An Information Theoretic Approach to Prevalence Estimation and Missing Data

Ola Hössjer, Daniel Andrés Díaz-Pachón, *Member, IEEE*, Chen Zhao, and J. Sunil Rao

***Abstract*— Many data sources, including tracking social behavior to election polling to testing studies for understanding disease spread, are subject to sampling bias whose implications are not fully yet understood. In this paper we study estimation of a given feature (such as disease, or behavior at social media platforms) from biased samples, treating non-respondent individuals as missing data. Prevalence of the feature among sampled individuals has an upward bias under the assumption of individuals' willingness to be sampled. This can be viewed as a regression model with symptoms as covariates and the feature as outcome. It is assumed that the outcome is unknown at the time of sampling, and therefore the missingness mechanism only depends on the covariates. We show that data, in spite of this, is missing at random only when the sizes of symptom classes in the population are known; otherwise data is missing not at random. With an information theoretic viewpoint, we show that sampling bias corresponds to external information due to individuals in the population knowing their covariates, and we quantify this external information by active information. The reduction in prevalence, when sampling bias is adjusted for, similarly translates into active information due to bias correction, with opposite sign to active information due to testing bias. We develop unified results that show that prevalence and active information estimates are asymptotically normal under all missing data mechanisms, when testing errors are absent and present respectively. The asymptotic behavior of the estimators is illustrated through simulations.**

***Index Terms*— Active information, asymptotic normality, biased estimate, missing data, testing errors.**

I. Introduction

According to the No Free Lunch Theorems, in a search problem, on average, no search does better than blind [1]. Therefore, when for a particular case one search does different than a uniform search (better or worse), it is because the programmer used her knowledge (good or bad) either of the target or the structure of the space, or both. Active information was introduced to measure the amount of information a programmer infuses in a search to reach the target with different probability than through a blind search [2], [3]. For a search space $\mathcal{X}$ and a target $A \subset \mathcal{X}$, active information is then naturally defined as $I^+ = \log(p/p_0)$, where p is the probability of reaching A under the algorithm devised by the programmer, and p_0 is the uniform probability of reaching A.

Another interpretation of active information will allow to see that a data set in $\mathcal{X}$, whose distribution is consistent with a probability p of reaching A, will have a local mode in the region A if $I^+ > 0$ [4], [5]. Montañez and collaborators have also used active information to analyze intention perception [6]. Díaz-Pachón and Hössjer have used active information to measure fine-tuning [7]. And Díaz-Pachón and Marks II used it to compare non-neutral to neutral population genetics models [8].

In this paper active information is used to unify estimation and bias correction of the prevalence p_0 of a particular feature in a population when data is missing. This corresponds to a setting where $\mathcal{X}$ is a population of individuals whereas A is the subpopulation of $p_0|\mathcal{X}|$ individuals that have the feature. It is assumed that a prevalence estimate is computed from a biased subsample of so-called "tested" individuals and that the data analyst does not control the sampling scheme. Due to external information among the tested individuals, the fraction of them that have the feature of interest, p, is typically larger than p_0, corresponding to a positive active information $I_T^+ = \log(p/p_0)$ due to testing bias.

For example, [9] studied the potential biases in big data focusing on features such as social behavior patterns from social media data platforms. They found, after analyzing survey data from a national sample of American adults' social network usage, that those individuals with higher socioeconomic status and greater internet skills are more likely to be on several different platforms. Another example was the use of social media platforms such as Twitter to share features such as personal health information, including their COVID-19 related sentiments and comments. Official agencies have also used these types of platforms to share policies and research progress. While such data sources can provide new opportunities for health-related research, potential significant sampling bias can appear whereby like-minded individuals seek each other out thereby limiting full assessment of opinions [10]. This can distort prevalence estimates greatly. Another example comes from online polling. Given that so many individuals

Manuscript received 22 February 2023; revised 13 September 2023; accepted 18 October 2023. Date of publication 25 October 2023; date of current version 23 April 2024. This work was supported in part by NSF under Grant 2137148. *(Corresponding author: Daniel Andrés Díaz-Pachón.)*

Ola Hössjer is with the Department of Mathematics, Stockholm University, 106 91 Stockholm, Sweden (e-mail: ola@math.su.se).

Daniel Andrés Díaz-Pachón and Chen Zhao are with the Division of Biostatistics, University of Miami, Miami, FL 33136 USA (e-mail: Ddiaz3@miami.edu; zhao.c@miami.edu).

J. Sunil Rao is with the Division of Biostatistics, University of Minnesota, Minneapolis, MN 55455 USA (e-mail: js-rao@umn.edu).

Communicated by L. Lai, Associate Editor for Signal Processing and Source Coding.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIT.2023.3327399.

Digital Object Identifier 10.1109/TIT.2023.3327399

are now online and the high cost of phone-based surveys, the exercise of querying people about features such as their voter preferences online has become increasingly popular. However, sampling biases for some of the reasons mentioned above means that these polls may not reflect views from the overall population resulting in biased prevalence estimates of say voter preference [11].

Finally, [12] examined the effect of sampling bias from COVID-19 testing studies on prevalence estimation. Since individuals with stronger symptoms are more likely to have the disease and get tested, their knowledge of these symptoms represents information that leads to an estimated prevalence with an upward bias $p - p_0$. From the point of view of the data analyst conducting the study, individuals' knowledge of their symptoms represents external information, and it quantifies the degree at which individuals' willingness to be tested correlates with their symptoms, which in turn correlate with disease status. In the more general settings of biased sampling discussed above, symptom status can indicate depth of support for a particular political candidate, or degree of attention that draws an individual to a social media platform. Inspired by this approach, in what follows we sometimes use the words "feature" or "infection/disease status" for the output, and "symptoms" for the input or covariates.

Our incomplete testing framework is regarded as a regression model with missing data [13], [14], [15], where the missingness mechanism depends on covariates but not on outcome variables. In the context of disease testing, for instance, disease-like symptoms are the covariates of the regression model, whereas disease status is the outcome. The missing data then consists of all or some information from those individuals that are not tested, analogously to non-respondents of surveys. For the simplest missingness mechanism, when data is missing completely at random (MCAR), there is no bias of the prevalence estimate due to testing, i.e. $I_T^+ = 0$. More realistically, some individuals are more likely to "test" themselves based on their "symptoms" (the covariates) not based on their unknown feature status (the outcome). In this context, we show that data is missing at random (MAR) only when the sizes of all symptom classes in the population are known, corresponding to a scenario where the data analyst knows symptom status in the whole population. On the other hand data is missing not a random (MNAR) when the sizes of these symptom classes are unknown, corresponding to a scenario where the data analyst does not know the symptom status of untested individuals. Note that this is a different type of MNAR scenario than typically dealt with in the missing data literature ([16]), where the missingness mechanism not only depends on covariates but also on outcomes, whereas the covariates are known not only among the respondents, but also among the non-respondents. Here we focus on the opposite MNAR scenario, with response probabilities only depending on covariates (not the outcomes), but on the other hand these covariates are only known among the respondents.

For many of the applications we have in mind, the MNAR scenario with unknown symptom classes is more realistic than the MAR scenario with known symptom classes. However, we can think of at least two situations when MAR sampling is useful: 1) When there are only two symptom classes and all individuals with stronger symptoms are required to be tested. 2) When symptoms are interpreted more generally, for instance as characteristics that can be retrieved from population registries. Such "symptoms" are known, whether the individuals of the population are tested for the feature of interest or not.

For sampling scenarios other than MCAR, the bias of the prevalence estimate due to testing bias can typically be corrected to some degree. This is quantified by means of a negative active information I_C^+ due to bias correction, and a total active information $I^+ = I_T^+ + I_C^+$ after bias correction that typically satisifes $0 \leq I^+ \leq I_T^+$. We show that the bias of the prevalence estimate can be removed for MAR scenarios ($I_C^+ = -I_T^+$ or $I^+ = 0$), whereas this is typically not the case of MNAR scenarios ($0 < I^+ < I_T^+$).

We derive asymptotic normality results for estimates $\hat{p}$ and $\hat{p}_0$ of the prevalence before and after correction for testing bias, as well as asymptotic normality for the estimate $\hat{I}_T^+ = \log(\hat{p}/\hat{p}_0)$ of active information due to testing bias. These results involve the above mentioned MCAR, MAR, and MNAR scenarios, with and without presence of testing errors. Whereas $\hat{p}$ is an asymptotically unbiased estimate of p for all sampling scenarios without systematic testing errors, $\hat{p}_0$ is an asymptotically unbiased estimate of p_0 only for MCAR and MAR sampling schemes. This implies that the associated confidence intervals of p_0 only have correct asymptotic coverage probabilities for MCAR and MAR scenarios.

II. Active Information Due to Testing Bias

Let $\mathcal{X}$ be a population of $N = |\mathcal{X}|$ individuals of which those in $A \subset \mathcal{X}$ have a specific value 1 of a binary feature (such as presence of a disease, or a specific type of behavior at a social media platform), whereas the other subjects in $A^c = \mathcal{X} \setminus A$ have the other feature value 0 (no disease or absence of the behavior). Let P_0 refer to the uniform probability measure on $\mathcal{X}$, which assigns a probability of $1/N$ to each individual. The objective is to estimate the population prevalence

$$p_0 = P_0(A) = \frac{|A|}{N} \tag{1}$$

of the feature value 1 from a subgroup of individuals that are tested. To this effect, first divide

$$\mathcal{X} = \cup_{s=0}^{S-1} \cup_{i=0}^{1} \mathcal{X}_{si}, \tag{2}$$

into a number of subpopulations of unknown sizes $|\mathcal{X}_{si}| = N\rho_{si}$, where $\mathcal{X}_{si}$ consists of those individuals with symptoms $s \in \{0, \ldots, S-1\}$ and feature status $i \in \{0, 1\}$. The first variable s is measured on an ordinal scale with increasingly stronger symptoms for feature value 1, so that $s = 0$ represents no symptoms whereas $s = S - 1$ codes for the strongest possible symptoms. Feature status, on the other hand, is a binary variable such that $i = 0$ and $i = 1$ correspond to a "non-infected" and "infected" individual, respectively. For each $x \in \mathcal{X}$ we let

$$I(x) = (I_1(x), I_2(x))$$

$$\in \{(0,0),\dots,(S-1,0),(0,1),\dots,(S-1,1)\} \quad (3)$$

signify the subpopulation $\mathcal{X}_{si}$ to which x belongs.

