiality of participants in a study (see, e.g., Gastwirth and Hammick (1989)). For other interesting applications, such as DNA screening or communication and security networks, see Malinovsky and Albert (2019).

In group testing applications, we are interested in estimating the prevalence conditional on an explanatory variable X (e.g., age). To do so, various techniques have been proposed, including parametric methods (Vansteelandt, Goetghebeur and Verstraeten (2000); Bilder and Tebbs (2009); Chen, Tebbs and Bilder (2009); Zhang, Bilder and Tebbs (2013); Lin, Wang and Zheng (2019); Chatterjee and Bandyopadhyay (2020)), nonparametric and semiparametric methods (Delaigle and Meister (2011); Delaigle and Hall (2012); Wang Zhou and Kulasekera (2013); Delaigle, Hall and Wishart (2014); Delaigle and Hall (2015); Delaigle and Zhou (2015); Lin and Wang (2018); Yuan et al. (2021)), and Bayesian methods (e.g., McMahan et al. (2017); Joyner et al. (2020); Liu et al. (2021)). However, these methods usually rely on the specimen and X being fully observed. In practice, these are sometimes missing for some individuals, and ignoring such missingness can introduce significant bias into the estimators. Delaigle, Huang and Lei (2020) developed nonparametric estimators that are valid when X is missing. In this work, we develop nonparametric consistent estimators of the conditional prevalence when individual specimens are missing.

Following Rubin (1976) and Little and Rubin (2002), we can distinguish three main types of missing mechanisms: missing completely at random (MCAR), where the missing data mechanism is independent of the variables of interest; missing at random (MAR), where the missingness depends only on the observed data; and missing not at random (MNAR), where the missingness depends also on unobserved data. With the MCAR assumption, a complete cases analysis that applies standard techniques to the fully observed individuals is usually consistent. However, this assumption is often too strong in practice. When a single variable is subject to missingness, the MAR assumption is often used to ensure identification (Little and Rubin (2002); Molenberghs et al. (2014)). There has also been growing interest in the MNAR assumption. However, in this case, to ensure identification, one usually requires additional observations, such as a validation sample (Kim and Yu (2011)), instrumental variables (Sun et al. (2018); Tchetgen Tchetgen and Wirth (2017)), or shadow variables (Miao, Tchetgen Tchetgen and Geng (2015)), which is often not possible in practice. Therefore, we develop our methodology assuming that before being grouped, the unobserved individual specimens are MAR.

The remainder of this paper is organized as follows. We introduce our model and data in Section 2, where we discuss three ways in which individual MAR specimens can affect the grouped data. In Section 3, we summarize existing nonparametric methods in the standard group testing setting. Then, we examine

the simplest MAR setting for grouped data in Section 4, where we show that, as in the nongrouped case, procedures for fully observed grouped data remain valid when some specimens are MAR before the others are pooled in groups of nonrandom size. In Section 5, we develop new nonparametric estimators of the conditional prevalence under the other two scenarios. We investigate the asymptotic properties of the proposed methods in Section 6, demonstrate the methods using simulated data in Section 7, and discuss an application in Section 8. We conclude by discussing some extensions such as the multivariate case and the use of auxiliary variables in Section 9. The online Supplementary Material contains our proofs and all technical details.

2. Model and Data

We are interested in estimating the conditional prevalence of a phenomenon

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

where X is a continuous explanatory random variable (e.g., age or weight), and D is a binary response random variable indicating the presence ($D = 1$) or absence ($D = 0$) of the phenomenon. Often, D is not observed directly and is assessed using a specimen (e.g., blood, urine, swab, or tissue) test, the outcome of which, $Y = \mathbb{1}\{\text{specimen tests positive}\}$, is typically errorprone (i.e., Y is not always equal to D).

In large population screenings, time constraints and limited resources often make it impossible to test all individuals. Note that throughout, we use individual to refer to a unit whose status D is of interest, for example a patient, a plant, or an animal. A useful approach for estimating the conditional prevalence in this case is to use group testing, where a sample of, say, N individuals is divided randomly into J groups of size $n_1, \ldots, n_J$, respectively. Using $_{i,j}$ to refer to the ith individual from the jth group (omitting the index when referring to generic individuals), we assume that $(X_{i,j}, D_{i,j})$ is independent and identically distributed (i.i.d.), where $D_{i,j}$ is the unobserved true status and $X_{i,j}$ is an observed covariate for individual i in group j. In standard group testing, instead of performing individual tests to assess $D_{i,j}$, for $j = 1, \ldots, J$, we assess the disease status

$$D^*_{\mathrm{st},j} = \max_{i=1,\ldots,n_j} D_{i,j} \tag{2.2}$$

of the jth group by testing pooled specimens of all individuals in the group, yielding the test result $Y^*_{\mathrm{st},j}$. As mentioned in the introduction, this technique is advantageous only when the overall prevalence $\theta = P(D = 1)$ in the population is relatively small, say up to 15%, or 30% if the groups are small (Kim et al. (2007); Bilder et al. (2020)). Indeed, because $P(D^*_{\mathrm{st},j} = 1) = 1 - (1 - \theta)^{n_j}$, if θ is large, we can expect most $D^*_{\mathrm{st},j}$ to be equal to one, which is not very useful or

informative; for example, if $\theta \geq 0.78$ and $n_j \geq 2$ or if $\theta \geq 0.64$ and $n_j \geq 3$, then $P(D^*_{\mathrm{st},j} = 1) > 0.95$. See also Remark 2.

In practice, specimens are not always available for all individuals. For example, in the case of a disease, some patients may be less likely to provide specimens because of their age or overall health condition, and in the case of detection in plants, some plants may die during the experiment. We let $R^D = \mathbb{1}\{\text{specimen is available}\}$ indicate whether an individual specimen is available or not. We know from the literature on nongrouped data that even in the parametric context, when a single variable is missing, the model is not generally identifiable without relatively strong identifiability assumptions; see Miao, Ding and Geng (2016). As noted in the introduction, a common approach to ensure identifiability is to assume that the missing variable is MAR. An alternative is to assume that it is MNAR, but this requires either strong additional assumptions or the availability of instrumental variables, which may not be feasible in practice (Miao, Ding and Geng (2016)). Following the first approach, we assume that the individual specimens are MAR, or equivalently, that the unobserved $D_{i,j}$ are MAR, that is,

$$P(R^D = r|X, D) = P(R^D = r|X)\,, \text{ for } r = 0 \text{ and } r = 1\,. \tag{2.3}$$

Thus, the unobserved individual $D_{i,j}$ are MAR. In particular, we do not make assumptions on their grouped versions defined below. Of course, the MAR D assumption is not always satisfied in practice, for example, when patients decide to provide their specimen based on their disease status (e.g., how well they feel). However, it is a popular approximation because it enables us to identify the model. It is also milder when more covariates are available; see Section 9 for a discussion of the multivariate case.

When some specimens are missing, only individuals with available specimens can contribute to the test performed on each group. If the missing status of all specimens is known before we start pooling the data, we can create the groups using only those individuals with nonmissing specimens. In this case, the sample size N' is random, where N' is the number of observed specimens in the original sample of size N. Given N', we fix the number of groups J' and their sizes $n_1, \ldots, n_{J'}$. For $j = 1, \ldots, J'$, we define the true status of group j as

$$\tilde{D}^*_j = \max_{i=1,\ldots,n_j} D_{i,j}|R^D_{i,j} = 1\,, \tag{2.4}$$

where $D_{i,j}$ denotes the unobserved true status of the ith individual from the jth group. Note that because we keep only individuals whose specimens are observed, $D_{i,j}$ is conditional on $R^D_{i,j} = 1$ (as is $X_{i,j}$).

If the groups are predetermined for practicality of the experiment or if it is difficult to identify missing individuals (e.g., because of confidentiality), then

only the subset of complete cases from each group contributes to the test for the group. Here, we fix the number of groups J and their respective sizes $n_1, \ldots, n_J$, and assume that the grouping is independent of the missing data mechanism. Then, for $j = 1, \ldots, J$, letting $I_j = \{i = 1, \ldots, n_j : R^D_{i,j} = 1\}$, the effective size of group j is $|I_j| = \sum_{i=1}^{n_j} R^D_{i,j}$, which is random. The true status for group j, computed from $|I_j|$ individuals, is defined as

$$D^*_j = \begin{cases} \max_{i \in I_j} D_{i,j} & |I_j| > 0\,, \\ -1 & |I_j| = 0\,. \end{cases} \tag{2.5}$$

We use the value -1 in (2.5) to code the case where D^*_j is missing, because there are no complete cases in group j.

Because the tests are usually imperfect, instead of reflecting perfectly the true group status $\tilde{D}^*_j$ (resp., D^*_j), the test result $\tilde{Y}^*_j$ (resp., Y^*_j) of group j (i.e., the result of the test applied to the nonmissing pooled specimens from group j) is prone to two types of errors: false positives, where $\tilde{Y}^*_j = 1$ when $\tilde{D}^*_j = 0$ (resp., $Y^*_j = 1$ when $D^*_j = 0$); and false negatives, where $\tilde{Y}^*_j = 0$ when $\tilde{D}^*_j = 1$ (resp., $Y^*_j = 0$ when $D^*_j = 1$). In the setting corresponding to (2.5), if no specimen is available for group j ($D^*_j = -1$), then no test is performed and we define $Y^*_j = -1$. Following Vansteelandt, Goetghebeur and Verstraeten (2000) and a large part of the literature on group testing, we assume that the known specificity $\mathrm{sp} = P(\tilde{Y}^*_j = 0|\tilde{D}^*_j = 0) = P(Y^*_j = 0|D^*_j = 0)$ and sensitivity $\mathrm{se} = P(\tilde{Y}^*_j = 1|\tilde{D}^*_j = 1) = P(Y^*_j = 1|D^*_j = 1)$ of the test do not depend on the group sizes, which is usually reasonable when the groups are not too large. Furthermore, we assume that the test results depend only on the true status. Specifically, for $y = 0, 1$,

$$P(\tilde{Y}^*_j = y|\tilde{D}^*_j, X_{i,j}, i = 1, \ldots, n_j) = P(\tilde{Y}^*_j = y|\tilde{D}^*_j) \tag{2.6}$$

in the setting in (2.4), whereas in the setting in (2.5), we assume that, for $y = 0, 1$,

$$P(Y^*_j = y|D^*_j, X_{i,j}, R^D_{i,j}, i = 1, \ldots, n_j) = P(Y^*_j = y|D^*_j)\,. \tag{2.7}$$

There is no test error when $Y^*_j = -1$, because no test is performed. In practice, sp and se are usually estimated before the test used for screening, for example using a medical diagnosis. This can be done at fast parametric rates, so that estimating sp and se has no first-order impact on the asymptotic properties of the nonparametric estimators of p; see, for example, Delaigle and Hall (2015), who derived such results in a group testing setting involving dilution effects. Because our results remain valid when sp and se are estimated, for simplicity, we assume throughout that sp and se are known. We also assume throughout that $\mathrm{sp} > 0.5$ and $\mathrm{se} > 0.5$, otherwise the test result would be less accurate than that obtained by tossing a coin.

Because the randomness of the missing specimens affects $\tilde{D}_j^*$ and D_j^* differently, these two settings require different estimation techniques. In Section 4, we show that in the first case, we can consistently estimate p by applying the technique of Delaigle, Hall and Wishart (2014) to the subset of individuals with nonmissing status. However, this estimator cannot be used in the second case, which is more widely applicable; in Section 5.1, we develop a consistent estimator that is valid in this case. In Section 5.2, we also develop a consistent estimator that can be computed even if we know how many specimens are missing from each group, but we do not know which ones are missing.

3. Review of Existing Methods Without Missing Data

In this section, we review existing local polynomial regression estimation techniques in standard settings without missing data.

3.1. Standard local polynomial estimators

In the standard setting with nongrouped data, to estimate a regression curve $g(x) = E(Y|X = x)$ from i.i.d. data $(X_1, Y_1), \ldots, (X_N, Y_N)$, a popular nonparametric estimator is the ℓth-order local polynomial regression estimator $\hat{g}_{\mathrm{LP},\ell}(x)$ (Fan and Gijbels (1996)), with $\ell \geq 0$ an integer. It is obtained by fitting, locally around x, a polynomial

$$g_\ell(z) = \sum_{0 \leq k \leq \ell} \alpha_{k,x}(z - x)^k \tag{3.1}$$

to (X_i, Y_i). It is equal to $\hat{g}_{\mathrm{LP},\ell}(x) = \hat{\alpha}_{0,x}$, where, for each x, $\hat{\alpha}_{k,x}$ is computed by minimizing the following with respect to $\alpha_{k,x}$:

$$\sum_{i=1}^{N} \left\{ Y_i - \sum_{0 \leq k \leq \ell} \alpha_{k,x}(X_i - x)^k \right\}^2 K_h(X_i - x)\,, \tag{3.2}$$

with K a kernel function, $h > 0$ a bandwidth, and $K_h(x) = h^{-1}K(x/h)$. This can be expressed as $\hat{g}_{\mathrm{LP},\ell}(x) = e_1^T \mathbf{S}^{-1}\mathbf{T}$, where $\mathbf{S} = (S_{k,k'})_{0 \leq k,k' \leq \ell}$ and $\mathbf{T} = (T_0, \ldots, T_\ell)^T$, with $S_{k,k'} = (Nh^{k+k'})^{-1} \sum_{i=1}^N K_h(X_i - x)(X_i - x)^{k+k'}$ and $T_k = (Nh^k)^{-1} \sum_{i=1}^N Y_i K_h(X_i - x)(X_i - x)^k$.