Let also T_x be a variable that equals 1 or 0 depending on whether x is "tested" for the feature or not. If the individual is tested, this could mean that he either enters a medical lab for getting to know his disease status, or enters a social media platform. In any case, prior to testing the individual knows his symptoms but not his feature. The collection $\{T_x;\, x \in \mathcal{X}\}$ is assumed to be formed by independent Bernoulli variables, with $P(T_x = 1) = \pi_{I(x)}$. This corresponds to an assumption whereby individuals in different groups are tested with different sampling probabilities π_{si}. Consequently, the weighted probability measure

$$P(x) = \frac{\pi_{I(x)}}{\sum_{y\in\mathcal{X}} \pi_{I(y)}}, \quad x \in \mathcal{X} \quad (4)$$

represents a prediction of the tested population, before testing has occurred. In particular, the testing prevalence

$$p = P(A) = \sum_{x\in A} P(x) \quad (5)$$

is the expected prevalence of feature value 1 in the tested subpopulation. The active information due to testing bias is defined as

$$I_T^+ = \log\frac{p}{p_0} = \log\frac{P(A)}{P_0(A)}. \quad (6)$$

To estimate p and I_T^+, the subpopulation

$$\mathcal{X}_T = \{x \in \mathcal{X};\, T_x = 1\} \quad (7)$$

of $N_T = |\mathcal{X}_T|$ tested individuals is introduced. Since N_T is known, this gives rise to an estimator

$$\hat{p} = \hat{p}(A) = \frac{|A \cap \mathcal{X}_T|}{N_T} \quad (8)$$

of p. The expected fraction of sampled individuals is also introduced as

$$\pi = \sum_{i=0}^{1}\sum_{s=0}^{S-1} \rho_{si}\pi_{si}, \quad (9)$$

which is estimated by

$$\hat{\pi} = \frac{N_T}{N}. \quad (10)$$

From the data analyst's point of view, symptom and feature status $I(x) = (I_1(x), I_2(x))$ is known for those individuals x that are tested ($x \in \mathcal{X}_T$), whereas feature status $I_2(x)$ is unknown for those individuals x that are not tested ($x \in \mathcal{X} \setminus \mathcal{X}_T$). This can be summarized by letting $\delta(x) = (\delta_1(x), \delta_2(x))$ be a binary vector of length two that indicates whether symptoms and feature status of $x \in \mathcal{X}$ is known (1) or not (0). Thus we have that

$$\begin{aligned} \delta_1(x) = \delta_2(x) &= 1, \quad \forall x \in \mathcal{X}_T, \\ \delta_2(x) &= 0, \quad \forall x \in \mathcal{X} \setminus \mathcal{X}_T. \end{aligned} \quad (11)$$

We further assume that symptom status is either known (for instance from public registers) for all individuals that are not tested, or unknown for all such individuals. This corresponds to

$$\delta_1(x) \text{ has the same value (0 or 1) } \forall x \in \mathcal{X} \setminus \mathcal{X}_T. \quad (12)$$

The missingness mechanism is defined as

$$\pi_{si} = P(\delta(x) = (1,1) | I(x) = (s,i)) \quad (13)$$

In the next section we will show that the value of $\delta_1(x)$ in (12) and the form of π_{si} in (13) influences whether data is missing at random (MAR) or not (MNAR). In the first case (MAR) feature status from $\mathcal{X} \setminus \mathcal{X}_T$ is regarded as missing data, whereas in the second case (MNAR) symptom and feature status from $\mathcal{X} \setminus \mathcal{X}_T$ is missing. Note in particular that if no data is missing, i.e. $\mathcal{X}_T = \mathcal{X}$, then $\hat{p} = p_0$. In the next section we will regard $\hat{p}$ as an estimator of p_0 that is biased whenever data is missing.

III. Active Information After Bias Correction

The relation between p and p_0 depends crucially on the sampling probabilities π_{si}. This can be seen by noting that the population and testing prevalences are different functions

$$p_0 = \sum_{s=0}^{S-1} \rho_{s1}, \quad p = \sum_{s=0}^{S-1} \rho_{s1}\pi_{s1} \Big/ \sum_{s,i} \rho_{si}\pi_{si} \quad (14)$$

of $\rho_{01},\dots,\rho_{S-1,1}$. Regarding non-tested individuals as missing data, concepts from the missing data literature [14] are helpful to explain the way in which data is missing. Random sampling, or data missing completely at random (MCAR), occurs when

$$\pi_{si} = \pi. \quad (15)$$

From (14), $p = p_0$ and $I_T^+ = 0$ whenever (15) holds. Condition (15) is usually very unrealistic, since people with stronger symptoms (larger s) are more likely to be tested (have larger π_{s0} and π_{s1}) than those with weaker symptoms. A weaker assumption of data missing at random (MAR) occurs when the sampling probabilities only depend on variables that are known. In an example of a MAR sampling scheme

$$\rho_s = \rho_{s0} + \rho_{s1} \text{ is known} \quad (16)$$

and

$$\pi_{si} = \pi_s \quad (17)$$

for $s = 0,\dots,S-1$. The first MAR condition (16) follows if (12) holds with $\delta_1(x) = 1$, that is, if symptom status is known among all non-tested individuals. The second MAR condition (17) implies that the sampling probability (13) only depends on symptom status, not the feature. The most challenging missingness mechanism (neither MCAR or MAR) is referred to as data missing not at random (MNAR). Note that data are MNAR when at least one of (16) and (17) fails.

Also from (14), typically $p \neq p_0$ and $I_T^+ \neq 0$ when data is MAR or MNAR. To construct a bias-corrected estimator $\hat{p}_0$ of p_0, notice first that the biased prevalence estimator (8) can be rewritten as

$$\hat{p} = \frac{\sum_{s=0}^{S-1} \rho_{s1}\tilde{\pi}_{s1}}{\sum_{s,i} \rho_{si}\tilde{\pi}_{si}} = \sum_{s=0}^{S-1} \rho_{Ts1}, \quad (18)$$

where the sampling fractions

$$\tilde{\pi}_{si} = \frac{|\mathcal{X}_T \cap \mathcal{X}_{si}|}{|\mathcal{X}_{si}|} = \frac{|\mathcal{X}_{Tsi}|}{|\mathcal{X}_{si}|} = \frac{N_{Tsi}}{N_{si}} = \frac{N_{Tsi}}{N\rho_{si}} \tag{19}$$

for different subpopulations approximate π_{si}, whereas

$$\rho_{Tsi} = \frac{\rho_{si}\tilde{\pi}_{si}}{\sum_{r,k}\rho_{rk}\tilde{\pi}_{rk}} = \frac{N_{Tsi}}{N_T} \tag{20}$$

are the known fractions at which the subpopulations appear in the sample. A comparison between (14) and (20) suggests an estimate

$$\hat{p}_0 = \frac{\sum_{s=0}^{S-1}\rho_{Ts1}\hat{\pi}_{s1}^{-1}}{\sum_{s,i}\rho_{Tsi}\hat{\pi}_{si}^{-1}} = \frac{\sum_{s=0}^{S-1}N_{Ts1}\hat{\pi}_{s1}^{-1}}{\sum_{s,i}N_{Tsi}\hat{\pi}_{si}^{-1}} \tag{21}$$

of the population prevalence p_0, where $\hat{\pi}_{si}$ is an estimate of $\tilde{\pi}_{si}$ (and thereby also an estimate of π_{si}). From the context of survey sampling [17] and survey methodology [18], it is possible to rewrite (21) as a Horvitz-Thompson type weighted average

$$\hat{p}_0 = \frac{\sum_{x\in\mathcal{X}} w_x I_2(x)}{\sum_{x\in\mathcal{X}} w_x} \tag{22}$$

of the outcome variables $I_2(x) \in \{0,1\}$ of all individuals in the population, with "inverse inclusion probability" weights

$$w_x = \delta_1(x)\delta_2(x)\hat{\pi}_{I(x)}^{-1} = \begin{cases}\hat{\pi}_{I(x)}^{-1}, & x \in \mathcal{X}_T \cap \mathcal{X}_{I(x)},\\ 0, & x \notin \mathcal{X}_T.\end{cases} \tag{23}$$

These weights are nonzero for all tested individuals (the respondents), with values depending on which strata $\mathcal{X}_{si}$ they belong to.

With estimates (8) and (21) of p and p_0 defined, we plug these estimates into (6), in order to obtain an estimator

$$\hat{I}_T^+ = \log\frac{\hat{p}}{\hat{p}_0} \tag{24}$$

of the active information I_T^+ due to testing bias. Introduce the notation $\bar{p}_0$ for $E(\hat{p}_0)$, or an asymptotic (large N) approximation of this expected value. We will refer to

$$I^+ = \log\frac{\bar{p}_0}{p_0} = \log\frac{p}{p_0} + \log\frac{\bar{p}_0}{p} = I_T^+ + I_C^+ \tag{25}$$

as the active information of the bias-adjusted prevalence estimate (21), which is a sum of two terms: the active information (6) due to testing bias and the active information I_C^+ due to bias correction. If the bias correction is completely successful ($I^+ = 0$), then $I_C^+ = -I_T^+$. This suggests an estimate

$$\hat{I}_C^+ = -\hat{I}_T^+ = -\log\frac{\hat{p}}{\hat{p}_0} \tag{26}$$

of I_C^+.

IV. Examples

In this section we will illustrate how to construct the bias-adjusted estimator $\hat{p}_0$ in (21), of the prevalence p_0 of feature value 1, for a number of sampling schemes.

*Example 1 (*MCAR*):* Whenever (15) holds, $I_T^+ = I_C^+ = 0$ follows from (6) and (14). In this context, to assume that the estimated sampling fractions $\hat{\pi}_{si}^{\text{MCAR}} = \hat{\pi}^{\text{MCAR}}$ are the same for all subpopulations $\mathcal{X}_{si}$ is natural. Since $\hat{\pi}^{\text{MCAR}}$ cancels out in the prevalence estimator (21), it simplifies to

$$\begin{aligned}\hat{p}_0^{\text{MCAR}} &= \frac{\sum_{s=0}^{S-1}\rho_{Ts1}}{\sum_{s,i}\rho_{Tsi}} = \sum_{s=0}^{S-1}\rho_{Ts1}\\ &= \frac{1}{N_T}\sum_{s=0}^{S-1}N_{Ts1} = \frac{N_{T\cdot 1}}{N_T}\\ &= \hat{p}.\end{aligned} \tag{27}$$

Note that the survey sampling estimator (22) of p_0 simplifies to (27) when $\hat{\pi}_{I(x)} = \hat{\pi}^{\text{MCAR}}$, independently of x, in the definition of the sampling weight w_x of x in (23). It also follows from (27) that $\hat{I}_T^+ = \hat{I}_C^+ = 0$ under MCAR sampling. From Fisher's exact test, $N_{T\cdot 1}$ has a hypergeometric distribution

$$N_{T\cdot 1} \mid N_T \sim \text{Hyp}(N, N_T, p_0) \tag{28}$$

conditionally on N_T. Taking expectations in both sides of (27), by (20) and (18), $E\left(\hat{p}_0^{\text{MCAR}}\right) = p_0$ and $I^+ = 0$.

*Example 2 (*MAR*):* The MAR sampling scheme (16)-(17) can be viewed as an instance of stratified sampling [18], where the relative sizes ρ_s of the strata (symptom classes) are known. Although the sampling fractions $\tilde{\pi}_{si}$ in (19) are unknown when (17) holds, they may be estimated consistently by means of

$$\hat{\pi}_{si}^{\text{MAR}} = \hat{\pi}_s^{\text{MAR}} = \frac{N_{Ts0}+N_{Ts1}}{N\rho_{s0}+N\rho_{s1}} = \frac{N_{Ts}}{N\rho_s}, \tag{29}$$

where in the last step $N_{Ts} = N_{Ts0} + N_{Ts1}$ was introduced. Plugging (29) into (21), the estimator

$$\begin{aligned}\hat{p}_0^{\text{MAR}} &= \frac{\sum_{s=0}^{S-1}\rho_{Ts1}(N\rho_s/N_{Ts})}{\sum_{s,i}\rho_{Tsi}(N\rho_s/N_{Ts})}\\ &= \frac{\sum_{s=0}^{S-1}\rho_s(N_{Ts1}/N_{Ts})}{\sum_{s=0}^{S-1}\rho_s}\\ &= \sum_{s=0}^{S-1}\rho_s\hat{p}_{0s}\end{aligned} \tag{30}$$

of p_0 is obtained. It is a weighted average of estimates

$$\hat{p}_{0s} = \frac{N_{Ts1}}{N_{Ts}} = \frac{\sum_{x\in\mathcal{X}} I_2(x)1_{\{x\in\mathcal{X}_{Ts}\}}}{\sum_{x\in\mathcal{X}} 1_{\{x\in\mathcal{X}_{Ts}\}}} \tag{31}$$

of the prevalences

$$p_{0s} = \frac{\rho_{s1}}{\rho_s} \tag{32}$$

in symptom classes $\mathcal{X}_s = \mathcal{X}_{s0} \cup \mathcal{X}_{s1}$, using data from cohorts $\mathcal{X}_{Ts} = \mathcal{X}_{Ts0} \cup \mathcal{X}_{Ts1}$. In the last step of (31) we also made use of notation $1_{\{x\in\mathcal{X}_{Ts}\}}$, a term that equals 1 when $x \in \mathcal{X}_{Ts}$ and 0 otherwise. The estimator (30) is well known from stratified sampling. It is also an instance of the survey sampling estimator (22), in the special case when the testing probabilities satisfy the second MAR condition $\pi_{s0} = \pi_{s1} = \pi_s$ in (17). Indeed, putting $I_1(x) = s$, this MAR condition (17) implies that (31) is equivalent to (22), when the inverse of $\hat{\pi}_{I(x)} = \hat{\pi}_{I_1(x)} = N_{Ts}/N_s$ is used in the definition of the inverse probability sampling weights w_x of (23). Since $\pi_{s0} = \pi_{s1} = \pi_s$, from Fisher's exact test,

$$N_{Ts1} \mid N_{Ts} \sim \text{Hyp}(N_s, N_{Ts}, p_{0s}) \tag{33}$$

for $s = 0, \ldots, S-1$. In view of (14), this implies

$$E\left(\hat{p}_0^{\mathrm{MAR}}\right) = \sum_{s=0}^{S-1} \rho_s E(\hat{p}_{0s}) = \sum_{s=0}^{S-1} \rho_s p_{0s} = p_0. \tag{34}$$

Consequently, $I^+ = 0$ under MAR sampling, although in general $I_T^+ = -I_C^+$ differs from zero. Thus $I^+ = 0$ is a consequence of the fact that $\hat{p}_0^{\mathrm{MAR}}$ is an unbiased method-of-moments estimator of p_0. In principle it is also possible to use an asymptotically unbiased maximum likelihood estimator of p_0. We prefer however to use (30), since it guarantees an unbiased estimate of p_0, so that $I^+ = 0$ under MAR conditions.

The next two examples treat sampling MNAR schemes, where one of the two MAR conditions (16)-(17) fail.

*Example 3 (*MNAR *with known strata sizes.):* Suppose the first MAR condition (16) of known strata sizes holds, which corresponds to covariates (symptoms) being known in the whole population $\mathcal{X}$. On the other hand we assume that the second MAR condition (17) fails. This corresponds to a MNAR scenario where the response probability π_{si} not only depends on covariate information (symptoms s) but also on the outcome variable (feature status i). Under this setting it is not possible to estimate prevalences within symptom classes as in (31). It is possible though to use a maximum likelihood approach that accounts for the missingness mechanism (in our case the response probabilities π_{si}). Either a joint likelihood is defined from a two-step procedure, where first a model for the joint distribution of the outcomes before non-response is defined, and second a model of non-response is imposed [19], [20], [21], [22]. A second option is to merge these two steps into a likelihood for one single model [16], [23], [24], [25]. A third likelihood-based approach [26] is to first estimate parameters of the regression model for the respondents (in our setting the prevalence within each symptom class, among those that are tested), then estimate response probabilities (in our case π_{si}) and finally estimate the parameter of interest (in our model the prevalence p_0).

*Example 4 (*MNAR *With Unknown Strata Sizes):* Even though the MNAR estimation techniques of Example 3 are well developed, their underlying assumptions are somewhat less suitable in the context of estimating the prevalence of a disease from medical labs or social behaviour from internet platforms. Indeed, it is usually the case that while each individuals knows his symptoms, these are usually not known by the data analyst in the whole population, so that (16) fails. In addition, it seems realistic to retain (17). That is, it seems plausible to assume that response probabilities only depend on symptoms, not on the feature itself, when each individual's feature is unknown to him at the time of sampling. Such a MNAR scenario is the topic of this example.

When symptom class sizes $N\rho_s$ are unknown, it is possible to employ a double sampling approach [27], where the first sample is used to find estimates of ρ_s, whereas the second sample is used to estimate prevalences p_{0s} of all symptom classes s. A second option is to apply the data integration integration method [28], whereby estimates of all ρ_s are obtained from another independent probability sampling. A third way of handling missing covariate information for non-respondents is to define a respondents' likelihood for the tested individuals [29].

Here we will rather use a Bayesian approach when symptom classes ρ_s are not known. We will illustrate this in the context of COVID-19 testing. As in [12] and [30], we consider a model of sampling with $S = 2$ symptoms, with strata fractions ρ_0 and ρ_1 being unknown, so that the first MAR condition (16) fails. On the other hand, it is assumed in these two articles that the second MAR condition (17) holds, with

$$\pi_{00} = \pi_{01} = \pi_0, \quad \pi_{10} = \pi_{11} = \pi_1, \quad \pi_1 > \pi_0. \tag{35}$$

The inequality of (35) adds another assumption to (17); that symptomatic individuals are more likely to get tested than asymptomatic ones. This implies that with high probability $\rho_1 < N_{T1}/N_T$. On the other hand, the presence of N_{T1} symptomatic individuals in the sample implies that $N_{T1}/N \leq \rho_1$.

From a Bayesian point of view, it is natural to interpret our incomplete knowledge about the proportion of symptomatic individuals, as a random variable ρ_1 whose distribution is supported on the interval $(N_{T1}/N, N_{T1}/N_T)$. If we are maximally ignorant about this distribution, the maximum entropy principle [31], [32] tells that ρ_1 should have a uniform distribution on $(N_{T1}/N, N_{T1}/N_T)$. Therefore, $\hat{\rho}_1 = E(\rho_1)$ is taken as the estimator of the proportion of symptomatic individuals in the population, and $\hat{\rho}_0 = 1 - \hat{\rho}_1 = E(\rho_0)$ estimates the proportion of the asymptomatic group. From this viewpoint, a modification of (29) produces

$$\hat{\pi}_{si}^{\mathrm{MaxEnt}} = \hat{\pi}_s^{\mathrm{MaxEnt}} = \frac{N_{Ts}}{NE(\rho_s)}, \tag{36}$$

and plugging (36) into (21), the estimator of prevalence is

$$\begin{aligned}\hat{p}_0 &= \frac{N_{T01}(\hat{\pi}_0^{\mathrm{MaxEnt}})^{-1} + N_{T11}(\hat{\pi}_1^{\mathrm{MaxEnt}})^{-1}}{N_{T0}(\hat{\pi}_0^{\mathrm{MaxEnt}})^{-1} + N_{T1}(\hat{\pi}_1^{\mathrm{MaxEnt}})^{-1}} \\ &= \frac{N_{T01}}{N_{T0}}(1-\hat{\rho}_1) + \frac{N_{T11}}{N_{T1}}\hat{\rho}_1,\end{aligned} \tag{37}$$

where (37) is obtained using the first two assumptions of (35), which imply that inside each group of symptoms the sampling of infected and non-infected is random.

V. Asymptotics

This section is focused on the asymptotic properties of the estimates $\hat{p}$ and $\hat{p}_0$ of the test-biased and population-based prevalences p and p_0 of feature value 1, as the population size N gets large. The second MAR condition (17) is assumed, so that sampling probabilities only depend on covariates, not on outcomes (features). On the other hand, the first MAR condition (16) (that symptom class sizes are known) may fail. The asymptotic theory will therefore cover Examples 1, 2, and 4, but not Example 3.

Equation (17) allows for a number of simplifications. It first of all implies that the expected fraction of tested individuals in (9) takes the form

$$\pi = \sum_{s=0}^{S-1} \rho_s \pi_s. \tag{38}$$

Secondly, (17) makes it possible, in conjunction with (14) and (32), to rewrite the expected prevalence among the tested individuals as

$$p = \sum_{s=0}^{S-1} \rho_{\pi s} p_{0s}, \tag{39}$$

where p_{0s} is the prevalence (32) among individuals with symptoms s, and

$$\rho_{\pi s} = \frac{\rho_s \pi_s}{\sum_{r=0}^{S-1} \rho_r \pi_r} = \frac{\rho_s \pi_s}{\pi} \tag{40}$$

is the expected proportion of tested individuals with symptoms s. The estimator of p in (8) and (18) can equivalently be expressed as

$$\hat{p} = \sum_{s=0}^{S-1} \hat{\rho}_{\pi s} \hat{p}_{0s} \tag{41}$$

whenever (17) holds, with

$$\hat{\rho}_{\pi s} = \rho_{Ts} = \frac{N_{Ts}}{N_T} = \rho_{Ts0} + \rho_{Ts1} \tag{42}$$

an estimate of $\rho_{\pi s}$. In view of (17), the requirement is made that $\hat{\pi}_{si} = \hat{\pi}_s$. Introducing

$$\hat{\rho}_s = \frac{N_{Ts} \hat{\pi}_s^{-1}}{\sum_{r=0}^{S-1} N_{Tr} \hat{\pi}_r^{-1}}, \tag{43}$$

the bias-corrected prevalence estimator in (21) simplifies to

$$\hat{p}_0 = \sum_{s=0}^{S-1} \hat{\rho}_s \hat{p}_{0s}. \tag{44}$$

The quality of $\hat{p}_0$ as an estimator of p_0 in (34), depends on how well $\hat{\rho}_s$ estimates ρ_s. In order to quantify this we introduce $\bar{\rho}_s$ as an asymptotic limit of $\hat{\rho}_s$, and

$$\bar{p}_0 = \sum_{s=0}^{S-1} \bar{\rho}_s p_{0s} \tag{45}$$

as the corresponding asymptotic limit of $\hat{p}_0$. It will further be helpful to introduce the S-dimensional vectors

$$\begin{aligned} \boldsymbol{p}_0 &= (p_{00}, \ldots, p_{0,S-1}), \\ \hat{\boldsymbol{p}}_0 &= (\hat{p}_{00}, \ldots, \hat{p}_{0,S-1}), \\ \boldsymbol{\rho}_\pi &= (\rho_{\pi 0}, \ldots, \rho_{\pi,S-1}), \\ \hat{\boldsymbol{\rho}}_\pi &= (\hat{\rho}_{\pi 0}, \ldots, \hat{\rho}_{\pi,S-1}), \\ \bar{\boldsymbol{\rho}} &= (\bar{\rho}_1, \ldots, \bar{\rho}_{S-1}), \\ \hat{\boldsymbol{\rho}} &= (\hat{\rho}_1, \ldots, \hat{\rho}_{S-1}) \end{aligned} \tag{46}$$

of (estimated) prevalences and expected relative sizes (before and after bias correction) of all symptom classes.