3.2. Local polynomial estimators for group testing data

In the standard group testing setting without missing data, considered by Delaigle and Meister (2011) and Delaigle, Hall and Wishart (2014), we observe $(X_{i,j}, Y_{\mathrm{st},j}^*)$, for $j = 1, \ldots, J$ and $i = 1, \ldots, n_j$, where $Y_{\mathrm{st},j}^*$ is the imperfect test result that measures the disease status $D_{\mathrm{st},j}^*$ of group j, defined in (2.2). Combining the fact that the test results depend only on the true disease status with the fact that $P(D_{\mathrm{st},j}^* = 1|X_{i,j} = x) = 1 - P(D_{i,j} = 0|X_{i,j} =$

$x)\prod_{k\neq i}P(D_{k,j}=0)=1-q^{n_j-1}\{1-p(x)\}$, where $q=P(D=0)$, and letting $Z^*_{\mathrm{st},j}=1-Y^*_{\mathrm{st},j}$, $\mathrm{sp}=P(Y^*_{\mathrm{st},j}=0|D_{\mathrm{st},j}=0)$, and $\mathrm{se}=P(Y^*_{\mathrm{st},j}=1|D_{\mathrm{st},j}=1)$, the aforementioned authors deduced that

$$g(x)=E\left\{q^{1-n_j}\frac{Z^*_{\mathrm{st},j}+\mathrm{se}-1}{\mathrm{sp}+\mathrm{se}-1}\Big|X_{i,j}=x\right\}=1-p(x)\,. \tag{3.3}$$

Similarly, $P(D^*_{\mathrm{st},j}=1)=1-q^{n_j}$, so that

$$P(Z^*_{\mathrm{st},j}=0)=1-P(Z^*_{\mathrm{st},j}=1)=\mathrm{se}-(\mathrm{sp}+\mathrm{se}-1)q^{n_j}. \tag{3.4}$$

To estimate p, they first estimated q using a maximum likelihood estimator (MLE), $\hat{q}$, obtained by maximizing, with respect to $q\in[0,1]$, the likelihood of $Z^*_{\mathrm{st},j}$:

$$\mathcal{L}(q;Z^*_{\mathrm{st},1},\ldots,Z^*_{\mathrm{st},J})=\prod_{j=1}^{J}P(Z^*_{\mathrm{st},j}=z^*_j)\,, \tag{3.5}$$

where z^*_j is the realization of $Z^*_{\mathrm{st},j}$ in the sample, and $P(Z^*_{\mathrm{st},j}=k)$, for $k=0,1$, as above.

Then, because g is a regression curve, they estimated it using the standard ℓth-order local polynomial estimator from Section 3.1, applied to the pairs $\{X_{i,j},\hat{q}^{1-n_j}(Z^*_{\mathrm{st},j}+\mathrm{se}-1)/(\mathrm{sp}+\mathrm{se}-1)\}$. Then, they added a group weight ψ_j to (3.2) that depends on the group size n_j, with the idea that larger groups blur the information more, and thus should be given less weight. Their estimator $\hat{g}_{\mathrm{st},\ell}(x)$ of $g(x)$ is obtained by fitting (3.1), locally around x, to the pairs $\{X_{i,j},\hat{q}^{1-n_j}(Z^*_{\mathrm{st},j}+\mathrm{se}-1)/(\mathrm{sp}+\mathrm{se}-1)\}$. Taking K and h as in (3.2), this is equal to $\hat{g}_{\mathrm{st},\ell}(x)=\hat{\alpha}_{0,x}$, where, for each x, $\hat{\alpha}_{k,x}$ is computed by minimizing the following with respect to $\alpha_{k,x}$:

$$\sum_{j=1}^{J}\sum_{i=1}^{n_j}\left\{\hat{q}^{1-n_j}\frac{Z^*_{\mathrm{st},j}+\mathrm{se}-1}{\mathrm{sp}+\mathrm{se}-1}-\sum_{0\le k\le\ell}\alpha_{k,x}(X_{i,j}-x)^k\right\}^2\psi_j\,K_h(X_{i,j}-x)\,. \tag{3.6}$$

This can be expressed as $\hat{g}_{\mathrm{st},\ell}(x)=e_1^T\mathbf{S}_{\mathrm{st}}^{-1}\mathbf{T}_{\mathrm{st}}$, where $\mathbf{S}_{\mathrm{st}}=(S_{\mathrm{st},k,k'})_{0\le k,k'\le\ell}$ and $\mathbf{T}_{\mathrm{st}}=(T_{\mathrm{st},0},\ldots,T_{\mathrm{st},\ell})^T$, with $S_{\mathrm{st},k,k'}=(Nh^{k+k'})^{-1}\sum_{j=1}^{J}\psi_j\sum_{i=1}^{n_j}K_h(X_{i,j}-x)(X_{i,j}-x)^{k+k'}$ and $T_{\mathrm{st},k}=(Nh^k)^{-1}\sum_{j=1}^{J}\psi_j\hat{q}^{1-n_j}(Z^*_{\mathrm{st},j}+\mathrm{se}-1)/(\mathrm{sp}+\mathrm{se}-1)\sum_{i=1}^{n_j}K_h(X_{i,j}-x)(X_{i,j}-x)^k$.

Finally, they estimated p as $\hat{p}_{\mathrm{st}}=1-\hat{g}_{\mathrm{st},\ell}$.

4. Estimator for Missing Data in the Setting at (2.4)

We start with the simplest case with missing specimens, where the groups are created after the data are collected, using only the N' individuals with nonmissing specimens out of the N individuals in the study. Here, the sample size N' is random and follows a binomial distribution $\mathrm{Bi}\{N,E(R^D)\}$. Given N', we fix the

number J' of groups and the group sizes $n_1, \ldots, n_{J'}$ such that $\sum_{j=1}^{J'} n_j = N'$, and for $i = 1, \ldots, n_j$, $j = 1, \ldots, J'$, we observe $X_{i,j}|R^D_{i,j} = 1$ and $\tilde{Y}^*_j$ defined under (2.5). To define an estimator for p in this case, a naive approach would be to apply the estimator $\hat{p}_{\text{st}}$ from Section 3.2, replacing $(X_{i,j}, Y^*_{\text{st},j})$, for $j = 1, \ldots, J$, $i = 1, \ldots, n_j$, with $(X_{i,j}|R^D_{i,j} = 1, \tilde{Y}^*_j)$, for $i = 1, \ldots, n_j$, $j = 1, \ldots, J'$, and replacing the definition of q in Section 3.2 with the quantity its MLE converges to when replacing $Z^*_{\text{st},j}$ with $\tilde{Z}^*_j = 1 - \tilde{Y}^*_j$ in (3.5). Compared with the standard setting from Section 3.2, all variables used here are defined conditional on $R^D_{i,j} = 1$ and the effective sample size is random; we need to check whether the results from Section 3.2 still hold in this case.

Recalling how $\hat{p}_{\text{st}}$ was constructed, to determine whether the naive approach is valid, we derive the relationship between $E(\tilde{Z}^*_j|X_{i,j} = x)$ and $p(x)$. Let $\tilde{Z}^*_{D,j} = 1 - \tilde{D}^*_j$, with $\tilde{D}^*_j$ given in (2.4). Using a standard decomposition (e.g., Delaigle and Meister (2011)), we show in Appendix A.2 that

$$\frac{E(\tilde{Z}^*_j + \text{se} - 1|X_{i,j} = x)}{\text{sp} + \text{se} - 1} = E(\tilde{Z}^*_{D,j}|X_{i,j} = x)\,. \tag{4.1}$$

Now, we also have

$$\begin{aligned} E(\tilde{Z}^*_{D,j}|X_{i,j} = x) &= P(\tilde{D}^*_j = 0|X_{i,j} = x) \\ &= P(D_{1,j} = \cdots = D_{n_j,j} = 0|X_{i,j} = x, R^D_{1,j} = \cdots = R^D_{n_j,j} = 1) \\ &= P(D_{i,j} = 0|X_{i,j} = x, R^D_{i,j} = 1) \prod_{k \neq i}^{n_j} P(D_{k,j} = 0|R^D_{k,j} = 1) \\ &= \{1 - p(x)\} q_{D|R}^{n_j - 1}\,, \end{aligned}$$

where $q_{D|R} = P(D = 0|R^D = 1)$, and we use $E(D|X = x, R^D = 1) = E(D|X = x) = p(x)$, which follows from (2.3). Multiplying these equations by $q_{D|R}^{1-n_j}$, we deduce that $\tilde{m}(x) \equiv E\{q_{D|R}^{1-n_j}(\tilde{Z}^*_j + \text{se} - 1)/(\text{sp} + \text{se} - 1)|X_{i,j} = x\} = 1 - p(x)$.

Here, $\tilde{m}, q_{D|R}, \tilde{Z}^*_j$, and $X_{i,j}$ satisfy the same equation as $g, q, Z^*_{\text{st},j}$, and $X_{i,j}$ do in (3.3). Similarly, we show in Appendix A.1 that $P(\tilde{Z}^*_j = 0) = 1 - P(\tilde{Z}^*_j = 1) = \text{se} - (\text{sp} + \text{se} - 1) q_{D|R}^{n_j}$, which are the same expressions as those in (3.4), but with $Z^*_{\text{st},j}$ and q replaced with $\tilde{Z}^*_j$ and $q_{D|R}$, respectively. Thus, although $q_{D|R} \neq q$ and $\tilde{Z}^*_j \neq Z^*_{\text{st},j}$, we can estimate $q_{D|R}$ by $\hat{q}_{D|R}$ obtained by applying to $\tilde{Z}^*_j$ the MLE for q from Section 3.2, that is, by maximizing $\mathcal{L}(q_{D|R}; \tilde{Z}^*_1 \ldots, \tilde{Z}^*_{J'}) = \prod_{j=1}^{J'} P(\tilde{Z}^*_j = z^*_j)$ with respect to $q_{D|R} \in [0, 1]$, and where z^*_j is the realization of $\tilde{Z}^*_j$ in the sample.

This suggests that we can estimate $p(x)$ using the "naive" estimator defined in Section 3.2, applied to the N' grouped individuals for which $R^D_{i,j} = 1$, that is,

$$\hat{p}_1(x) = 1 - \hat{\tilde{m}}(x)\,, \tag{4.2}$$

where the ℓth-order local polynomial estimator of $\tilde{m}(x)$ is given by $\hat{\tilde{m}}(x) = e_1^T \mathbf{S'}^{-1}\mathbf{T'}$, with $e_1^T = (1, 0, \ldots, 0)$, $\mathbf{S'} = (S'_{k,k'})_{0\le k,k'\le \ell}$, $\mathbf{T'} = (T'_0, \ldots, T'_\ell)^T$, and

$$\begin{aligned} S'_{k,k'} &= \frac{1}{N' h^{k+k'}} \sum_{j=1}^{J'} \psi_j \sum_{i=1}^{n_j} K_h(X_{i,j} - x)(X_{i,j} - x)^{k+k'}, \\ T'_k &= \frac{1}{N' h^k} \sum_{j=1}^{J'} \psi_j \hat{q}_{D|R}^{1-n_j} \frac{\tilde{Z}_j^* + \mathrm{se} - 1}{\mathrm{sp} + \mathrm{se} - 1} \sum_{i=1}^{n_j} K_h(X_{i,j} - x)(X_{i,j} - x)^k. \end{aligned} \tag{4.3}$$

5. Estimators for Missing Data in the Setting in (2.5)

5.1. Known individual missing status

Next, we develop a nonparametric estimator for p when the groups are determined before knowing the missing status of the specimens. We observe $(X_{i,j}, Y_j^*, R_{i,j}^D)$, for $i = 1, \ldots, n_j$, $j = 1, \ldots, J$, where $\sum_{j=1}^J n_j = N$, Y_j^* is the imperfect test result measuring D_j^* in (2.5), and $R_{i,j}^D = 1$ if the corresponding specimen is observed, and zero otherwise. Unlike Section 4, the number of tested specimens per group, that is, the effective size $|I_j| = \sum_{i=1}^{n_j} R_{i,j}^D$ of each group j is random, because we test only the subset I_j of the n_j individuals whose specimens are available.