We will first establish a lemma on the joint asymptotic normality of the two vectors $\hat{\boldsymbol{p}}_0$ and $\hat{\boldsymbol{\rho}}_\pi$ in (46).

Lemma 1: Suppose $N \to \infty$ in such a way that all $\rho_{si} = N_{si}/N$ are kept fixed, and that the second MAR condition (17) holds for fixed $\pi_0, \ldots, \pi_{S-1}$. The vectors of estimated prevalences $\hat{\boldsymbol{p}}_0$ and estimated relative symptom classes $\hat{\boldsymbol{\rho}}_\pi$, among the tested individuals, are then asymptotically normally distributed in the sense that

$$N^{1/2}(\hat{\boldsymbol{p}}_0 - \boldsymbol{p}_0, \hat{\boldsymbol{\rho}}_\pi - \boldsymbol{\rho}_\pi) \longrightarrow_{\mathcal{L}} N\left(0, \begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix}\right) \tag{47}$$

as $N \to \infty$, with $\longrightarrow_{\mathcal{L}}$ referring to weak convergence, whereas $A = (A_{rs})_{r,s=0}^{S-1}$ and $B = (B_{rs})_{r,s=0}^{S-1}$ are square matrices of order S with elements

$$A_{rs} = \begin{cases} (1-\pi_s) p_{0s}(1-p_{0s})/(\rho_s \pi_s), & r = s, \\ 0, & r \neq s, \end{cases} \tag{48}$$

and

$$\begin{aligned} \pi^4 B_{rs} = {} & \pi^2 \rho_s \pi_s (1-\pi_s) \mathbf{1}_{\{r=s\}} \\ & - \pi \rho_r \pi_r (1-\pi_r) \rho_s \pi_s \\ & - \pi \rho_r \pi_r \rho_s \pi_s (1-\pi_s) \\ & + \rho_r \pi_r \rho_s \pi_s \Sigma_\pi \end{aligned} \tag{49}$$

respectively, where $\mathbf{1}_A$ is the indicator function over the set A and

$$\Sigma_\pi = N \mathrm{Var}(\hat{\pi}) = \sum_{s=0}^{S-1} \rho_s \pi_s (1-\pi_s). \tag{50}$$

Remark 1: a) Lemma 1 states that the vector of estimated prevalences $\hat{\boldsymbol{p}}_0 = (N_{Ts1}/N_{Ts})_{s=0}^{S-1}$ is asymptotically independent of the vector of estimated relative symptom classes $\hat{\boldsymbol{\rho}}_\pi = (N_{Ts}/N_T)_{s=0}^{S-1}$. This is a consequence of the second MAR condition (17). Indeed, when the willingness to be tested only depends on symptoms, the number of tested individuals with different types of symptoms carry no information about the prevalences p_{0s} within the symptom classes $\mathcal{X}_s$. b) In order to simplify the proof of Lemma 1 we assumed that symptom sizes $N_s = N\rho_s$ are non-random. However, Lemma 1 also holds when $(N_s)_{s=0}^{S-1} \sim \mathrm{Mult}\left(N; (\rho_s)_{s=0}^{S-1}\right)$ has a multinomial distribution. This result is obtained by first repeating the proof of Lemma 1 conditionally on $(N_s)_{s=0}^{S-1}$, and then averaging over $(N_s)_{s=0}^{S-1}$.

The following theorem provides asymptotic properties of $\hat{p}_0$, $\hat{p}$, and $\hat{I}_T^+$:

Theorem 1: Suppose that the conditions of Lemma 1 hold. Assume additionally that the vector $\hat{\boldsymbol{\rho}}$ of estimated symptom class sizes is such that (47) extends to

$$\begin{aligned} & N^{1/2}(\hat{\boldsymbol{p}}_0 - \boldsymbol{p}_0, \hat{\boldsymbol{\rho}}_\pi - \boldsymbol{\rho}_\pi, \hat{\boldsymbol{\rho}} - \bar{\boldsymbol{\rho}}) \\ & \longrightarrow_{\mathcal{L}} N\left(0, \begin{pmatrix} A & 0 & 0 \\ 0 & B & D \\ 0 & D^T & C \end{pmatrix}\right) \end{aligned} \tag{51}$$

as $N \to \infty$, for some square matrices $C = (C_{rs})_{r,s=0}^{S-1}$, and $D = (D_{rs})_{r,s=0}^{S-1}$ of dimension S. Then $\hat{p}$, $\hat{p}_0$, $\hat{I}_T^+$ are asymptotically normal estimators of p, $\bar{p}_0$ and $\bar{I}_T^+ = \log(p/\bar{p}_0) = I_T^+ - \log(\bar{p}_0/p_0)$ as $N \to \infty$, in the sense that

$$N^{1/2}(\hat{p} - p) \longrightarrow_{\mathcal{L}} N(0, V_1 + V_2), \tag{52}$$

$$N^{1/2}(\hat{p}_0 - \bar{p}_0) \longrightarrow_{\mathcal{L}} N(0, V_3 + V_4), \tag{53}$$

and

$$N^{1/2}\left(\hat{I}_T^+ - \bar{I}_T^+\right)$$

$$\longrightarrow_{\mathcal{L}} N\left(0, \frac{V_1+V_2}{p^2}+\frac{V_3+V_4}{\bar{p}_0^2}-\frac{2(V_5+V_6)}{p\bar{p}_0}\right), \quad (54)$$

with

$$\begin{aligned}
V_1 &= \sum_s \rho_{\pi s}^2 A_{ss} \\
&= \sum_s \rho_{\pi s}^2 (1-\pi_s) p_{0s}(1-p_{0s})/(\rho_s \pi_s) \\
&= \sum_s \rho_s \pi_s (1-\pi_s) p_{0s}(1-p_{0s})/\pi^2, \\
V_2 &= \sum_{r,s} p_{0r} p_{0s} B_{rs} \\
&= \sum_s \rho_s \pi_s (1-\pi_s)(p_{0s}-p)^2/\pi^2, \\
V_3 &= \sum_s \bar{\rho}_s^2 A_{ss} \\
&= \sum_s \bar{\rho}_s^2 (1-\pi_s) p_{0s}(1-p_{0s})/(\rho_s \pi_s), \\
V_4 &= \sum_{r,s} p_{0r} p_{0s} C_{rs}, \\
V_5 &= \sum_s \rho_{\pi s} \bar{\rho}_s A_{ss} \\
&= \sum_s \rho_{\pi s} \bar{\rho}_s (1-\pi_s) p_{0s}(1-p_{0s})/(\rho_s \pi_s), \\
V_6 &= \sum_{r,s} p_{0r} p_{0s} D_{rs}. \quad (55)
\end{aligned}$$

Remark 2: It is assumed in Theorem 1 that the vector $\hat{\boldsymbol{\rho}}$ of bias corrected estimates of symptom class sizes is asymptotically independent of the vector $\hat{\boldsymbol{p}}_0$ of prevalence estimates within the symptom classes. This relies on the second MAR condition (17) and a tacit assumption that $\hat{\boldsymbol{\rho}}$ is a known function of $\{N_{Ts}\}_{s=0}^{S-1}$ (cf. Remark 1).

Corollary 1 (Standard Errors and Confidence Intervals): The asymptotic variances $\sigma_p^2 = (V_1+V_2)/N$, $\sigma_{p_0}^2 = (V_3+V_4)/N$, and $\sigma_{I_T^+}^2$ in formulas (52)-(54) are functions of p_{0s}, $\bar{p}_0$, p, π_s, ρ_s, and $\bar{\rho}_s$. If estimates $\hat{p}_{0s}$, $\hat{p}_0$, $\hat{p}$, $\hat{\pi}_s$, $\hat{\rho}_s$, and $\hat{\rho}_s$ of these quantities are plugged into the asymptotic variances in (52)-(54), it is possible to obtain standard errors $\hat{\sigma}_p$, $\hat{\sigma}_{p_0}$, and $\hat{\sigma}_{I_T^+}$ of $\hat{p}$, $\hat{p}_0$, and $\hat{I}_T^+$, respectively. The corresponding confidence interval of I_T^+, with asymptotic coverage probability $1-\alpha$ when $\bar{I}_T^+ = I_T^+$, is

$$\mathrm{CI}_{I_T^+} = \left(\hat{I}_T^+ - \lambda_{\alpha/2}\hat{\sigma}_{I_T^+}, \hat{I}_T^+ + \lambda_{\alpha/2}\hat{\sigma}_{I_T^+}\right),$$

where $\lambda_{\alpha/2}$ is the $(1-\alpha/2)$-quantile of a standard normal distribution. As for prevalences, the delta method is first used to determine confidence intervals for logit transformed versions $g(p) = \mathrm{logit}(p) = \log[p/(1-p)]$ and $g(p_0) = \mathrm{logit}(p_0)$ of the prevalence parameters [33], [34]. Denoting

$$\begin{aligned}
L_{\hat{p}}^{\alpha} &= \lambda_{\alpha/2}\hat{\sigma}_p/[\hat{p}(1-\hat{p})], \\
L_{\hat{p}_0}^{\alpha} &= \lambda_{\alpha/2}\hat{\sigma}_{p_0}/[\hat{p}_0(1-\hat{p}_0)],
\end{aligned}$$

a logistic back-transformation $g^{-1}(z) = \mathrm{logit}^{-1}(z) = \exp(z)/(1+\exp(z))$ yields confidence intervals

$$\mathrm{CI}_p = \left(g^{-1}\left(g(\hat{p}) - L_{\hat{p}}^{\alpha}\right), g^{-1}\left(g(\hat{p}) + L_{\hat{p}}^{\alpha}\right)\right), \quad (56)$$

and

$$\mathrm{CI}_{p_0} = \left(g^{-1}\left(g(\hat{p}_0) - L_{\hat{p}_0}^{\alpha}\right), g^{-1}\left(g(\hat{p}_0) + L_{\hat{p}_0}^{\alpha}\right)\right), \quad (57)$$

of p and p_0, respectively. The asymptotic coverage probability is $1-\alpha$ for CI_p, and for CI_{p_0} as well whenever $\bar{p}_0 = p_0$.