As in Section 4, a naive way to estimate $p(x)$ would be to apply $\hat{p}_{\mathrm{st}}(x)$ from Section 3.2 to these data, but omitting the groups for which $Z_j^* = 2$, where $Z_j^* = 1 - Y_j^*$, because $\hat{p}_{\mathrm{st}}$ is defined only for $Z_{\mathrm{st},j}^* = 0$ or 1. This gives $\hat{p}_{\mathrm{nai}}(x) = 1 - \hat{g}_{\mathrm{nai},\ell}(x)$, where $\hat{g}_{\mathrm{nai},\ell}(x) = e_1^T \hat{\mathbf{S}}_{\mathrm{nai}}^{-1} \hat{\mathbf{T}}_{\mathrm{nai}}$ with $\hat{\mathbf{S}}_{\mathrm{nai}} = (S_{\mathrm{nai},k,k'})_{0\le k,k'\le \ell}$ and $\hat{\mathbf{T}}_{\mathrm{nai}} = (\hat{T}_{\mathrm{nai},0}, \ldots, \hat{T}_{\mathrm{nai},\ell})^T$, and with $\hat{S}_{\mathrm{nai},k,k'} = (Nh^{k+k'})^{-1} \sum_{j=1}^J \mathbb{1}\{Z_j^* \ne 2\}$ $\psi_j \sum_{i=1}^{n_j} K_h(X_{i,j} - x)(X_{i,j} - x)^{k+k'}$ and $\hat{T}_{\mathrm{nai},k} = (Nh^k)^{-1} \sum_{j=1}^J \mathbb{1}\{Z_j^* \ne 2\}\psi_j$ $\hat{q}_{\mathrm{nai}}^{1-n_j}(Z_j^* + \mathrm{se} - 1)/(\mathrm{sp} + \mathrm{se} - 1) \sum_{i=1}^{n_j} K_h(X_{i,j} - x)(X_{i,j} - x)^k$. Here, $\hat{q}_{\mathrm{nai}}$ is the naive estimator of q obtained by maximizing $\mathcal{L}(q; Z_1^*, \ldots, Z_J^*) = \prod_{j=1}^J P_{\mathrm{nai}}(Z_j^* = z_j^*)^{\mathbb{1}\{z_j^* \ne 2\}}$ with respect to $q \in [0,1]$, where $P_{\mathrm{nai}}(Z_j^* = 0) = 1 - P_{\mathrm{nai}}(Z_j^* = 1) = \mathrm{se} - (\mathrm{sp} + \mathrm{se} - 1)q^{n_j}$ are formulae valid for $Z_{\mathrm{st},j}^* = 0$ or 1 from Section 3.2.

However, using the derivations below, this naive estimator does not consistently estimate $p(x)$, because in this case, our data do not satisfy the same equations as those of the data from Section 3.2. For example, unlike for $\tilde{Z}_j^*$ in Section 4, here $P_{\mathrm{nai}}(Z_j^* = 0)$ and $P_{\mathrm{nai}}(Z_j^* = 1)$ are not valid for Z_j^*. To derive a consistent estimator for p, we need to express p in terms of a regression curve that depends only on the observed data. Then, we can estimate that regression curve using a standard local polynomial estimator.

Mimicking the derivations in the standard case from Section 3.2, another approach would be to express $E(Z_j^* | X_{i,j} = x)$ in terms of $p(x)$. However, using the results from Appendix A.3, it can be proved that $E(Z_j^* | X_{i,j} = x) = q_{RD}^{n_j - 1}\{1 - b(x)\}(\mathrm{sp} + \mathrm{se} - 1) + q_R^{n_j - 1}\{1 - d(x)\}(2 - \mathrm{sp}) + 1 - \mathrm{se}$, where $b(x) = E(R_{i,j}^D D_{i,j} | X_{i,j} =$

$x)$, $d(x) = E(R^D_{i,j}|X_{i,j} = x)$, $q_{RD} = P(R^D D = 0)$, and $q_R = P(R^D = 0)$, which does not seem helpful for estimating $p(x)$. Instead, our idea is to condition also on the missing status. Using this approach and the same standard decomposition as in Section 4 (see Appendix A.2), we first express the test results in terms of D^*_j by writing, for all $i \in I_j$,

$$\frac{E(Z^*_j + \mathrm{se} - 1|X_{i,j} = x, R^D_{i,j} = 1)}{\mathrm{sp} + \mathrm{se} - 1} = 1 - m_j(x)\,, \tag{5.1}$$

where $m_j(x) = P(D^*_j = 1|X_{i,j} = x, R^D_{i,j} = 1)$. The difficulty in expressing this in terms of $p(x)$ comes from the randomness of the missing specimens within groups and the missing indicators, thus requiring combinatorial arguments. First, note that

$$\begin{aligned} m_j(x) =& 1 - P(D^*_j = -1|X_{i,j} = x, R^D_{i,j} = 1) - P(D^*_j = 0|X_{i,j} = x, R^D_{i,j} = 1) \\ =& 1 - P(D^*_j = 0|X_{i,j} = x, R^D_{i,j} = 1) \\ =& 1 - \sum_{w=1}^{n_j} P\Big(\max_{k\in I_j} D_{k,j} = 0, |I_j| = w\Big|X_{i,j} = x, R^D_{i,j} = 1\Big)\,, \end{aligned}$$

because, using (2.5), $D^*_j = -1 \Rightarrow R^D_{i,j} = 0$. Letting C^k_n denote the combination of k items among n, and noting that $P(D_{k,j} = 0, R^D_{k,j} = 1) = q_{RD} - q_R$, we deduce that

$$\begin{aligned} m_j(x) =& 1 - P\big(D_{i,j} = 0\big|X_{i,j} = x, R^D_{i,j} = 1\big) \sum_{w=1}^{n_j} C^{w-1}_{n_j-1}(q_{RD} - q_R)^{w-1} q_R^{n_j - w} \\ =& 1 - q_{RD}^{n_j-1}\{1 - E(D_{i,j}|X_{i,j} = x, R^D_{i,j} = 1)\} = 1 - q_{RD}^{n_j-1}\{1 - p(x)\}\,, \end{aligned} \tag{5.2}$$

where we use the binomial theorem and (2.3).

We can remove the dependence on j by multiplying the equations by $q_{RD}^{1-n_j}$, yielding

$$p(x) = 1 - m(x)\,, \tag{5.3}$$

where $m(x) = E\{q_{RD}^{1-n_j}(Z^*_j + \mathrm{se} - 1)/(\mathrm{sp} + \mathrm{se} - 1)|X_{i,j} = x, R^D_{i,j} = 1\}$. Because m is a regression curve that depends only on the observed data, we can estimate it using an ℓth-order local polynomial, as in Section 3.2, but this time constructed from the subset of the pairs $\big(X_{i,j}, \hat{q}_{RD}^{1-n_j}(Z^*_j + \mathrm{se} - 1)/(\mathrm{sp} + \mathrm{se} - 1)\big)$ corresponding to individuals for which $R^D_{i,j} = 1$, and with $\hat{q}_{RD}$ an MLE of q_{RD}, defined below. This suggests estimating $p(x)$ using

$$\hat{p}_2(x) = 1 - e_1^T \hat{\mathbf{S}}^{-1} \hat{\mathbf{T}}\,, \tag{5.4}$$

where $\hat{m}(x) = e_1^T \hat{\mathbf{S}}^{-1} \hat{\mathbf{T}}$, $\hat{\mathbf{S}} = (\hat{S}_{k,k'})_{0 \le k,k' \le \ell}$, and $\hat{\mathbf{T}} = (\hat{T}_0, \dots, \hat{T}_\ell)^T$, with

$$\hat{S}_{k,k'} = \frac{1}{Nh^{k+k'}} \sum_{j=1}^{J} \psi_j \sum_{i=1}^{n_j} R_{i,j}^D K_h(X_{i,j} - x)(X_{i,j} - x)^{k+k'}, \tag{5.5}$$

$$\hat{T}_k = \frac{1}{Nh^k} \sum_{j=1}^{J} \psi_j \hat{q}_{RD}^{1-n_j} \frac{Z_j^* + \mathrm{se} - 1}{\mathrm{sp} + \mathrm{se} - 1} \sum_{i=1}^{n_j} R_{i,j}^D K_h(X_{i,j} - x)(X_{i,j} - x)^k, \tag{5.6}$$

where ψ_j is a weight that depends on n_j (in Section 7.1, we show how to choose these weights in practice). Note that the individuals with $R_{i,j}^D = 0$ do not contribute to the estimator, that is, we do not use their $X_{i,j}$, because they do not add any information about $p(x) = E(D|X = x)$.

It remains to show how to estimate q_{RD}. In Appendix A.1, we show that $P(Z_j^* = 2) = q_R^{n_j}$, $P(Z_j^* = 1) = 1 - \mathrm{se} + (\mathrm{sp} + \mathrm{se} - 1) q_{RD}^{n_j} - \mathrm{sp}\, q_R^{n_j}$, and $P(Z_j^* = 0) = 1 - P(Z_j^* = 1) - P(Z_j^* = 2)$. For $r = 0, 1, 2$, define $\hat{P}(Z_j^* = r)$ obtained by replacing q_R with $\hat{q}_R = 1 - \sum_{j=1}^{J} \sum_{i=1}^{n_j} R_{i,j}^D / N$ in $P(Z_j^* = r)$. We estimate q_{RD} using the MLE $\hat{q}_{RD}$ obtained by maximizing $\mathcal{L}(q_{RD}, \hat{q}_R; Z_1^*, \dots, Z_J^*) = \prod_{j=1}^{J} \hat{P}(Z_j^* = z_j^*)$ with respect to $q_{RD} \in [\hat{q}_R, 1]$, with z_j^* the realization of Z_j^* in the sample.

5.2. Unknown individual missing status

In some cases, we may not know which individual specimens are missing. For example, the information may have been masked or may be lost or missing. Here, we show that it is possible to construct a consistent estimator in this case too. We observe $(X_{i,j}, Y_j^*, |I_j|)$, for $i = 1, \dots, n_j$, $j = 1, \dots, J$, with Y_j^* and the number $|I_j|$ of observed specimens in group j as in Section 2. Because we do not observe $R_{i,j}^D$, we cannot estimate p in (2.1) as we did in Section 5.1.

Again, the main difficulty is in expressing p in terms of the observed data. We already know that $E(Z_j^*|X_{i,j} = x)$ is not useful for estimating $p(x)$. Thus, instead of focusing directly on p, we start by studying functions that depend on the observed data, and then relate them to p. Because $|I_j| = \sum_{k=1}^{n_j} R_{k,j}^D$, we can write $E(|I_j||X_{i,j} = x) = (n_j - 1)(1 - q_R) + d(x)$, where $d(x) = E(R_{i,j}^D|X_{i,j} = x)$. Recalling that $Y_j^* = -1 \iff D_j^* = -1$, using the combinatorial derivations from Appendix A.3, we also have $P(Y_j^* = -1|X_{i,j} = x) = P(D_j^* = -1|X_{i,j} = x) = q_R^{n_j - 1}\{1 - d(x)\}$. Furthermore, recalling (2.7) and the definition of sp and se,

$$\begin{aligned} P(Y_j^* = 0|X_{i,j} = x) &= \sum_{k=0,1} P(Y_j^* = 0, D_j^* = k|X_{i,j} = x) \\ &= \mathrm{sp}\, P(D_j^* = 0|X_{i,j} = x) + (1 - \mathrm{se}) P(D_j^* = 1|X_{i,j} = x) \\ &= (\mathrm{sp} + \mathrm{se} - 1) q_{RD}^{n_j - 1}\{1 - b(x)\} - \mathrm{sp}\, q_R^{n_j - 1}\{1 - d(x)\} + 1 - \mathrm{se}, \end{aligned}$$

where $b(x) = E(R_{i,j}^D D_{i,j}|X_{i,j} = x)$, with q_{RD} as in (5.2). Using (2.3) and that

R^D and D are Bernoulli variables, we have $p(x) = E(D_{i,j}|X_{i,j} = x, R^D_{i,j} = 1) = b(x)/d(x)$. Together with the above calculations, this suggests that we can estimate p from our data.

Specifically, it follows from the results above that

$$d(x) = E\{|I_j| - (n_j - 1)(1 - q_R)|X_{i,j} = x\} \tag{5.7}$$

$$b(x) = E\left\{1 - q_{RD}^{1-n_j} \frac{W_j + 1 - \text{se}}{\text{sp} + \text{se} - 1}\middle| X_{i,j} = x\right\}, \tag{5.8}$$

where $W_j = \mathbb{1}\{Y_j^* = 0\} + \text{sp}\, \mathbb{1}\{Y_j^* = -1\}$. We can estimate q_R using $\hat{q}_R$ from Section 5.1, because we can write $\hat{q}_R = 1 - N^{-1} \sum_{j=1}^J |I_j|$, which depends only on the observed $|I_j|$. Therefore, we can estimate q_{RD} using the MLE $\hat{q}_{RD}$ from Section 5.1. Then, the regression curves b and d can be estimated from our data using ℓth-order local polynomial estimators $\hat{b}$ and $\hat{d}$, respectively, similar to those in Section 5.1. We take $\hat{b}(x) = e_1^T(\hat{\mathbf{S}}^p)^{-1}\hat{\mathbf{T}}^b$ and $\hat{d}(x) = e_1^T(\hat{\mathbf{S}}^p)^{-1}\hat{\mathbf{T}}^d$, where $\hat{\mathbf{S}}^p = (S^p_{k,k'})_{0\le k,k'\le \ell}$, $\hat{\mathbf{T}}^b = (\hat{T}^b_0, \ldots, \hat{T}^b_\ell)^T$, and $\hat{\mathbf{T}}^d = (\hat{T}^d_0, \ldots, \hat{T}^d_\ell)^T$, with, for $s = b$ and d and letting $U_{b,j} = 1 - \hat{q}_{RD}^{1-n_j}(W_j - 1 + \text{se})/(\text{sp} + \text{se} - 1)$ and $U_{d,j} = |I_j| - (n_j - 1)(1 - \hat{q}_R)$,

$$S^p_{k,k'} = \frac{1}{Nh^{k+k'}} \sum_{j=1}^J \psi_j \sum_{i=1}^{n_j} K_h(X_{i,j} - x)(X_{i,j} - x)^{k+k'},$$

$$\hat{T}^s_k = \frac{1}{Nh^k} \sum_{j=1}^J U_{s,j}\, \psi_j \sum_{i=1}^{n_j} K_h(X_{i,j} - x)(X_{i,j} - x)^k. \tag{5.9}$$

Note that, unlike the estimator $\hat{p}_2$ in Section 5.1, we use all $X_{i,j}$, even those for individuals whose specimens are missing, because we do not know whether $R^D_{i,j} = 0$ or 1. Here, we use the same h and ψ_j for $\hat{b}(x)$ and $\hat{d}(x)$ (see Section 7.1 for how to choose them in practice). Finally, we estimate $p(x)$ using the following ratio of two correlated local polynomial estimators:

$$\hat{p}_3(x) = \frac{\hat{b}(x)}{\hat{d}(x)} = \frac{e_1^T(\hat{\mathbf{S}}^p)^{-1}\hat{\mathbf{T}}^b}{\{e_1^T(\hat{\mathbf{S}}^p)^{-1}\hat{\mathbf{T}}^d\}}.$$

6. Asymptotic Properties

In this section, we investigate the asymptotic properties of our estimators. We treat $q_R = P(R^D = 0)$ and $q_{RD} = P(R^D D = 0)$ as parameters with unknown true values q_{R0} and q_{RD0}, respectively, and denote the corresponding value of $q_{D|R} = (q_{RD} - q_R)/(1 - q_R)$ by $q_{D|R0}$.