Corollary 2 (MAR): Since ρ_s is known under MAR sampling, it follows that $\hat{\rho}_s = \bar{\rho}_s = \rho_s$ and $\bar{p}_0 = p_0$. Then $C_{rs} = D_{rs} = 0$ for all $0 \le r, s \le S-1$, so that the two variance components $V_4 = V_6 = 0$ vanish. In particular, $\hat{I}_T^+$ is an asymptotically unbiased estimator of I_T^+, and (54) simplifies to

$$N^{1/2}(\hat{I}_T^+ - I_T) \longrightarrow_{\mathcal{L}} N\left(0, \frac{V_1+V_2}{p^2}+\frac{V_3}{p_0^2}-\frac{2V_5}{pp_0}\right)$$

as $N \to \infty$.

Corollary 3 (A Conditional Version of Active Information): Suppose that the interest is in active information due to sampling bias *conditionally* on the number $N_{T0}, \ldots, N_{T,S-1}$ of individuals with different symptoms that are tested. The corresponding prevalence and active information are

$$\bar{p}_N = E\left(\hat{p} | \{N_{Ts}\}_{s=0}^{S-1}\right) = \sum_{s=0}^{S-1} \hat{\rho}_{\pi s} p_{0s}. \quad (58)$$

and

$$\bar{I}_{TN}^+ = \log \frac{\bar{p}_N}{p_0} \quad (59)$$

respectively. Using the same type of argument as in the proof of Theorem 1, it can be shown that

$$N^{1/2}(\hat{p} - \bar{p}_N) \longrightarrow_{\mathcal{L}} N(0, V_1) \quad (60)$$

and

$$\begin{aligned}
&N^{1/2}\left(\hat{I}_T^+ - \left(\bar{I}_{TN}^+ - \log\frac{\bar{p}_0}{p_0}\right)\right) \\
&\longrightarrow_{\mathcal{L}} N\left(0, \frac{V_1}{p^2}+\frac{V_3+V_4}{\bar{p}_0^2}-\frac{2V_5}{p\bar{p}_0}\right) \quad (61)
\end{aligned}$$

as $N \to \infty$.

Remark 3: The V_2 and V_6 terms are missing in (60) and (61), compared to (52) and (54). These terms correspond to the fact that the actual proportions $\hat{\rho}_{\pi s}$ of tested individuals with different symptoms deviate slightly from the corresponding expected proportions $\rho_{\pi s}$. Because of these missing variance terms of (60) and (61), the standard errors of $\hat{p}$ and $\hat{I}_T^+$ are smaller when a conditional rather than an unconditional approach is used, and the confidence intervals for $\bar{p}_N$ and $\bar{I}_{TN}^+$ are shorter than those for p and I_T^+.

Example 5 (MNAR With Unknown Strata Sizes, Contd): Let us generalize Example 4 and consider an MNAR sampling scheme where the sizes ρ_s of symptom classes might not be known, although the other MAR condition (17) holds. It is assumed that lower and upper bounds $0 \le a_{sN} \le \rho_s \le b_{sN} \le 1$ of ρ_s are known. The maximum entropy approach of Example 4 will be generalized. To this end, assume that the vector $\boldsymbol{\rho} = (\rho_0, \ldots, \rho_{S-1})$ is a random variable supported on

$$\mathcal{R} = \left\{\boldsymbol{\rho};\, a_{sN} \le \rho_s \le b_{sN};\, \sum_{s=0}^{S-1} \rho_s = 1\right\}, \quad (62)$$

a subset of the $(S-1)$-simplex of dimension $0 \le d \le S-1$, where $d = \max(|\{s; a_{sN} < b_{sN}\}| - 1, 0)$. By the maximum entropy principle, $\boldsymbol{\rho}$ has a uniform density f_ρ on $\mathcal{R}$, which degenerates to a point mass at $\mathcal{R}$ when $d = 0$. This gives rise to estimates

$$\hat{\pi}_s = \frac{N_{Ts}}{NE(\rho_s)} \tag{63}$$

of the sampling probabilities π_s. Inserting (63) into (43),

$$\hat{\rho}_s = \frac{E(\rho_s)}{\sum_{r=0}^{S-1} E(\rho_s)} = E(\rho_s) = \int_{\mathcal{R}} r_s f_\rho(r) dr, \tag{64}$$

with $r = (r_0, \ldots, r_{S-1})$. Since $\mathcal{R}$ in (62) is convex, it follows that $\hat{\rho}_s \in \mathcal{R}$.

The MAR sampling scheme of Example 2 corresponds to the special case $a_{sN} = b_{sN} = E(\rho_s) = \rho_s = \hat{\rho}_s$ and $d = 0$. For the MNAR COVID-19 model of Example 4, we recall that $S = 2$. Use (36) to rewrite (43) according to

$$\hat{\rho}_0 = 1 - N_{T1} / \left(N \hat{\pi}_1^{\text{MaxEnt}}\right), \quad \hat{\rho}_1 = N_{T1} / \left(N \hat{\pi}_1^{\text{MaxEnt}}\right). \tag{65}$$

With an apriori assumption $\hat{\pi}_0^{\text{MaxEnt}} \le \hat{\pi}_1^{\text{MaxEnt}}$, equation (65) implies that $\mathcal{R}$ has the maximal dimension $d = S - 1 = 1$ as long as $N_T < N$. Indeed, since

$$\begin{aligned} a_{0N} &= 1 - N_{T1}/N_T, \quad & b_{0N} &= 1 - N_{T1}/N, \\ a_{1N} &= N_{T1}/N, & b_{1N} &= N_{T1}/N_T, \end{aligned}$$

we then have that $a_{sN} < b_{sN}$ for $s = 0, 1$. It then follows from (64) that

$$\begin{aligned} \hat{\rho}_0 &= 1 - N_{T1}/(2N_T) \cdot (N_T/N + 1), \\ \hat{\rho}_1 &= N_{T1}/(2N_T) \cdot (N_T/N + 1). \end{aligned} \tag{66}$$

Insertion of (31) and (66) into (44) finally leads to (37). It is shown in the Appendix that Theorem 1 holds with

$$\begin{aligned} C_{11} &= C_{00} = -C_{01} = -C_{10} \\ &= \left[(1+\pi)^2 B_{11} + \rho_{\pi 1}^2 \Sigma_\pi \right. \\ &\quad \left. + 2(1+\pi)\rho_{\pi 1} \Sigma_{\rho\pi 1}\right]/4, \end{aligned} \tag{67}$$

$$\begin{aligned} D_{11} &= D_{00} = -D_{01} = -D_{10} \\ &= \left[(1+\pi)B_{11} + \rho_{\pi 1}\Sigma_{\pi\rho 1}\right]/2, \end{aligned} \tag{68}$$

and

$$\begin{aligned} \Sigma_{\pi\rho s} &= \lim_{N\to\infty} N\text{Cov}\left(\hat{\pi}^{\text{MaxEnt}}, \hat{\rho}_{\pi s}\right) \\ &= \rho_s \pi_s (1 - \pi_s)/\pi - \rho_s \pi_s \Sigma_\pi / \pi^2. \end{aligned} \tag{69}$$

*Example 6 (*MNAR *With Unknown Strata Sizes, Contd):* In the previous example it was assumed that the set $\mathcal{R}$ of symptom size classes in (62) was convex. We will now consider models where

$$\mathcal{R} = \{\boldsymbol{\rho}(\theta) = (\rho_0(\theta), \ldots, \rho_{S-1}(\theta)); \theta \in \Theta\} \tag{70}$$

is not necessarily convex, but the d-dimensional parameter set Θ is. The maximum entropy principle is used to assign a uniform prior distribution to θ on Θ. This gives rise to estimates

$$\hat{\boldsymbol{\rho}} = \boldsymbol{\rho}(E(\theta)) = (\hat{\rho}_0, \ldots, \hat{\rho}_{S-1}).$$

The parameter set Θ is determined by some apriori assumptions as well as constraints imposed by data. Since Θ is convex it follows that $\hat{\theta} = E(\theta) \in \Theta$ and hence $\hat{\boldsymbol{\rho}} \in \mathcal{R}$. In order to illustrate this approach, assume that the testing probability of the symptom classes satisfy

$$\pi_s(\theta) = \frac{\pi_0}{1 - \theta s} \tag{71}$$

for $s = 0, \ldots, S-1$, with θ a $d = 1$-dimensional parameter that satisfies $0 \le \theta \le \bar{\theta}$ for some appropriately chosen upper bound $\bar{\theta}$. The lower zero bound on θ ensures that testing probability π_s is a monotone and non-decreasing function of symptom strength s. Equation (71) gives rise to bias-corrected symptom class sizes

$$\rho_s(\theta) = \frac{N_{Ts}(1 - \theta s)}{\sum_{r=0}^{S-1} N_{Tr}(1 - \theta r)}$$

when θ is the true parameter. The corresponding estimated symptom class sizes are

$$\hat{\rho}_s = \rho_s\left(\hat{\theta}\right) = \frac{N_{Ts}\left(1 - \hat{\theta} s\right)}{\sum_{r=0}^{S-1} N_{Tr}\left(1 - \hat{\theta} r\right)}, \tag{72}$$

with $\hat{\theta} = E(\theta) = \bar{\theta}/2$. Note that (72) is consistent with (43), when the estimated sampling probabilities are chosen as $\hat{\pi}_s = \pi_0/(1 - \hat{\theta} s)$. The upper bound $\bar{\theta}$ on θ is determined by the requirement $\hat{\rho}_s \ge N_{Ts}/N$ for $s = 0, \ldots, S-1$. It can be seen that this leads to

$$\bar{\theta} = \min\left\{\frac{N - N_T}{sN - \sum_r r N_{Tr}}; s \text{ such that } sN > \sum_r N_{Tr}\right\}.$$

VI. Testing Errors

Following [30], we now extend the model and allow for testing errors, when the disease status at a medical lab or the behavioral characteristics at a social platform are registered imperfectly. We assume that the probability is α_s of falsely classifying an individual of symptom group s with feature value 0 ($i = 0$) as having feature value 1 ($i = 1$), whereas the probability of falsely classifying an individual with symptoms s and feature value 1, as having feature value 0, is β_s. Let $\check{\alpha}_s$ and $\check{\beta}_s$ be the corresponding fractions of wrongly classified subjects among the N_{Ts} individuals with symptoms s that were tested. Whereas a fraction N_{Ts1}/N_{Ts} of tested individuals with symptoms s have the feature 1, the corresponding reported fraction is

$$\begin{aligned} \check{p}_{0s} &= \check{\alpha}_s(1 - N_{Ts1}/N_{Ts}) + \left(1 - \check{\beta}_s\right) N_{Ts1}/N_{Ts} \\ &= \check{\alpha}_s + \left(1 - \check{\alpha}_s - \check{\beta}_s\right) N_{Ts1}/N_{Ts}. \end{aligned} \tag{73}$$

Although N_{Ts} is known, the number N_{Ts1} of tested individuals with symptoms s that have the feature ($i = 1$), is unknown. Consequently N_{Ts1}/N_{Ts} is unknown as well, but on the other hand it has the correct expected value $E(N_{Ts1}/N_{Ts}) = p_{0s}$. The reported fraction $\check{p}_{0s}$ of diseased individuals with symptoms s, on the other hand, is known and therefore an estimator of p_{0s} that is biased due to the presence of testing errors. Notice also that the testing error fractions $\check{\alpha}_s$ and $\check{\beta}_s$ are

random variables that approximate the expected testing error rates α_s and β_s. But since $\check{\alpha}_s$ and $\check{\beta}_s$ are unknown they are not viewed as estimators of α_s and β_s.