We need the following conditions to establish the theoretical properties of the estimators $\hat{p}_1(x)$ and $\hat{p}_2(x)$ from Sections 4 and 5.1, where $x \in \mathbb{R}$.

Condition 1.

(A1) $f_{X|R^D}(u|1)$ *is twice differentiable for all* u, $\|f^{(k)}_{X|R^D}(\cdot|1)\|_\infty < \infty$, *for* $k = 0, 1, 2$, *and* $f_{X|R^D}(x|1) > 0$.

(A2) K *is an even density function such that* $\int |u|^{2\ell+3}K(u)\,du < \infty$, *and for some* $\delta > 0$, $\int |u^{2\ell}K(u)|^{2+\delta}\,du < \infty$.

(A3) p *is* $\ell + 3$ *times differentiable and* $\|p^{(k)}\|_\infty < \infty$, *for* $k = 0, \ldots, \ell + 3$.

(A4) $h \to 0$ *and* $Nh \to \infty$ *as* $N \to \infty$.

(A5) $0 < \inf_j \psi_j \le \sup_j \psi_j < \infty$.

(A6) $\sup_j n_j < \infty$, $q_{R0} < q_{RD0} < 1$.

Conditions (A1) to (A4) are standard in nonparametric regression, and (A5) and (A6) are standard in group testing. (A1) and (A3) assume only that the functions are smooth, and (A2), (A4), and (A5) are satisfied easily because we choose K, h, and ψ_j (see Section 7.1). In (A6), the boundedness of n_j is always satisfied in practice, and $q_{R0} < q_{RD0} < 1$ is a mild condition used to prevent pathological cases in which all nonmissing data have the same disease status.

Given N', $\hat{p}_1$ from Section 4 is the same as in the nonmissing case studied in Delaigle and Meister (2011) and Delaigle, Hall and Wishart (2014), except that we replace (X, D) with $(X, D)|R^D = 1$. Then, the asymptotic normality of $\hat{p}_1$ follows from the results in those papers, combined with the fact that $N'/N \xrightarrow{P} 1 - q_{R0}$ as $N \to \infty$, because $N' \sim \text{Bi}(N, 1 - q_{R0})$. The central limit theorem for a random sum can be found in, among others, Bethmann (1989). Specifically, let $N'_\psi = \sum_{j=1}^{J'} n_j\psi_j$, $\mu_{K,j} = \int u^j K(u)\,du$, $\nu_j = \int u^j K^2(u)\,du$, $\boldsymbol{\mu} = (\mu_{K,\ell+1}, \ldots, \mu_{K,2\ell+1})^T$, $\tilde{\boldsymbol{\mu}} = (\mu_{K,\ell+2}, \ldots, \mu_{K,2\ell+2})^T$, $\mathbf{m}(x) = \{m(x), \ldots, h^\ell(\ell!)^{-1}m^{(\ell)}(x)\}$, where $m = 1 - p$, and let $\mathbf{S}$, $\tilde{\mathbf{S}}$, and $\mathbf{S}^*$ be $(\ell+1)\times(\ell+1)$ matrices with the $(k+1, k'+1)$th element defined by $\mathbf{S}_{k,k'} = \mu_{K,k+k'}$, $\tilde{\mathbf{S}}_{k,k'} = \mu_{K,k+k'+1}$, and $\mathbf{S}^*_{k,k'} = \nu_{k+k'}$, respectively, for $k, k' = 0, \ldots, \ell$. Under Conditions (A1)–(A6), it follows from Delaigle and Meister (2011) and Delaigle, Hall and Wishart (2014) that

$$\hat{p}_1(x) = p(x) + B(x) + \sqrt{V_1(x)}\,\mathcal{N}_N + o_p\{B(x)\} + o_p\{\sqrt{V_1(x)}\}\,,$$

where $\mathcal{N}_N \xrightarrow{\mathcal{D}} N(0,1)$ as $N \to \infty$, $V_1(x) = e_1^T\mathbf{S}^{-1}\mathbf{S}^*\mathbf{S}^{-1}e_1 \sum_{j=1}^{J'} n_j\psi_j^2\mathbb{V}_{1,j}(x)/\{(N'_\psi)^2 h f_{X|R^D}(x|1)\}$,

$$\text{with} \quad \mathbb{V}_{1,j}(x) = \frac{(2\,\text{se} - 1)m(x)}{q_{D|R0}^{n_j-1}(\text{sp} + \text{se} - 1)} + \frac{\text{se} - \text{se}^2}{q_{D|R0}^{2n_j-2}(\text{sp} + \text{se} - 1)^2} - m^2(x)\,,$$

and for ℓ odd, $B(x) = -e_1^T \mathbf{S}^{-1} \boldsymbol{\mu}\, m^{(\ell+1)}(x) h^{\ell+1}/(\ell+1)!$, while for ℓ even,

$$B(x) = \; e_1^T \mathbf{S}^{-1} \left\{ (\tilde{\mathbf{S}} \mathbf{S}^{-1} \boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}) \frac{m^{(\ell+1)}(x) f'_{X|R^D}(x|1)}{(\ell+1)! f_{X|R^D}(x|1)} - \tilde{\boldsymbol{\mu}} \frac{m^{(\ell+2)}(x)}{(\ell+2)!} \right\} h^{\ell+2} .$$

Comparing these results with those without missing data from Delaigle and Meister (2011) and Delaigle, Hall and Wishart (2014), the only difference is that, here, quantities that depend on X and D are conditional on $R^D = 1$, and our sample size $N' \sim \text{Bi}(N, 1 - q_{R0})$. The "bias" term B is of the same order as in the case without missing data: $B(x) \asymp h^{\ell+1}$ for ℓ odd, $B(x) \asymp h^{\ell+2}$ for ℓ even. The "variance" term is also of the same order as in the case without missing data, because $V_1(x) \asymp (N'h)^{-1} = (N(1-q_{R0})h)^{-1}\{1 + o_P(1)\}$. The convergence rate of $\hat{p}_1$ is optimized by taking $B(x) \asymp \sqrt{V_1(x)}$, that is $h \asymp N^{-1/(2\ell+3)}$ for ℓ odd and $h \asymp N^{-1/(2\ell+5)}$ for ℓ even, which gives a rate of order $N^{-(\ell+1)/(2\ell+3)}$ for ℓ odd and $N^{-(\ell+2)/(2\ell+5)}$ for ℓ even, as in the case without missing data.

The following theorem establishes the asymptotic normality of $\hat{p}_2(x)$ from Section 5.1. See Appendix B.1 for a proof.

Theorem 1. *Let $N_\psi = \sum_{j=1}^J n_j \psi_j$. Under Conditions* (A1)–(A6)*, we have*

$$\hat{p}_2(x) = p(x) + B(x) + \sqrt{V_2(x)}\, \mathcal{N}_N + o_p\{B(x)\} + o_p\{\sqrt{V_2(x)}\} ,$$

where $\mathcal{N}_N \xrightarrow{\mathcal{D}} N(0,1)$ as $N \to \infty$, $B(x)$ is as above, and $V_2(x) = e_1^T \mathbf{S}^{-1} \mathbf{S}^* \mathbf{S}^{-1} e_1$ $\sum_{j=1}^J n_j \psi_j^2 \mathbb{V}_{2,j}(x) / \{N_\psi^2 h (1-q_{R0}) f_{X|R^D}(x|1)\}$, with

$$\mathbb{V}_{2,j}(x) = \frac{(2\,\text{se} - 1) m(x)}{q_{RD0}^{n_j - 1}(\text{sp} + \text{se} - 1)} + \frac{\text{se} - \text{se}^2}{q_{RD0}^{2n_j - 2}(\text{sp} + \text{se} - 1)^2} - m^2(x) .$$

Here too, the "bias" and "variance" terms, B and V_2, respectively, are of the same order as in the case without missing data (B is the same as for $\hat{p}_1$, and $V_2(x) \asymp 1/(Nh)$). The optimal convergence rate of $\hat{p}_2$ is the same as that of $\hat{p}_1$, with h of the same order as for $\hat{p}_1$.

Recall that the advantage of $\hat{p}_2$ is that the groups can be created regardless of the missing status of the specimens, but it is interesting to compare its performance with that of $\hat{p}_1$. Both have the same asymptotic bias term B, but, in general, it is difficult to compare their variance terms V_1 and V_2, which differ in the number of groups, n_j, $q_{D|R0}$, and q_{RD0}. We can compare them when all groups are of equal size, $n_j = n$, because $\psi_j = 1$, $N_\psi = N$, and $N'_\psi = N(1-q_{R0})\{1 + o_P(1)\}$. In that case, $\hat{p}_2$ outperforms $\hat{p}_1$, because $q_{D|R0} = 1 - (1 - q_{RD0})/(1 - q_{R0}) \le 1 - (1 - q_{RD0}) = q_{RD0}$ and $V_2(x)/V_1(x) = \mathbb{V}_{2,1}(x)/\mathbb{V}_{1,1}(x) + o_P(1)$, with $\mathbb{V}_{2,1}(x) \le \mathbb{V}_{1,1}(x)$. However if n_j is smaller for $\hat{p}_1$ than it is for $\hat{p}_2$ and both estimators use the same number of groups, $J' = J$, then $\hat{p}_1$ usually outperforms $\hat{p}_2$.

We need the following conditions to derive the theoretical properties of $\hat{p}_3(x)$ from Section 5.2.

Condition 2.

(B1) *f_X is twice differentiable, $||f_X^{(k)}||_\infty < \infty$, for $k = 0, 1, 2$ and $f_X(x) > 0$.*

(B2) *K is an even density function, $\int |u|^{2\ell+3} K(u)\, du < \infty$ and $\int |u^{2\ell} K(u)|^3\, du < \infty$.*

(B3) *b and d defined in* (5.8) *and* (5.7)*, respectively, are $\ell+3$ times differentiable, $\|b^{(k)}\|_\infty < \infty$, and $\|d^{(k)}\|_\infty < \infty$, for $k = 0, \ldots, \ell+3$, and $d(x) > 0$.*

(B4) *to* (B6) *are defined in the same way as* (A4) *to* (A6)*, respectively.*

(B7) *$\operatorname{cov}\{(U_{b0,j}, |I_j|)|X_{i,j} = x\} = \big(\Sigma_{j,k\ell}(x)\big)_{k,\ell=1,2}$ is invertible, for $j = 1, \ldots, J$, where $U_{b0,j}$ is the version of $U_{b,j}$ with $\hat{q}_{RD}$ replaced with q_{RD0}, and the expressions for $\Sigma_{j,k\ell}$ are given in Appendices B.3 and B.4.*

Conditions (B1)–(B6) are similar to (A1)–(A6), respectively, in Condition A. Condition (B7) is mild: $U_{b0,j}$ is a function of Y_j^* and $|I_j|$ is a function of $R_{k,j}^D$, so it would be very unusual for their conditional covariance matrix to be noninvertible. This condition plays the role of the assumption of an invertible covariance matrix used in the standard multivariate central limit theorem (Rao (1973); Serfling (2009)), and is used only to establish the asymptotic normality of $\hat{p}_3$ but is not needed for $\hat{p}_3$ to be consistent.

The next theorem establishes the asymptotic properties of $\hat{p}_3$. Here, for ℓ even, the bias term of the asymptotic expansion is much more involved than in the case of $\hat{p}_2$. Therefore, and because in practice it is standard to use odd-order local polynomial estimators (they have better properties, e.g., near boundaries; see Remark 3), we establish our theorem only for estimators of odd order. See Appendix B.2 for a proof.