In order to construct a bias adjusted estimator $\hat{p}_{0s}$ of p_{0s}, suppose $\hat{\alpha}_s$ and $\hat{\beta}_s$ are estimators of α_s and β_s that make use of other data. Formula (73) suggests

$$\hat{p}_{0s} = \frac{\check{p}_{s0} - \hat{\alpha}_s}{1 - \hat{\alpha}_s - \hat{\beta}_s} = \frac{\check{\alpha}_s - \hat{\alpha}_s + \left(1 - \check{\alpha}_s - \check{\beta}_s\right) \cdot N_{Ts1}/N_{Ts}}{1 - \hat{\alpha}_s - \hat{\beta}_s}. \tag{74}$$

We can also view $\hat{\alpha}_s$ and $\hat{\beta}_s$ as predictors of the random quantities $\tilde{\alpha}_s$ and $\tilde{\beta}_s$. In the ideal case when $\hat{\alpha}_s = \tilde{\alpha}_s$ and $\hat{\beta}_s = \tilde{\beta}_s$, it would be possible to eliminate the effect of testing errors, so that $\hat{p}_{0s} = N_{Ts1}/N_{Ts}$.

The estimators $\hat{p}$ and $\hat{p}_0$ of the prevalences p and p_0, before and after correction for testing bias, are defined as in (41) and (44), but with the symptom specific prevalence estimates $\hat{p}_{0s}$ given by (74) instead of (31). In order to study the asymptotic properties of $\hat{p}$, $\hat{p}_0$ and $\hat{I}_T^+$, we need to know the asymptotic behaviour of the actual testing error fractions $\check{\alpha}_s$ and $\check{\beta}_s$, as well as the asymptotics of the estimated testing error fractions $\hat{\alpha}_s$ and $\hat{\beta}_s$. For the actual testing error rates $\check{\alpha}_s$ and $\check{\beta}_s$ we assume that type I and II testing errors occur, independently between individuals, with probabilities α_s and β_s. From this it follows that $\check{\alpha}_s$ and $\check{\beta}_s$ are binomial proportions, whose asymptotics is summarized in the following lemma:

Lemma 2: The fractions $\check{\alpha}_s$ and $\check{\beta}_s$ of tested individuals with errors of type I and II, for symptom classes $s = 0, \ldots, S-1$ are all independent random variables. Moreover, they satisfy

$$N^{1/2}(\check{\alpha}_s - \alpha_s) \longrightarrow_{\mathcal{L}} N(0, \Sigma_{\alpha\alpha s}) \tag{75}$$

and

$$N^{1/2}(\check{\beta}_s - \beta_s) \longrightarrow_{\mathcal{L}} N(0, \Sigma_{\beta\beta s}) \tag{76}$$

respectively as $N \to \infty$, with

$$\Sigma_{\alpha\alpha s} = \alpha_s(1-\alpha_s)/[\rho_{\pi s}(1-p_{0s})],$$
$$\Sigma_{\beta\beta s} = \beta_s(1-\beta_s)/[\rho_{\pi s}p_{0s}].$$

Regarding the vectors $\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_0, \ldots, \hat{\alpha}_{S-1})$ and $\hat{\boldsymbol{\beta}} = \left(\hat{\beta}_0, \ldots, \hat{\beta}_{S-1}\right)$ of estimated type I and II errors, their asymptotics depend on the way these estimators are constructed. We will *assume* that they are independent of all $\check{\alpha}_s$ and $\check{\beta}_s$ and converge to asymptotic limits $\bar{\boldsymbol{\alpha}} = (\bar{\alpha}_0, \ldots, \bar{\alpha}_{S-1})$ and $\bar{\boldsymbol{\beta}} = \left(\bar{\beta}_0, \ldots, \bar{\beta}_{S-1}\right)$ respectively. In more detail, we assume asymptotic normality

$$N^{1/2}\left(\hat{\boldsymbol{\alpha}} - \bar{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}\right) \longrightarrow_{\mathcal{L}} N\left(0, \begin{pmatrix} \Omega_{\alpha\alpha} & \Omega_{\alpha\beta} \\ \Omega_{\alpha\beta}^T & \Omega_{\beta\beta} \end{pmatrix}\right) \tag{77}$$

as $N \to \infty$, with an asymptotic covariance matrix of order $2S$ that involves the matrices $\Omega_{\alpha\alpha} = (\Omega_{\alpha\alpha rs})_{r,s=0}^{S-1}$, $\Omega_{\beta\beta} = (\Omega_{\beta\beta rs})_{r,s=0}^{S-1}$, and $\Omega_{\alpha\beta} = (\Omega_{\alpha\beta rs})_{r,s=0}^{S-1}$ of order S.

Formulas (75)-(77) suggest that the estimated prevalence (74) within symptom classes $s = 0, \ldots, S-1$ converge to the elements of $\bar{\boldsymbol{p}}_0 = (\bar{p}_0, \ldots, \bar{p}_{S-1})$ as $N \to \infty$, where

$$\bar{p}_{0s} = \frac{\alpha_s - \bar{\alpha}_s + (1 - \alpha_s - \beta_s)p_{0s}}{1 - \bar{\alpha}_s - \bar{\beta}_s} \tag{78}$$

Note in particular that $\bar{p}_{0s} = p_{0s}$ when the type I and II error rates are estimated consistently, that is, when $\bar{\alpha}_s = \alpha_s$ and $\bar{\beta}_s = \beta_s$. With these preliminaries we are ready to formulate the following extension of Theorem 1:

Theorem 2: Suppose the conditions of Theorem 1 hold, and additionally that the estimated prevalances $\hat{\boldsymbol{p}}_0 = (\hat{p}_{00}, \ldots, \hat{p}_{0,S-1})$ of all symptom classes involve correction for testing errors, as defined in (73)-(74), and with testing error rate estimates $\hat{\alpha}_s$ and $\hat{\beta}_s$ being asymptotically normal, according to (77), and independent of testing error rates $\check{\alpha}_s$ and $\check{\beta}_s$ of the sample. Then

$$N^{1/2}\left(\hat{\boldsymbol{p}}_0 - \bar{\boldsymbol{p}}_0, \hat{\boldsymbol{\rho}}_\pi - \boldsymbol{\rho}_\pi, \hat{\boldsymbol{\rho}} - \bar{\boldsymbol{\rho}}\right) \longrightarrow_{\mathcal{L}} N\left(0, \begin{pmatrix} \bar{A} & 0 & 0 \\ 0 & B & D \\ 0 & D^T & C \end{pmatrix}\right) \tag{79}$$

as $N \to \infty$, with $B = (B_{rs})_{r,s=0}^{S-1}$, $C = (C_{rs})_{r,s=0}^{S-1}$, and $D = (D_{rs})_{r,s=0}^{S-1}$ the same square matrices of dimension S as in Theorem 1, whereas $\bar{A} = (\bar{A}_{rs})_{r,s=0}^{S-1}$ is a matrix whose elements are defined in the appendix. Moreover, $\hat{p}$, $\hat{p}_0$ and $\hat{I}_T^+ = \log(\hat{p}/\hat{p}_0)$ are asymptotically normal estimators of $\bar{p} = \sum_s \rho_{\pi s}\bar{p}_{0s}$, $\bar{p}_0 = \sum_s \bar{\rho}_s\bar{p}_{0s}$ and $\bar{I}_T^+ = \log(\bar{p}/\bar{p}_0)$ respectively, in the sense that

$$N^{1/2}(\hat{p} - \bar{p}) \longrightarrow_{\mathcal{L}} N(0, V_1 + V_2),$$
$$N^{1/2}(\hat{p}_0 - \bar{p}_0) \longrightarrow_{\mathcal{L}} N(0, V_3 + V_4),$$
$$N^{1/2}\left(\hat{I}_T^+ - \bar{I}_T^+\right) \longrightarrow_{\mathcal{L}} N(0, V) \tag{80}$$

as $N \to \infty$, with

$$V = \frac{V_1 + V_2}{p^2} + \frac{V_3 + V_4}{\bar{p}_0^2} - \frac{2(V_5 + V_6)}{p\bar{p}_0}$$

and

$$V_1 = \sum_{r,s} \rho_{\pi r}\rho_{\pi s}\bar{A}_{rs}, \qquad V_2 = \sum_{r,s} \bar{p}_{0r}\bar{p}_{0s}B_{rs},$$
$$V_3 = \sum_{r,s} \bar{\rho}_r\bar{\rho}_s\bar{A}_{rs}, \qquad V_4 = \sum_{r,s} \bar{p}_{0r}\bar{p}_{0s}C_{rs},$$
$$V_5 = \sum_{r,s} \rho_{\pi s}\bar{\rho}_s\bar{A}_{rs}, \qquad V_6 = \sum_{r,s} \bar{p}_{0r}\bar{p}_{0s}D_{rs}. \tag{81}$$

Remark 4: The asymptotic bias of the estimator $\hat{I}_T^+$ of active information I_T^+, in Theorem 2, is

$$\bar{I}_T^+ - I_T^+ = \log\left(\frac{\bar{p}}{p}\right) - \log\left(\frac{\bar{p}_0}{p_0}\right). \tag{82}$$

The first term $\log(\bar{p}/p)$ on the right hand side of (82) is due to error in testing, whereas the second term $\log(\bar{p}_0/p_0)$ is due to testing bias as well as errors in testing. Only the second bias term is present when there are no errors in testing (cf. Theorem 1).

Remark 5: Suppose testing error rates are estimated independently between symptom classes, so that $\Omega_{\alpha\alpha}$, $\Omega_{\beta\beta}$, and

$\Omega_{\alpha\beta}$ in (77) are diagonal matrices. It follows from the proof of Theorem 2 that $\bar{A}$ is diagonal under this assumption.

Remark 6: Theorem 2 differs from Theorem 1 in that $\bar{A}_{rs}$, $\bar{p}_{0s}$, and $\bar{p}$ replace A_{rs}, p_{0s}, and p respectively. In the special case of no testing errors ($\alpha_s = \beta_s = 0$) and perfect estimation of testing errors ($\bar{\alpha}_s = \alpha_s$, $\bar{\beta}_s = \beta_s$, and $\Omega_{\alpha\alpha} = \Omega_{\beta\beta} = \Omega_{\alpha\beta} = 0$), we have that $\bar{A}_{rs} = A_{rs}$, $\bar{p}_{0s} = p_{0s}$, and $\bar{p} = p$. Theorem 2 then reduces to Theorem 1, and the variance components $V_1, \ldots, V_6$ of both theorems are then the same.