Theorem 2. *Under Conditions* (B1)–(B7)*, if ℓ is odd, we have $\hat{p}_3(x) = p(x) + B_3(x) + \sqrt{V_3(x)}\mathcal{N}_N + o_p\{B_3(x)\} + o_p\{\sqrt{V_3(x)}\}$, where $\mathcal{N}_N \xrightarrow{\mathcal{D}} N(0,1)$ as $N \to \infty$,*

$$B_3(x) = e_1^T \mathbf{S}^{-1} \boldsymbol{\mu} h^{\ell+1} \frac{d(x) b^{(\ell+1)}(x) - b(x) d^{(\ell+1)}(x)}{(\ell+1)! d^2(x)},$$

and $V_3(x) = \{N_\psi^2 h(1 - q_{R0}) f_{X|R^D}(x|1)\}^{-1} e_1^T \mathbf{S}^{-1} \mathbf{S}^* \mathbf{S}^{-1} e_1 \sum_{j=1}^J n_j \psi_j^2 \mathbb{V}_{3,j}(x)/d(x)$, with $\mathbb{V}_{3,j}(x) = \Sigma_{j,11}(x) - 2p(x)\Sigma_{j,12}(x) + p^2(x)\Sigma_{j,22}(x)$ and $\Sigma_{j,k\ell}$ is defined as in Condition (B7)*.*

As for Theorem 1, the rate of the "bias" term B_3 and the "variance" term V_3 in Theorem 2 are the same as in the case without grouping for odd ℓ, that is, $B_3(x) \asymp h^{\ell+1}$ and $V_3(x) \asymp (Nh)^{-1}$. Again, the optimal convergence rate $N^{-(\ell+1)/(2\ell+3)}$ of $\hat{p}_3$ is obtained by taking $B_3(x) \asymp \sqrt{V_3(x)}$, that is, $h \asymp N^{-1/(2\ell+3)}$.

However, $\hat{p}_3(x)$ is a ratio of two correlated local polynomial estimators, which makes the asymptotic expressions more involved and difficult to compare with those for $\hat{p}_1$ and $\hat{p}_2$. We compare these estimators numerically in Section 7.

Remark 1. (Integrated squared error). For each estimator $\hat{p}_k$, for $k = 1, 2, 3$, we can also compute an asymptotic weighted mean integrated squared error, $\text{AMISE}_w = \int \{B^2(x) + V_k(x)\} f_{X|R^D}(x|1) w(x)\, dx$, where w is an integrable weight function. AMISE_w is commonly used in nonparametric regression problems to compute a plug-in (PI) bandwidth (see Section 7.1), and is of the same asymptotic order as its pointwise version, in our case, the quantity $B^2(x) + V_k(x)$. For example, for ℓ odd and for our three estimators, it is optimized at the rate $N^{-(\ell+1)/(2\ell+3)}$, obtained by taking $h \asymp N^{-1/(2\ell+3)}$.

Remark 2. (Group sizes). The choice of n_j depends on a number of factors, and involves a trade-off between optimizing the main goal of the study and remaining within its time, budget, and other constraints. If the main goal is to estimate p, then an optimal strategy could be to minimize AMISE_w from Remark 1, computed using its optimal bandwidth, under the various constraints (for $\ell = 1$, the optimal bandwidth is derived in Section 7.1). For example, if the only constraint is that the number tests that can be performed is equal to a given number J, then the optimal AMISE_w-based strategy is to take $n_j = n$ so that $N_\psi = N$ and $\psi_j = 1$, and there is a corresponding value n that minimizes AMISE_w. As in the parametric case without missing data studied in Section 3 of Vansteelandt, Goetghebeur and Verstraeten (2000), finding this n requires a preliminary estimator of p, for example, computed from a small sample. However, if the main goal of the study is to estimate the nonconditional prevalence, and p is a side result, then we can replace AMISE_w with a criterion for that nonconditional estimator.

Remark 3. (Boundary case). If $f_{X|R^D}(\cdot|1)$ is compactly supported and not continuous at the endpoints of its support, then unlike kernel density estimators, we find that local polynomial estimators, and particularly our three estimators, remain consistent. However, while local polynomial estimators of odd order ℓ converge at the same rate as in the absence of boundaries, the rate degrades if ℓ is even. In the latter case, the bias component is of order $h^{\ell+1}$ instead of $h^{\ell+2}$, and the convergence rate of the estimator is of order $N^{-(\ell+1)/(2\ell+3)}$ instead of $N^{-(\ell+2)/(2\ell+5)}$. For example, a local constant estimator ($\ell = 0$) converges at the rate $N^{-1/3}$ in the boundary case, instead of the $N^{-2/5}$ rate in the no-boundary case, whereas a local linear estimator ($\ell = 1$) converges at the rate $N^{-2/5}$ in both cases.

7. Simulation Study

7.1. Computing the estimators in practice

The estimators $\hat{p}_1$, $\hat{p}_2$, and $\hat{p}_3$ all include weight functions ψ_j and a tuning parameter h. In this section, we show how to choose these values for the local linear version ($\ell = 1$) of the estimators, which is the most popular version of the local polynomial estimators, owing to its nice properties at boundaries (see Remark 3).

As in Delaigle, Hall and Wishart (2014), because ψ_j does not affect the asymptotic bias of $\hat{p}_1$, $\hat{p}_2$, and $\hat{p}_3$, we choose it by minimizing $\int v(x) f_{X|R^D}(x|1)$ $w(x)\,dx$ with respect to ψ_j, with w a weight function (see Section 7.2 for its choice) and where, for $k = 1$ to 3, $v = V_k$ is defined as in Section 6. This gives $\psi_{k,j} = \big\{\int \mathbb{V}_{k,j}(x)w(x)\,dx\big\}^{-1}$ for $\hat{p}_k$, $k = 1, 2$, and $\psi_{3,j} = \big\{\int \mathbb{V}_{3,j}(x)w(x)/d(x)\,dx\big\}^{-1}$ for $\hat{p}_3$, with $\mathbb{V}_{k,j}(x)$ as in Section 6 for $k = 1$ to 3; see Appendix C.1. In practice, for $k = 1, 2$, we estimate $\psi_{k,j}$ as $\hat{\psi}_{k,j} = \big\{\int \hat{\mathbb{V}}_{k,j}(x)w(x)\,dx\big\}^{-1}$, with $\hat{\mathbb{V}}_{1,j}$ and $\hat{\mathbb{V}}_{2,j}$ obtained by replacing $q_{D|R}$ with $\hat{q}_{D|R}$ given in Section 4, q_{RD} with $\hat{q}_{RD}$ given in Section 5.1, and $\tilde{m}$ and m with the pilot estimators $\hat{\tilde{m}}_{\text{PILOT}}$ and $\hat{m}_{\text{PILOT}}$, respectively, defined by $\hat{\tilde{m}}$ and $\hat{m}$ in Sections 4 and 5.1, respectively, with $\ell = 0$, $\psi_j \equiv 1$, and the cross-validation (CV) h from Appendix C.2. Similarly, we estimate $\psi_{3,j}$ as $\hat{\psi}_{3,j} = \big\{\int \hat{\mathbb{V}}_{3,j}(x)w(x)/\hat{d}_{\text{PILOT}}(x)\,dx\big\}^{-1}$, with $\hat{\mathbb{V}}_{3,j}$ obtained by replacing, in $\mathbb{V}_{3,j}$, q_R and q_{RD} with $\hat{q}_R$ and $\hat{q}_{RD}$, respectively, from Section 5.1, and b and d with $\hat{b}_{\text{PILOT}}$ and $\hat{d}_{\text{PILOT}}$, defined by $\hat{b}$ and $\hat{d}$, respectively, above (5.9), with $\ell = 0$, $\psi_j \equiv 1$, and the CV bandwidth h from Appendix C.2.

To choose h for $\hat{p}_2$, we use a PI approach, as in Delaigle and Meister (2011). Let B and V_2 be defined as in Theorem 1, w be defined as for ψ_j, and $\Theta_{2,1} = \int\{p''(x)\}^2 f_{X|R^D}(x|1)w(x)\,dx$. We choose h by minimizing, with respect to h, an estimator of $\text{AMISE}_w = \int\{B^2(x) + V_2(x)\}f_{X|R^D}(x|1)w(x)\,dx = \mu_{K,2}^2\Theta_{2,1}h^4/4 + \nu_0\sum_{j=1}^J n_j\psi_j^2\int \mathbb{V}_{2,j}(x)w(x)\,dx/\{h(1-q_R)N_\psi^2\}$, obtained by estimating $\Theta_{2,1}$ by $\hat{\Theta}_{2,1}$ (Appendix C.3), ψ_j by $\hat{\psi}_{2,j}$, and q_R by $\hat{q}_R$ (Section 5.1), resulting in our PI bandwidth $\hat{h}_{\text{PI},2} = \nu_0^{1/5}\{(1-\hat{q}_R)\mu_{K,2}^2\hat{\Theta}_{2,1}\sum_{j=1}^J n_j\hat{\psi}_{2,j}\}^{-1/5}$. Similarly, replacing V_2 with V_1 for $\hat{p}_1$, and B and V_2 with B_3 and V_3, respectively, for $\hat{p}_3$, and following the same arguments, our PI bandwidth for $\hat{p}_1$ is equal to $\hat{h}_{\text{PI},1} = \nu_0^{1/5}\big(\mu_{K,2}^2\tilde{\Theta}_{2,1}\sum_{j=1}^{J'} n_j\hat{\psi}_{1,j}\big)^{-1/5}$, and for $\hat{p}_3$ is equal to $\hat{h}_{\text{PI},3} = \nu_0^{1/5}\{(1-\hat{q}_R)\mu_{K,2}^2\hat{\Theta}_{2,2}\sum_{j=1}^J n_j\hat{\psi}_{3,j}\}^{-1/5}$, where $\tilde{\Theta}_{2,1}$, $\Theta_{2,2}$, and $\hat{\Theta}_{2,2}$ are defined in Appendix C.3.

7.2. Simulation results

Here, we apply the local linear versions ($\ell = 1$) of our estimators of p from Sections 4 and 5 to simulated data, with h and ψ_j chosen as in Section 7.1. We use the same n_j for $\hat{p}_1$ and $\hat{p}_2$ (the groups for $\hat{p}_2$ are created without knowing the number of missing specimens, and so there is no sensible way to use a different

Table 1. Simulation results for five nonparametric estimators of p with MAR D for models (i) to (iv). We show the median (interquartile range) of ISE$\times 10^3$ computed from 200 samples.