VII. Numerical Illustrations

A. Examples 1-3

This section illustrates with simulations the methodology under the framework of Examples 1-3. In these simulations, N denotes known population size which is increased from 1000 to 1000000. The true population prevalence of feature value 1 is set at $p_0 = 0.20$. Only two levels of symptoms will be considered, $s \in \{0, 1\}$. The proportion ρ_1 of people with symptoms in the population is 0.20 and without symptoms it is $\rho_0 = 0.80$. The proportion of positive cases ($i = 1$) with symptoms $\rho_{11} = 0.15$, and the proportion of positive cases without symptoms $\rho_{01} = 0.05$. Notice that $\rho_{01} + \rho_{11} = p_0$.

For Examples 1-2, the testing group within each symptom class is assumed to be independent of the feature status ($\pi_{si} = \pi_s$), in accordance with (17). Let π_1 be the probability of testing the symptomatic group, and π_0 be the probability of testing the asymptomatic group.

In the case of MCAR (Example 1), the sampling probability π is set to 0.6. Thus, estimated sampling fractions $\hat{\pi}_{00}^{\text{MCAR}} = \hat{\pi}_{01}^{\text{MCAR}} = \hat{\pi}_{10}^{\text{MCAR}} = \hat{\pi}_{11}^{\text{MCAR}} = \hat{\pi}^{\text{MCAR}}$ are same. The overall prevalence rate $\hat{p}_0^{\text{MCAR}} = \hat{p}$ can be estimated by the positive rate (27) in the testing sample.

For the MAR scenario of Example 2, the probability of testing in the symptomatic group π_1 is set to 0.90, while the probability of testing in the asymptomatic group is $\pi_0 = 0.10$. The estimated sampling fractions are assumed to be constant by symptoms from (29). Thus, the estimated population prevalence $\hat{p}_0$ in (30) differs from the sum of $\hat{p}_{0s}$, and it is a weighted average of the positive test rate by symptom proportions in the population.

Finally an MNAR situation of Example 3 is considered. Unlike MAR, the simulations were repeated without assuming $\pi_{si} = \pi_s$. Here, $\pi_{00} = 0.20$, $\pi_{01} = 0.30$, $\pi_{10} = 0.70$, $\pi_{11} = 0.80$. Thus, using the weighted positive test rate $\hat{p}_0$ from equation (30), as for MAR, biased results, for which the bias will not vanish asymptotically, are expected.

Each experiment is repeated $M = 500$ times. Let $\hat{p}^{(m)}$ and $\hat{p}_0^{(m)}$ refer to the prevalence estimates of iteration $m \in \{1, \ldots, M\}$ before and after correction for testing bias. We will use tildes to denote Monte Carlo estimates, and put

$$\tilde{E}(\hat{p}) = \sum_{m=1}^{M} \hat{p}^{(m)}/M,$$
$$\tilde{E}(\hat{p}_0) = \sum_{m=1}^{M} \hat{p}_0^{(m)}/M.$$

TABLE I
Monte Carlo-Estimated Active Information Under MCAR

Population	1000	10000	100000	1000000
$\tilde{I}_T^+$	0.0022	0.00002	0.00003	0.00003
$\tilde{I}_C^+$	0	0	0	0
$\tilde{I}^+$	0.0022	0.00002	0.00003	0.00003

TABLE II
Monte Carlo-Estimated Active Information Under MAR

Population	1000	10000	100000	1000000
$\tilde{I}_T^+$	0.9873	0.9901	0.9903	0.9905
$\tilde{I}_C^+$	-0.9917	-0.9914	-0.9892	-0.9907
$\tilde{I}^+$	-0.0044	-0.0013	0.0011	-0.0002

The corresponding Monte Carlo estimates of active informations I_T^+, I_C^+ and I^+, are defined as

$$\tilde{I}_T^+ = \log\left[\tilde{E}(\hat{p})/p_0\right],$$
$$\tilde{I}_C^+ = \log\left[\tilde{E}(\hat{p}_0)/\tilde{E}(\hat{p})\right],$$
$$\tilde{I}^+ = \log\left[\tilde{E}(\hat{p}_0)/p_0\right] = \tilde{I}_T^+ + \tilde{I}_C^+. \tag{83}$$

These estimates should not be confused with $\hat{I}_T^{+(m)} = \log\left(\hat{p}^{(m)}/\hat{p}_0^{(m)}\right)$ and $\hat{I}_C^{+(m)} = -\hat{I}_T^{+(m)}$, which are computed for each simulation. The Law of Large Numbers implies that $\tilde{I}_T^+ \to \log[E(\hat{p})/p_0]$ and $\tilde{I}^+ \to \log[E(\hat{p}_0)/p_0]$ as the number of simulations $M \to \infty$, for each fixed N. Note also that $\tilde{I}_T^+ \to I_T^+$ and $\tilde{I}^+ \to \log(\bar{p}_0/p_0)$ as $N \to \infty$, for each fixed M, with $\bar{p}_0$ the asymptotic limit of $\hat{p}_0$. In particular, the closer to zero $\tilde{I}^+$ is, the more successful the bias correction of the prevalence estimate is.

Table I shows the active information of MCAR. As described above, the probabilities were averaged over the $M = 500$ realizations before calculating the active information values in (83). The estimated active information of the correction, $\tilde{I}_C^+$, is 0 because $\hat{p}_0^{(m)} = \hat{p}^{(m)}$ in MCAR. Thus, the active information of the bias-adjusted prevalence estimate for MCAR, $\tilde{I}^+$, is obtained from $\tilde{I}_T^+$. Notice that $\tilde{I}_T^+ = \tilde{I}^+$ converges to 0 with increasing N. This is to be expected from Remark 2, since $\bar{p}_0 = p_0$ for MCAR schemes.

Next, active information values under the MAR simulation were obtained, as shown in Table II. The active information of the bias-adjusted prevalence estimate in MAR is seen to increase as population increases, removing asymptotically the effect of a small overcorrection, with $\tilde{I}^+$ converging to 0 as N gets larger. Again, this is to be expected from Remark 2, since $\bar{p}_0 = p_0$ for any MAR scheme.

For MNAR, the active information $\tilde{I}^+$ of the bias-adjusted prevalence estimate for this simulation is displayed in Table III, showing that the strategy partially corrects the sampling bias. However, this bias correction does not improve with increasing N, since $\tilde{I}^+$ does not converge to 0. This is to be expected, since $\bar{p}_0 \neq p_0$ for a MNAR scheme.

TABLE III
MONTE CARLO-ESTIMATED ACTIVE INFORMATION UNDER MNAR

Population	1000	10000	100000	1000000
$\tilde{I}_T^+$	0.994	0.990	0.990	0.990
$\tilde{I}_C^+$	-0.398	-0.396	-0.396	-0.396
$\tilde{I}^+$	0.596	0.594	0.594	0.594

TABLE IV
EMPIRICAL RMSE FOR THREE SAMPLING MODEL SIMULATIONS

N	Empirical RMSE (SD)		
	MCAR	MAR	MNAR
1000	0.0058 (0.0042)	0.0218 (0.0129)	0.164 (0.006)
10000	0.0020 (0.0015)	0.0072 (0.0045)	0.162 (0.002)
100000	0.0006 (0.0005)	0.0023 (0.0014)	0.162 (0.001)
1000000	0.0002 (0.0001)	0.0007 (0.0004)	0.162 (0.001)

TABLE V
FRACTION OF CONFIDENCE INTERVALS $\mathrm{CI}_{p_0}^{(m)}$ FOR THE TWO MARSCENARIOS OF FIGURES 1 AND 2 THAT COVER THE TRUE PREVALENCE p_0 OUT OF $M = 500$ RUNS OF THE SIMULATION. THE NOMINAL COVERAGE IS $1 - \alpha = 0.95$

Population size N	ρ_1=0.1, p_0=0.05	ρ_1=0.2, p_0=0.15
1000	0.850	0.916
10000	0.898	0.928
100000	0.912	0.950
1000000	0.916	0.942

Empirical root mean squared errors

$$\mathrm{RMSE} = \sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(\hat{p}_0^{(m)} - p_0\right)^2}$$

for the bias-corrected population prevalence estimates under each scenario are reported in Table IV, together with their standard deviations in parentheses. Clearly, the empirical RMSEs drop to zero with increasing N under MAR and MCAR but not under MNAR where even for very large population sizes, the estimation of population prevalence cannot be improved. This is in line with the comments below Tables I-III, since $\mathrm{RMSE} \to |\bar{p}_0 - p_0|$ as $N \to \infty$.

B. Asymptotics

Section V develops the asymptotic limiting distribution of the bias-corrected population prevalence estimator $\hat{p}_0$ in (44). Two MAR-scenarios are explored here: i) small p_0, where the proportion of symptomatic individuals in the population ρ_1 is set to 0.1, the population prevalence p_0 is set to 0.05, and the proportion of positive cases with symptoms ρ_{11} equals 0.07; and ii) large p_0, where $\rho_1 = 0.2$, $p_0 = 0.15$, and $\rho_{11} = 0.1$.

Corollary 1 and Corollary 2 are used to estimate $\sigma_{p_0}^2 = V_3/N$. Figures 1 and 2 show 95% CIs for p_0 over $M = 500$ realizations of the simulations for increasing N, for each of the two MAR-scenarios, with the red dashed lines indicating the true value of p_0. Table V gives the empirical coverage probabilities for these scenarios.

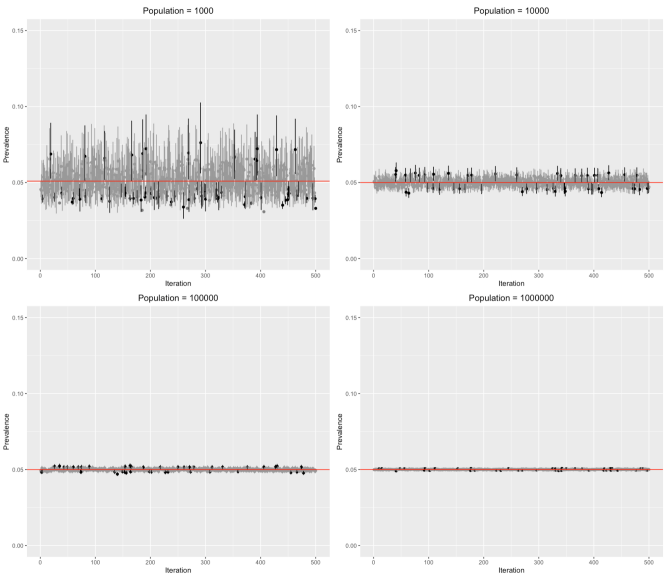

Fig. 1. Confidence interval plots of $\mathrm{CI}_{p_0}^{(m)}$ for $M = 500$ simulations and increasing population sizes under the MAR scenario 1, with $\rho_1 = 0.1$ and $p_0 = 0.05$.