Model			$\hat{p}_{\text{nai}}$		$\hat{p}_2$		$\hat{p}_1$		$\hat{p}_3$		$\hat{p}_{\text{ungr},N}$		$\hat{p}_{\text{ungr},J}$	
							$J = 250$							
(1)	(i)	(A)	8.29	(6.12)	5.68	(6.27)	6.66	(6.79)	7.46	(9.58)	2.96	(3.09)		
		(B)	7.35	(6.76)	5.80	(5.56)	6.19	(8.12)	7.08	(9.40)	3.00	(3.14)	14.17	(17.39)
		(C)	6.99	(7.02)	6.10	(6.32)	6.22	(6.27)	7.80	(7.87)	1.62	(1.65)		
	(ii)	(A)	21.21	(14.59)	9.57	(9.47)	11.99	(12.42)	11.67	(14.22)	2.86	(3.10)		
		(B)	18.67	(14.62)	9.05	(6.93)	10.52	(10.83)	10.89	(13.16)	3.28	(3.19)	13.34	(15.42)
		(C)	24.14	(21.52)	18.17	(17.14)	20.67	(25.07)	20.29	(26.02)	1.65	(1.45)		
	(iii)	(A)	14.90	(12.58)	5.30	(7.76)	5.87	(7.95)	5.95	(9.52)	1.65	(2.48)		
		(B)	13.77	(12.67)	4.41	(5.80)	4.86	(7.67)	6.44	(8.37)	1.88	(2.67)	8.49	(14.82)
		(C)	14.91	(15.64)	5.85	(7.74)	7.14	(9.51)	8.04	(13.66)	1.06	(1.16)		
(2)	(i)	(A)	11.03	(8.58)	6.20	(6.29)	7.92	(9.41)	9.56	(10.76)	3.24	(3.40)		
		(B)	11.18	(9.37)	6.86	(6.80)	8.33	(8.37)	10.45	(12.87)	3.62	(4.53)	18.46	(21.22)
		(C)	10.74	(8.41)	6.76	(6.94)	10.62	(12.08)	9.64	(14.83)	1.75	(1.95)		
	(ii)	(A)	37.50	(21.97)	11.07	(13.10)	14.36	(16.39)	14.55	(19.74)	3.83	(4.66)		
		(B)	40.83	(19.56)	11.68	(11.44)	14.10	(18.15)	14.91	(18.29)	4.69	(4.64)	15.41	(20.12)
		(C)	42.64	(21.78)	12.88	(12.52)	25.64	(28.92)	18.04	(27.22)	2.14	(2.65)		
	(iii)	(A)	33.19	(16.81)	5.87	(8.66)	5.96	(7.89)	7.14	(11.69)	2.65	(4.11)		
		(B)	34.81	(17.60)	5.78	(7.27)	6.46	(9.71)	8.55	(10.62)	3.19	(5.31)	13.08	(22.93)
		(C)	37.21	(17.37)	4.31	(6.57)	8.16	(10.66)	8.17	(16.50)	1.55	(1.80)		
							$J = 2000$							
(1)	(i)	(A)	5.48	(2.67)	0.97	(0.89)	1.22	(1.05)	1.54	(1.39)	0.54	(0.57)		
		(B)	5.59	(2.54)	1.08	(0.87)	1.12	(0.99)	1.36	(1.51)	0.59	(0.65)	2.39	(2.11)
		(C)	5.27	(2.78)	1.17	(1.02)	1.44	(1.47)	1.70	(1.86)	0.33	(0.30)		
	(ii)	(A)	15.23	(5.71)	2.21	(2.02)	2.48	(2.25)	2.26	(3.04)	0.62	(0.48)		
		(B)	14.92	(6.36)	1.95	(1.75)	2.24	(1.63)	2.31	(2.80)	0.65	(0.61)	2.42	(2.43)
		(C)	16.31	(8.88)	3.22	(2.77)	4.84	(4.30)	3.56	(4.57)	0.36	(0.32)		
	(iii)	(A)	13.98	(4.94)	0.94	(1.06)	1.03	(1.29)	1.48	(1.72)	0.37	(0.42)		
		(B)	14.29	(5.46)	0.99	(1.15)	0.89	(1.12)	1.49	(2.23)	0.39	(0.49)	1.35	(1.54)
		(C)	13.59	(5.98)	1.10	(1.07)	1.30	(1.81)	1.90	(2.32)	0.21	(0.25)		
(2)	(i)	(A)	8.93	(3.31)	1.12	(1.09)	1.42	(1.15)	1.82	(2.05)	0.61	(0.63)		
		(B)	9.07	(3.41)	1.16	(1.08)	1.48	(1.41)	1.58	(1.83)	0.70	(0.71)	2.52	(2.89)
		(C)	9.73	(3.54)	1.23	(1.13)	1.80	(1.79)	1.81	(1.93)	0.32	(0.30)		
	(ii)	(A)	35.12	(7.62)	2.15	(1.96)	3.52	(3.51)	3.23	(3.01)	0.80	(0.63)		
		(B)	34.88	(7.65)	2.07	(1.89)	3.51	(2.70)	2.98	(2.95)	0.91	(0.89)	3.11	(3.18)
		(C)	35.73	(9.49)	2.72	(2.41)	6.29	(6.56)	3.81	(3.44)	0.49	(0.46)		
	(iii)	(A)	31.63	(5.70)	0.84	(0.99)	1.15	(1.34)	1.31	(1.52)	0.48	(0.52)		
		(B)	32.11	(7.38)	0.94	(1.11)	1.24	(1.35)	1.18	(1.51)	0.56	(0.67)	2.07	(2.51)
		(C)	32.70	(5.84)	0.84	(0.90)	1.45	(1.42)	1.18	(1.87)	0.28	(0.31)		

n_j). Therefore, the number of groups J' for $\hat{p}_1$ is smaller than that, J, for $\hat{p}_2$, and we expect $\hat{p}_2$ to outperform $\hat{p}_1$ (see the discussion under Theorem 1). Because $\hat{p}_2$

exploits $R^D_{i,j}$, whereas $\hat{p}_3$ uses the less informative $|I_j| = \sum_{i=1}^{n_j} R^D_{i,j}$, we also expect $\hat{p}_2$ to outperform $\hat{p}_3$.

Because group testing is based on less information, it is clear that estimators constructed from J groups of N aggregated specimens are less accurate than an estimator constructed from N nongrouped specimens. To illustrate how much information is lost by grouping, we computed the estimator $\hat{p}_{\mathrm{ungr},N}$, constructed from N nongrouped specimens, which is equal to $\hat{p}_2$ with $n_j = 1$ and $J = N$. We also computed the estimator $\hat{p}_{\mathrm{ungr},J}$ constructed from J nongrouped specimens, which is equal to $\hat{p}_2$ with $n_j = 1$ and $N = J$. Here, the estimator $\hat{p}_2$ computed from J groups of N aggregated specimens can outperform $\hat{p}_{\mathrm{ungr},J}$, because only a small fraction of individuals are positive, and so we need the sample to contain enough positives to obtain a good estimator, and $\hat{p}_2$ uses N individuals rather than J.

To illustrate why we need to take MAR into account, we compared $\hat{p}_2$ with the naive estimator $\hat{p}_{\mathrm{nai}}$ of p introduced in Section 5.1. Note that we cannot compare $\hat{p}_{\mathrm{nai}}$ with $\hat{p}_3$, which we use only when $R^D_{i,j}$ is not available (i.e., $\hat{p}_{\mathrm{nai}}$ is not computable). There does not seem to be an obvious naive version of $\hat{p}_3$ using the same data as $\hat{p}_3$. For all estimators, we took the kernel K as the standard normal density, and w from Section 7.1 equal to $w(x) = \mathbb{1}_{[q_{0.1},q_{0.9}]}(x)$, with q_α the empirical α-quantile of X. For the CV criterion used in Section 7.1, we took $[a, b] = [q_{0.1}, q_{0.9}]$.

To generate $(X_{i,j}, \tilde{Y}^*_j, R^D_{i,j})$ and $(X_{i,j}, Y^*_j, R^D_{i,j})$, we first generated $(X_{i,j}, D_{i,j}, R^D_{i,j})$ and then obtained $\tilde{Y}^*_j$ following (2.4) and (2.6), and Y^*_j following (2.5) and (2.7), where we took $\mathrm{sp} = 0.99$ and $\mathrm{se} = 0.85$, that is, within ranges reported from Covid-19 testing (e.g., Arevalo-Rodriguez et al. (2020); Surkova, Nikolayevskyy and Drobniewski (2020)). We generated $(X_{i,j}, D_{i,j})$ from three models: (i) $p(x) = \min(x^2/8, 1)$; (ii) $p(x) = \mathbb{1}_{(-\infty,-3)}(x) + \big[1/\{1+\exp(2x+4)\} + (x-0.4)^2 \sin(\pi x)/20 + 0.1\big]\mathbb{1}_{[-3,3.08]}(x)$; and (iii) $p(x) = 1/\{1+\exp(2x+3)\}$, where $X \sim N(0, 0.75^2)$ and $D|X \sim \mathrm{Be}\{p(X)\}$, a Bernoulli distribution with parameter $p(X)$. Model (i) was used by Delaigle and Meister (2011) and Delaigle, Hall and Wishart (2014), model (iii) is a logistic curve, and model (ii) has a few more features. In (i) and (ii), p is nondifferentiable at two points far in the tails of f_X, which does not affect the overall performance of the estimators. In each case, we generated $R^D_{i,j}$ in two ways similar to the method of Zhou, Wan and Wang (2008): (1) $R^D|X \sim \mathrm{Be}[0.7 + 0.3\sin\{(X-1)^2\}]$; and (2) $R^D|X \sim \mathrm{Be}\big(\exp\{\sin(X) + 0.5\}/[1 + \exp\{\sin(X) + 0.5\}]\big)$. The average percentage of missing data is 20% (resp. 39%) in case (1) (resp. (2)); thus, case (2) is the most challenging.

We generated data from all combinations of models (i) to (iii) and (1) and (2), for $J = 250, 500, 1000$, and 2000 groups of sizes n_j chosen in three ways: (A) $J/2$ groups of size $n_j = 4$ and $J/2$ groups of size $n_j = 8$; (B) J groups of size $n_j = 5$; and (C) J groups of size $n_j = 12$ (for $\hat{p}_1$ we took the same n_j but replaced J with the random J' in each sample). We ran simulations from each

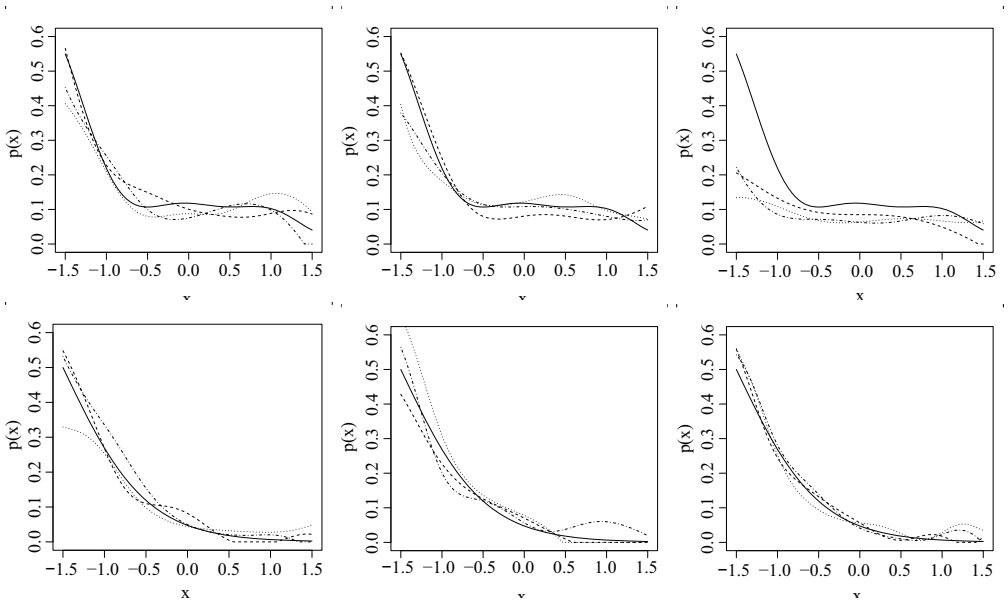

Figure 1. True curve (—), first (- - -), second (- · - · -), and third (· · ·) quartile estimated curves in the MAR D case. Top: $\hat{p}_2$ (left), $\hat{p}_{\text{ungr},J}$ (middle), and $\hat{p}_{\text{nai}}$ (right) for model (ii) in case (2) with $J = 1{,}000$ and grouping (A). Bottom: model (iii) in case (1) with grouping (A), when $J = 500$ for $\hat{p}_2$ (left) or $\hat{p}_3$ (middle), and when $J = 2{,}000$ for $\hat{p}_3$ (right).

combination 200 times and summarized the results using the integrated squared error, ISE $= \int_{-1.5}^{1.5} \{\check{p}(x) - p(x)\}^2 \, dx$, where $\check{p}$ denotes any estimator of p, truncated to [0,1], because we know that $p \in [0, 1]$; note that $[-1.5, 1.5]$ contains about 95% of X_i.

Table 1 shows, for each estimator, the median and interquartile range of the 200 ISE$\times 10^3$ for $J = 250$ and 2000; see Table D.1 in Appendix D for the other values of J. To see what this corresponds to for $\hat{p}_1$, recall that the number J' of groups used by $\hat{p}_1$ is random, because it is computed from the number $N' \sim \text{Bi}(N, 1 - q_R)$ of individuals with nonmissing specimens in each sample, where $N = \sum_{j=1}^{J} n_j$ and $N' = \sum_{j=1}^{J'} n_j$. Unsurprisingly, in general, for all estimators, the difficulty of the estimating task increases with the amount of missing data. As expected, $\hat{p}_1$, $\hat{p}_2$, $\hat{p}_3$, and $\hat{p}_{\text{ungr},J}$ improved as J increased, but the inconsistent $\hat{p}_{\text{nai}}$ was very biased and performed poorly. Consistent with our theory in Section 6, $\hat{p}_2$ performed slightly better (or even much better for grouping (C)) than $\hat{p}_1$ in all cases. Although $\hat{p}_3$ requires only $|I_j| = \sum_{i=1}^{n_j} R_{i,j}^D$ for each j, its performance was not much worse than that of $\hat{p}_2$, which needs the individual $R_{i,j}^D$. An exception is model (ii) with grouping (C), where the larger prevalence and group sizes were more difficult to deal with for $\hat{p}_3$. Furthermore, $\hat{p}_2$ outperformed $\hat{p}_{\text{ungr},J}$ in all cases ($\hat{p}_1$ and $\hat{p}_3$ did in most cases, but again, not for model (ii) with grouping (C)), which can be expected in these low prevalence settings where we need to observe

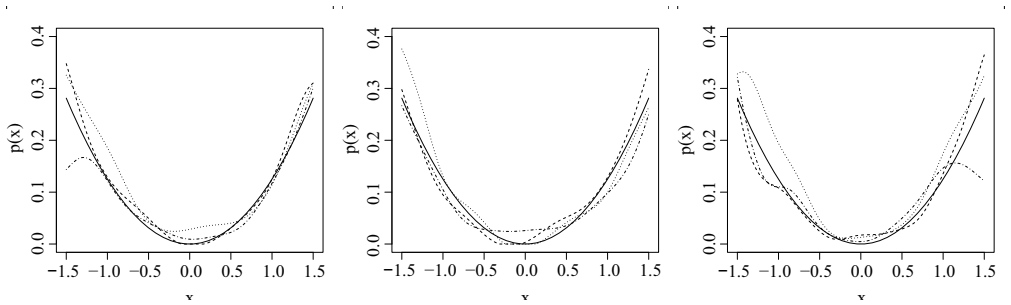

Figure 2. True curve (—), first (- - -), second (-·-·-), and third (· · ·) quartile estimated curves in the MAR D case for model (i) with grouping (B) and $J = 1000$ in case (2) for $\hat{p}_2$ (left), and in case (1) for $\hat{p}_2$ (middle) and $\hat{p}_{\mathrm{ungr},J}$ (right).

many individuals to find some positives. Finally, as expected, the estimator $\hat{p}_{\mathrm{ungr},N}$ that uses N nongrouped individuals significantly outperformed the other estimators, although the estimators constructed from grouped data performed well; see the figures below for an illustration. Note that for $\hat{p}_{\mathrm{ungr},N}$, the sample size N is larger with grouping (C) than it is with (A), which is itself larger than that of (B). This explains why it performed much better for grouping (C), and a bit better for grouping (A), than it did for grouping (B).

To illustrate some of these results visually, we show, for a few cases, the true curve and three estimated curves corresponding to the samples that gave the first, second, and third quartile values, respectively, out of the 200 ISEs. We refer to them as the first, second, and third quartile estimated curves, respectively. The top row of Fig. 1 compares $\hat{p}_2$, $\hat{p}_{\mathrm{ungr},J}$, and $\hat{p}_{\mathrm{nai}}$ for model (ii) in case (2) with grouping (A) and $J = 1{,}000$. It illustrates the large bias of the inconsistent $\hat{p}_{\mathrm{nai}}$, which performed poorly in most cases. It also illustrates how $\hat{p}_2$ can outperform $\hat{p}_{\mathrm{ungr},J}$. As illustrated in the second row of Fig. 1, $\hat{p}_2$ and $\hat{p}_3$ often performed similarly; here, we show them for model (iii) in case (1) with grouping (A) and $J = 500$. To illustrate that our estimators improve as J increases, we also show $\hat{p}_3$ for $J = 2{,}000$ (we obtained similar results for $\hat{p}_1$ and $\hat{p}_2$). Fig. 2 illustrates the finite-sample advantage of $\hat{p}_2$ over $\hat{p}_{\mathrm{ungr},J}$; we show them for model (i), grouping (B); and $J = 1{,}000$ in case (1). We also see that $\hat{p}_2$ performed worse in case (2) than it did in case (1), which illustrates the degradation of the estimators when more data are missing.

8. Real-Data Illustration

As is usual in real-data analyses from the group testing literature, our goal is to compare our estimators based on grouped data with estimators based on nongrouped data, to show that group testing can be applied in practice. As in that literature (e.g., Xie (2001); Chen, Tebbs and Bilder (2009); Zhang, Bilder and Tebbs (2013)), our data sets include individual test results, which we treated as

perfect, that is, $D_{i,j} \equiv Y_{i,j}$ (the documentation available for those data suggests that this is reasonable; see, for example, Maheu-Giroux et al. (2017) for HIV data). Then, as in the literature, we grouped the individuals into $J-1$ (resp. $J'-1$) groups of equal size $n_j = n$ (we considered two cases (D): $n = 8$ and (E): $n = 4$), and one group of size $N - n(J-1)$, where $J = \lfloor N/n \rfloor$ (resp. $N' - n(J'-1)$, where $J' = \lfloor N'/n \rfloor$), and generated Y_j^* (resp., $\tilde{Y}_j^*$) following (2.5) and (2.7) (resp. (2.4) and (2.6)), for different values of sp and se.

Our data set comes from the National Health and Nutrition Examination Survey carried out in the United States from 2015 to 2016 (NHANES (2017)). Note that we use this data set merely for illustration purposes, and ignore the sampling weights, as often occurs in this case. Our goal is to estimate $p(x) = E(D|X = x)$, where D is an indicator of the presence of the hepatitis B core antibody (HBcAb) for a patient, and X is the patient's age, ranging from 6 to 80 years. The sample size is $N = 8{,}021$, D is missing for 897 individuals, so that $N' = 7{,}124$, and no X is missing. The results of a point-biserial correlation coefficient test suggested a strong relationship between X and R^D. Thus, it seems reasonable to assume that the missing data mechanism depends on X, and we illustrate our techniques with MAR D on these data. Because p is unknown, we took our target curve to be $\hat{p}_{\text{ideal}}$, the estimator $\hat{p}_2$ computed from $Y_{i,j}$, with $\text{sp} = \text{se} = n = 1$.

The presence of the HBcAb indicates current or past infection by the hepatitis B virus. Several factors can influence prevalence in the general population; for example, baby boomers (people born during 1945–1965, aged 50 to 70 in the data set) are known to have higher prevalence, because the vaccine was approved in the United States only in 1982, and further infection controls started around 1992 (see Shing et al. (2020)). Moreover, all other factors being equal, older individuals have a greater chance of having been exposed to the virus. Reflecting this, the prevalence curve $\hat{p}_{\text{ideal}}$ increases with age, with a striking peak for patients in the age bracket 50–70, before decreasing again as age increases to 80.

We grouped the individuals as described above, either with $\text{sp} = \text{se} = 1$ or, to illustrate the effect of imperfect tests, with $\text{sp} = 0.995$ and $\text{se} = 0.95$, as in White et al. (2003). In each case, we randomly created 200 samples of $(X_{i,j}, \tilde{Y}_j^*, R_{i,j}^D)$ and $(X_{i,j}, Y_j^*, R_{i,j}^D)$, and calculated our estimators, $\hat{p}_1$, $\hat{p}_2$, and $\hat{p}_3$, as well as $\hat{p}_{\text{ungr},J}$ computed based on J nongrouped individuals selected randomly among N, with J equal to the number of groups used by $\hat{p}_2$, and the naive estimator $\hat{p}_{\text{nai}}$ from Section 7.2. Recall that $\hat{p}_1$ and $\hat{p}_2$ require knowing each missing status $R_{i,j}^D$, whereas $\hat{p}_3$ requires only $|I_j| = \sum_{i=1}^n R_{i,j}^D$.

We chose h, ψ_j, and K as in Section 7.2. To assess the performance of the estimators, denoted here generically by $\check{p}$, we calculated the integrated squared difference $\text{ISD} = \int_a^b \{\check{p}(x) - \hat{p}_{\text{ideal}}(x)\}^2 \, dx$, with a and b the 2.5% and 97.5% empirical quantiles, respectively of X. We summarize the ISDs in Table 2. In this example, $\hat{p}_2$ and $\hat{p}_3$ outperformed $\hat{p}_1$, $\hat{p}_{\text{ungr},J}$, and $\hat{p}_{\text{nai}}$, especially when $n = 4$

Table 2. Estimators of p for the hepatitis B data set with groupings (D) and (E). The numbers shown are the median (interquartile range) of the ISD$\times 10^3$ computed from 200 samples.

Grouping	$\hat{p}_1$	$\hat{p}_2$	$\hat{p}_3$	$\hat{p}_{\mathrm{ungr},J}$	$\hat{p}_{\mathrm{nai}}$
			sp = se = 1		
(D)	12.78 (12.77)	11.63 (10.95)	11.85 (11.93)	15.76 (17.01)	13.37 (14.33)
(E)	5.65 (5.26)	4.62 (4.14)	5.00 (4.79)	7.54 (7.69)	7.26 (6.54)
			sp = 0.995, se = 0.95		
(D)	13.51 (13.75)	11.68 (13.09)	12.82 (14.63)	17.30 (20.59)	13.58 (14.38)
(E)	6.09 (5.55)	5.29 (4.34)	5.49 (5.07)	8.67 (8.34)	7.38 (6.66)

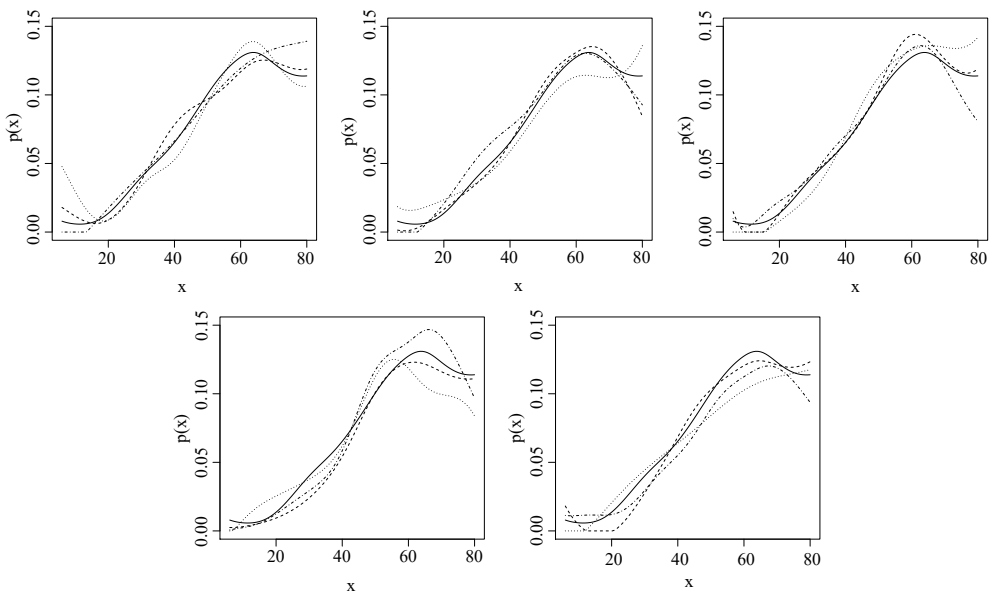

Figure 3. $\hat{p}_{\mathrm{ideal}}$ (—) for the hepatitis B data set, first (- - -), second (-·-·-), and third (· · ·) quartile estimated curves with sp = 0.995 and se = 0.95 for, from left to right, $\hat{p}_1$, $\hat{p}_2$, and $\hat{p}_3$ in the first row, and $\hat{p}_{\mathrm{ungr},J}$ and $\hat{p}_{\mathrm{nai}}$ in the second row, for grouping (E).

and $J = 2006$. Because the prevalence is low and the sample size is not extremely large, $\hat{p}_{\mathrm{ungr},J}$ fared worst, because very few of the J individuals were positive, making the estimation challenging. The same conclusions can be drawn from Fig. 3, which shows the estimated quartile curves corresponding to the samples for the first, second, and third quartiles of the 200 ISDs of $\hat{p}_1$, $\hat{p}_2$, $\hat{p}_3$, $\hat{p}_{\mathrm{ungr},J}$, and $\hat{p}_{\mathrm{nai}}$ with grouping (E). Overall, all estimators captured the increasing trend of prevalence with a peak in the bracket 50–70, followed by a decreasing trend. However, the naive estimator $\hat{p}_{\mathrm{nai}}$, which is biased, tended to flatten the peak, and $\hat{p}_3$ and $\hat{p}_{\mathrm{ungr},J}$ were more variable, especially $\hat{p}_{\mathrm{ungr},J}$ (because the prevalence is low, the J individuals contain too few positives to produce reliable estimators).

9. Extensions

In this section, we discuss a few interesting potential extensions of our methods. Note that we discuss only the main ideas; details, such as a fully data-driven implementation, are left for future research.

Our methods can be extended to the multivariate case of a $\mathfrak{d}$-dimensional covariate $\mathbf{X} \in \mathbb{R}^{\mathfrak{d}}$ by using a purely nonparametric approach, as in Delaigle and Meister (2011), or, to avoid the curse of dimensionality, by using single-index or partially linear models, as in Delaigle, Hall and Wishart (2014). These extensions are technical, but conceptually straightforward, because the main difficulty is expressing p in terms of a regression curve estimable from the data, which is identical to the univariate case explored here. For example, in the purely nonparametric case, to extend the local linear version of $\hat{p}_2$ to $\mathfrak{d}$ dimensions, it suffices to replace X with $\mathbf{X} = (X_1, \ldots, X_{\mathfrak{d}})^T$ in (2.3) and (2.7). Then, for $\mathbf{x} = (x_1, \ldots, x_{\mathfrak{d}})^T$, we can estimate $p(\mathbf{x}) = E(D_{i,j}|\mathbf{X}_{i,j} = \mathbf{x})$ by $\hat{p}_2(\mathbf{x}) = 1 - e_1^T \hat{\mathbf{S}}^{-1} \hat{\mathbf{T}}$, where $e_1^T = (1, 0, \ldots, 0)$, $\hat{\mathbf{S}} = (\hat{S}_{k,k'})_{0 \le k,k' \le \mathfrak{d}}$, and $\hat{\mathbf{T}} = (\hat{T}_0, \ldots, \hat{T}_{\mathfrak{d}})^T$, with $\hat{S}_{k,k'} = \sum_{j=1}^{J} \psi_j \sum_{i=1}^{n_j} R_{i,j}^D \mathbf{K}_{\mathbf{H}}(\mathbf{X}_{i,j} - \mathbf{x})(X_{i,j,k} - x_k)^{\delta_k}(X_{i,j,k'} - x_{k'})^{\delta_{k'}}$, and $\hat{T}_k = \sum_{j=1}^{J} \hat{q}_{RD}^{*\,n_j - 1}(Z_j^* + \mathrm{se} - 1)/(\mathrm{sp} + \mathrm{se} - 1)\psi_j \sum_{i=1}^{n_j} R_{i,j}^D \mathbf{K}_{\mathbf{H}}(\mathbf{X}_{i,j} - \mathbf{x})(X_{i,j,k} - x_k)^{\delta_k}$, where $\delta_k = \mathbb{1}(k > 0)$, $\mathbf{H} = \mathrm{diag}(h_1, \ldots, h_{\mathfrak{d}})$ is the bandwidth matrix (often taken to be a diagonal rescaled by the standard deviations of $X_{i,j,k}$), $\mathbf{K}$ is a $\mathfrak{d}$-dimensional kernel (e.g., a $\mathfrak{d}$-dimensional standard normal density), and $\mathbf{K}_{\mathbf{H}}(\mathbf{x}) = |\mathbf{H}|^{-1/2}\mathbf{K}(\mathbf{H}^{-1/2}\mathbf{x})$, with $|\mathbf{H}|$ the determinant of $\mathbf{H}$.

Such multivariate estimators may be useful when we can observe additional auxiliary variables $\mathbf{U} \in \mathbb{R}^{\mathfrak{d}-1}$ for the MAR assumption. See, for example, Wang, Rotnitzky and Lin (2010) for examples with nongrouped data, where the authors are interested in estimating a curve $p(x) = E(D|X = x)$, and assume that the MAR assumption holds with X and $\mathbf{U}$, that is,

$$P(R^D = r|X, \mathbf{U}, D) = P(R^D = r|X, \mathbf{U}) \quad \text{for } r = 0, 1\,.$$

They use a parametric model for $P(R^D = r|X, \mathbf{U})$ and a doubly robust method to mitigate the effect of incorrect parametric assumptions. In our case with grouped data, to avoid this parametric specification, we can use $p_{\mathrm{mult}}(x, \mathbf{u}) = E(D|X = x, \mathbf{U} = \mathbf{u})$, which can be estimated by $\hat{p}_{\mathrm{mult}}(x, \mathbf{u})$, one of the multivariate estimators discussed in the previous paragraph. Then, noting that $p(x) = E\{p_{\mathrm{mult}}(X, \mathbf{U})|X = x\}$, we can estimate $p(x)$ using a locally smoothed version of $\hat{p}_{\mathrm{mult}}(x, \mathbf{u})$, for example, $\hat{p}(x) = \sum_{j=1}^{J} \sum_{i=1}^{n_j} \hat{p}_{\mathrm{mult}}(X_{i,j}, \mathbf{U}_{i,j}) K_{h'}(x - X_{i,j}) \big/ \sum_{j=1}^{J} \sum_{i=1}^{n_j} K_{h'}(x - X_{i,j})$, with $h' > 0$ a bandwidth.

The local constant ($\ell = 0$) versions of our three estimators can also be extended to the case where X is discrete, by replacing the local weights $K_h(x - X_{i,j})$ with discrete weights $L(x, X_{i,j}, h)$. For example, if X takes c values $0, 1, \ldots, c - 1$, then following Racine and Li (2004), we can use $L(x, X_{i,j}, h) =$

$1\{X_{i,j} = x\} + h \cdot 1\{X_{i,j} \neq x\}$, where $h \in [0,1]$. More generally, if X has a natural ordering and $|X_{i,j} - x|$ is well defined, following Racine and Li (2004), we can use $L(x, X_{i,j}, h) = h^{|X_{i,j}-x|}$. In the bivariate case, where $\mathbf{X} = (X_1, X_2)$ with X_1 continuous and X_2 discrete, to estimate $p(\mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x})$, we can rather replace $K_h(X_{i,j} - x)$ with $K_h(X_{i,j,1} - x_1)L(x_2, X_{i,j,2}, \lambda)$, where $\lambda \in [0,1]$ and $h > 0$ are bandwidths.

Another interesting extension is the estimation of the prevalence conditional on X lying within a range of values $[a,b]$, that is, $p(a,b) = P(D = 1|X \in [a,b])$, where $a, b \in \mathbb{R}$. For example, when X denotes age, it is often of interest to consider prevalence given an age range. We have $p(a,b) = \int_a^b P(D = 1|X = x) f_X(x)\,dx / \int_a^b f_X(x)\,dx = E\{p(X)\mathbb{1}_{[a,b]}(X)\}/\{F_X(b) - F_X(a)\}$, where F_X denotes the distribution function of X. Therefore, we can estimate $p(a,b)$ by $\hat{p}(a,b) = \sum_{j=1}^{J}\sum_{i=1}^{n_j} \hat{p}(X_{i,j})\mathbb{1}_{[a,b]}(X_{i,j}) / \sum_{j=1}^{J}\sum_{i=1}^{n_j} \mathbb{1}_{[a,b]}(X_{i,j})$, where $\hat{p} = \hat{p}_1$, $\hat{p}_2$, or $\hat{p}_3$, depending on whether the setting is that of Section 4, 5.1, or 5.2, respectively.

Supplementary Material

The online supplementary material contains all technical proofs and numerical details.

Acknowledgments

The authors thank the referees for their helpful comments and suggestions. Delaigle's research was supported by the discovery project DP170102434 of the Australian Research Council (ARC). Tan's research was supported by the ARC Center of Excellence for Mathematical and Statistical Frontiers CE140100049 and by the China Scholarship Council.

References

Arevalo-Rodriguez, I., Buitrago-Garcia, D., Simancas-Racines, D., Zambrano-Achig, R., Del Campo, R., Ciapponi, A. et al. (2020). False-negative results of initial RT-PCR assays for COVID-19: A systematic review. *PloS one* **15**, e0242958.

Bethmann, J. (1989). The Lindberg-Feller theorem for sums of a random number of independent random variables in a triangular array. *Theory Probab. Appl.* **33**, 334–339.

Bilder, C., Iwen, P., Abdalhamid, B., Tebbs, J. and McMahan, C. (2020). Tests in short supply? Try group testing. *Significance* **17**, 15–16.

Bilder, C. R. and Tebbs, J. M. (2009). Bias, efficiency, and agreement for group-testing regression models. *J. Statist. Comput. Simul.* **79**, 67–80.

Chatterjee, A. and Bandyopadhyay, T. (2020). Regression models for group testing: Identifiability and asymptotics. *J. Statist. Plann. Inference* **204**, 141–152.

Chen, P., Tebbs, J. M. and Bilder, C. R. (2009). Group testing regression models with fixed and random effects. *Biometrics* **65**, 1270–1278.

Delaigle, A. and Hall, P. (2012). Nonparametric regression with homogeneous group testing data. *Ann. Statist.* **40**, 131–158.

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

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

Delaigle, A., Huang, W. and Lei, S. (2020). Estimation of conditional prevalence from group testing data with missing covariates. *J. Amer. Statist. Assoc.* **115**, 467–480.

Delaigle, A. and Meister, A. (2011). Nonparametric regression analysis for group testing data. *J. Amer. Statist. Assoc.* **106**, 640–650.

Delaigle, A. and Zhou, W.-X. (2015). Nonparametric and parametric estimators of prevalence from group testing data with aggregated covariates. *J. Amer. Statist. Assoc.* **110**, 1785–1796.

Dorfman, R. (1943). The detection of defective members of large populations. *Ann. Math. Statist.* **14**, 436–440.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications: Monographs on Statistics and Applied Probability 66*. CRC Press.

Gastwirth, J. and Hammick, P. (1989). Estimation of 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. Statist. Plann. Inference* **22**, 15–27.

Joyner, C. N., McMahan, C. S., Tebbs, J. M. and Bilder, C. R. (2020). From mixed effects modeling to spike and slab variable selection: A Bayesian regression model for group testing data. *Biometrics* **76**, 913–923.

Kim, H., Hudgens, M., Dreyfuss, J., Westreich, D. and Pilcher, C. (2007). Comparison of group testing algorithms for case identification in the presence of test error. *Biometrics* **63**, 1152—1163.

Kim, J. K. and Yu, C. L. (2011). A semiparametric estimation of mean functionals with nonignorable missing data. *J. Amer. Statist. Assoc.* **106**, 157–165.

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

Lin, J., Wang, D. and Zheng, Q. (2019). Regression analysis and variable selection for two-stage multiple-infection group testing data. *Stat. Med.* **38**, 4519–4533.

Little, R. J. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. Wiley, New York.

Liu, Y., McMahan, C. S., Tebbs, J. M., Gallagher, C. M. and Bilder, C. R. (2021). Generalized additive regression for group testing data. *Biostatistics* **22**, 873–889.

Maheu-Giroux, M., Joseph, L., Belisle, P., Lancione, S. and Eaton, J. W. (2017). Assessing the impact of imperfect immunoassays on HIV prevalence estimates from surveys conducted by the DHS Program. *DHS Methodological Reports No. 22*.

Malinovsky, Y. and Albert, P. S. (2019). Revisiting nested group testing procedures: New results, comparisons, and robustness. *Amer. Statist.* **73**, 117–125.

Mallapaty, S. (2020). The mathematical strategy that could transform coronavirus testing. *Nature* **583**, 504–505.

McMahan, C. S., Tebbs, J. M., Hanson, T. E. and Bilder, C. R. (2017). Bayesian regression for group testing data. *Biometrics* **73**, 1443–1452.

Miao, W., Ding, P. and Geng, Z. (2016). Identifiability of normal and normal mixture models with nonignorable missing data. *J. Amer. Statist. Assoc.* **111**, 1673–1683.

Miao, W., Tchetgen Tchetgen, E. J. and Geng, Z. (2015). Identification and doubly robust estimation of data missing not at random with a shadow variable. *arXiv:1509.02556*.

Molenberghs, G., Fitzmaurice, G., Kenward, M. G., Tsiatis, A. and Verbeke, G. (2014). *Handbook of Missing Data Methodology*. CRC Press.

Montesinos-López, O. A., Eskridge, K., Montesinos-López, A., Crossa, J., Cortés-Cruz, M. and Wang, D. (2016). A regression model for pooled data in a two-stage survey under informative sampling with application for detecting and estimating the presence of transgenic corn. *Seed Sci. Res.* **26**, 182–197.

Mutesa, L., Ndishimye, P., Butera, Y., Souopgui, J., Uwineza, A., Rutayisire, R. et al. (2021). A pooled testing strategy for identifying SARS-CoV-2 at low prevalence. *Nature* **589**, 276–280.

NHANES (2017). Hepatitis B: Core antibody, Surface antigen, and Hepatitis D antibody. Hyattsville, MD: HHS, CDC. Web: `wwwn.cdc.gov/nchs/nhanes/Search/DataPage.aspx?Component=Laboratory&CycleBeginYear=2015`.

Racine, J. and Li, Q. (2004). Nonparametric estimation of regression functions with both categorical and continuous data. *J. Econometrics* **119,** 99–130.

Rao, C. R. (1973). *Linear Statistical Inference and its Applications*. Wiley, New York.

Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63**, 581–592.

Serfling, R. J. (2009). *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons.

Shing, J. Z., Ly, K. N., Xing, J., Teshale, E. H. and Jiles, R. B. (2020). Prevalence of hepatitis B virus infection among US adults aged 20-59 years with a history of injection drug use: National health and nutrition examination survey, 2001–2016. *Clin. Infect. Dis.* **70**, 2619–2627.

Sun, B., Liu, L., Miao, W., Wirth, K., Robins, J. and Tchetgen Tchetgen, E. J. (2018). Semiparametric estimation with data missing not at random using an instrumental variable. *Statist. Sinica* **28**, 1965–1983.

Surkova, E., Nikolayevskyy, V. and Drobniewski, F. (2020). False-positive COVID-19 results: hidden problems and costs. *Lancet Respir. Med.* **8**, 1167–1168.

Tchetgen Tchetgen, E. J. and Wirth, K. E. (2017). A general instrumental variable framework for regression analysis with outcome missing not at random. *Biometrics* **73**, 1123–1131.

Vansteelandt, S., Goetghebeur, E. and Verstraeten, T. (2000). Regression models for disease prevalence with diagnostic tests on pools of serum samples. *Biometrics* **56**, 1126–1133.

Wang, D., Zhou, H. and Kulasekera, K. (2013). A semi-local likelihood regression estimator of the proportion based on group testing data. *J. Nonparametr. Stat.* **25**, 209–221.

Wang, L., Rotnitzky, A. and Lin, X. (2010). Nonparametric regression with missing outcomes using weighted kernel estimating equations. *J. Amer. Statist. Assoc.* **105**, 1135–1146.

White, R., Delieu, E., Perry, K. and Parry, J. (2003). *Four anti-HBc Assays*. Medicines and Healthcare products Regulatory Agency.

Xie, M. (2001). Regression analysis of group testing samples. *Stat. Med.* **20**, 1957–1969.

Yuan, A., Piao, J., Ning, J. and Qin, J. (2021). Semiparametric isotonic regression modelling and estimation for group testing data. *Canad. J. Statist.* **49**, 659–677.

Zhang, B., Bilder, C. R. and Tebbs, J. M. (2013). Regression analysis for multiple-disease group testing data. *Stat. Med.* **32**, 4954–4966.

Zhou, Y., Wan, A. T. K. and Wang, X. (2008). Estimating equations inference with missing data. *J. Amer. Statist. Assoc.* **103**, 1187–1199.

Aurore Delaigle
School of Mathematics and Statistics, University of Melbourne, Parkville, Vic, 3010, Australia.
E-mail: aurored@unimelb.edu.au
Ruoxu Tan
School of Mathematics and Statistics, University of Melbourne, Parkville, Vic, 3010, Australia.
E-mail: ruoxut@outlook.com

(Received November 2021; accepted May 2022)