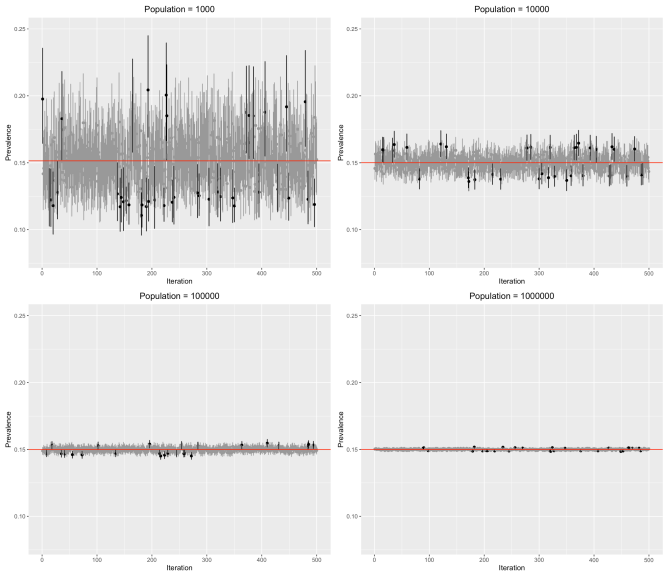

Fig. 2. Confidence interval plots of $\mathrm{CI}_{p_0}^{(m)}$ for $M = 500$ simulations and increasing population sizes under the MAR scenario 2, with $\rho_1 = 0.2$ and $p_0 = 0.15$.

VIII. DISCUSSION

A. Summary

In this paper we study prevalence estimation of a binary-valued feature when individuals with various degrees of symptoms voluntarily decide to be "tested" for this feature. Active information is used to quantify the testing bias due to the fact that individuals with stronger symptoms, who are more likely to have value 1 of the feature, are also more likely to be tested. Incomplete testing is treated as a missing data problem, analogous to survey sampling, with non-tested individuals treated as non-respondents. Bias-corrected estimators are defined, and their asymptotic properties are derived, for a wide range of missingness mechanisms where data is

either missing at random (MAR) or data is missing not at random (MNAR). In particular, we focus on a non-standard type of MNAR scheme where i) response probabilities depend on covariates (symptoms) but not on outcomes (feature status), and ii) all non-respondents' covariates and outcome variables are unknown to the data analyst.

B. Interpretation of Information Theoretic Approach

We have assumed that individuals voluntarily are tested for feature 1, and based on this assumption we quantified the increased prevalence p of feature 1 among the tested individuals, compared to the total population prevalance p_0, as active information $I_T^+ = \log(p/p_0)$ due to testing. From the point of view of the data analyst, since testing is voluntary, $I_T^+ > 0$ represents external information that individuals bring in, in order to simplify the analyst's search for those that have feature 1.

Suppose however that it is possible for the data analyst to control the testing protocol, and that his goal is to find as many individuals as possible with feature 1, at the smallest possible cost. If the cost is proportional to the number $N_T \approx N\pi$ of tested individuals, this amounts to maximizing the number $N_{T1} \approx N_T \cdot p$ of individuals with feature 1, subject to an upper bound $\pi \leq \Pi$ on the fraction of tested individuals. Assume that testing probabilities only depend on symptoms (17), with π_s the probability that individuals with symptoms $s \in \{0, 1, \ldots, S-1\}$ are tested. We also postulate that the symptom strata fractions ρ_s as well the symptom specific prevalences p_{0s} are fixed. Since the total population prevalence p_0 only depends on $\{\rho_s\}_{s=0}^{S-1}$ and $\{p_{0s}\}_{s=0}^{S-1}$ (cf. (34)), it follows that p_0 is fixed as well. The prevalence p among the tested individuals, on the other hand, additionally depends on $\{\pi_s\}_{s=0}^{S-1}$. Thus the task of the data analyst is equivalent to finding a testing protocol $\{\pi_s\}_{s=0}^{S-1}$ that maximizes $I_T^+ = \log(p/p_0)$ subject to $\pi = \sum_s \rho_s \pi_s \leq \Pi$. If prevalences $p_{00} \leq \cdots \leq p_{0,S-1}$ increase with strength of symptoms, it can be seen that the optimal testing procedure is given by

$$\pi_s(\Pi) = \begin{cases} 0; & s = 0, 1, \ldots, r(\Pi) - 1, \\ \frac{\Pi - \bar{F}(r(\Pi)+1)}{\rho_{r(\Pi)}}; & s = r(\Pi), \\ 1; & s = r(\Pi) + 1, \ldots, S-1, \end{cases} \tag{84}$$

where $r(\Pi)$ is the solution of $\bar{F}(r+1) < \Pi \leq \bar{F}(r)$ for the survival function $\bar{F}(r) = \sum_{s=r}^{S-1} \rho_s$ of the symptom strata fraction distribution $\{\rho_s\}_{s=0}^{S-1}$. The optimal testing protocol (84) corresponds to an active information

$$I_{T,\max}^+(\Pi) = \log \frac{p_{\max}(\Pi)}{p_0} \tag{85}$$

due to testing, with $p_{\max}(\Pi)$ the testing prevalence obtained from protocol (84). This protocol is such that individuals are being tested in order of the strength of their symptoms, until a fraction Π of all individuals have been tested. It can be seen that the maximal active information of (85) is a decreasing function of Π, converging to $\log(p_{0,S-1}/p_0)$ and 0, as $\Pi \to 0$ and $\Pi \to 1$ respectively. That is, the smaller the resources Π of the data analyst, the more active information (85) the optimal sampling protocol (84) represents.

C. Extensions

The results of this paper can be extended in various ways. A first type of extension is to consider more elaborate covariates. For instance, if prevalence estimation in different localities (such as medical laboratories or social platforms) $l = 1, \ldots, L$, it is appropriate to divide the population into various subpopulations $\mathcal{X}_{lsi}$ with different combinations of localities l, symptoms s and feature status i. In this context, the prevalence of feature value 1 can be made to not only depend on symptoms but also on localities. That is,

$$p_{0ls} = \frac{|\mathcal{X}_{ls1}|}{|\mathcal{X}_{ls0}| + |\mathcal{X}_{ls1}|},$$

within each locality-symptom stratum (l, s), reflects that different localities have different medical testing procedures or different rules for social behaviour. This amounts to treating $z = (l, s)$ as a two-dimensional covariate, with feature status i as the binary outcome variable.

A second extension is to consider more general regression models, with data $x = (z, y) \in \mathcal{X}$ that consists of covariates z and a an arbitrary type of outcome variable y (such a continuous or a count variable). Suppose our goal is to estimate the expected response

$$\mu_0 = E_0(Y) = \sum_{x \in \mathcal{X}} y P_0(x) = E_0[\mu_0(Z)],$$

with P_0 the population distribution of $X = (Z, Y)$ and $\mu_0(z) = E_0(Y|Z = z)$. Let $\hat{\mu}$ be the sample mean of the response variable among the respondents. If the sampling probability $\pi_x = \pi_z$ is a function of covariates only, it follows that the asymptotic limit of of $\hat{\mu}$ is

$$\mu = E_\pi(Y) = \frac{\sum_{x \in \mathcal{X}} y \pi_z P_0(z, y)}{\sum_{x \in \mathcal{X}} \pi_z P_0(z, y)} = E_\pi[\mu_0(Z)],$$

rather than μ_0, where E_π refers to expectation with respect to the size-biased covariate distribution

$$P(z) = \frac{\pi_z P_0(z)}{\sum_{z'} \pi_{z'} P_0(z')},$$

whereas $P_0(z) = \sum_y P_0(z, y)$ is the population distribution of the covariate. A number of examples of size-biased distributions are provided in [35]. A corrected estimate $\hat{\mu}_0$ of μ_0 is asymptotically unbiased only for MAR schemes for which the covariate distribution P_0 is known. When P_0 is unknown, we have a sampling scheme with MNAR data. It is possible for such a scheme to define $\hat{\mu}$ from a Bayesian prior on P_0, similarly as in Example 4.

Hence, in a more general setting, in this paper we consider how to estimate the expected outcome $\mu_0 = E(Y)$ from a sample, when the sampling probabilities depend on some covariate Z, and the data analyst does not know the covariate nor the outcome Y among the non-respondents. We believe this framework has applications within a number of areas beyond epidemiology, such as behavioral science, quality control and market research.

ACKNOWLEDGMENT

The authors are grateful to Michael Sverchkov and Danny Pfefferman for providing valuable references on missing data. They also thank three anonymous reviewers for very valuable comments that considerably improved the quality of the article.

APPENDIX A
PROOFS

Proof of Lemma 1: In order to prove (47), we start by analysing the elements of $N^{1/2}(\hat{\boldsymbol{p}}_0 - \boldsymbol{p}_0)$. From (31)-(33) and properties of the hypergeometric distribution [36], it follows that the conditional mean and variance of the estimated prevalence of symptom class s equal

$$\begin{aligned} E(\hat{p}_{0s}|N_{Ts}) &= \frac{N_{Ts}p_{0s}}{N_{Ts}} = p_{0s}, \\ \mathrm{Var}(\hat{p}_{0s}|N_{Ts}) &= \frac{p_{0s}(1-p_{0s})(N_s - N_{Ts})}{(N_s-1)N_{Ts}} \end{aligned} \tag{86}$$

respectively. Averaging (86) over N_{Ts}, the corresponding unconditional expected value and normalized variance are

$$\begin{aligned} E(\hat{p}_{0s}) &= p_{0s}, \\ N\mathrm{Var}(\hat{p}_{0s}) &= p_{0s}(1-p_{0s})\frac{N}{N_s-1}E\left(\frac{N_s - N_{Ts}}{N_{Ts}}\right) \\ &\to p_{0s}(1-p_{0s})\frac{1-\pi_s}{\rho_s\pi_s} = A_{ss}, \end{aligned} \tag{87}$$

with the limit taken as $N \to \infty$, and with A_{ss} the diagonal element of symptom class s for the matrix A defined in (48). In the last step of (87) we made use of the fact that $N_s = N\rho_s$ and $N_{Ts}/(N\rho_s\pi_s) \to 1$ as $N \to \infty$. Equation (87), and asymptotic normality of the hypergeometric distribution when p_{0s} is kept fixed and $N_{Ts}/(N\rho_s\pi_s) \to 1$, implies

$$N^{1/2}(\hat{p}_{0s} - p_{0s}) \longrightarrow_{\mathcal{L}} N(0, A_{ss}) \tag{88}$$

as $N \to \infty$, for $s = 0, \ldots, S-1$. Next we consider several estimated prevalences jointly. It follows from (31), (86), and the fact that $\{\hat{p}_{0s}\}_{s=0}^{S-1}$ are conditionally independent given $\{N_{Ts}\}_{s=0}^{S-1}$, that

$$\begin{aligned} \mathrm{Cov}(\hat{p}_{0r}, \hat{p}_{0s}) &= E[\mathrm{Cov}(\hat{p}_{0r}, \hat{p}_{0s})|N_{Tr}, N_{Ts}] \\ &\quad + \mathrm{Cov}[E(\hat{p}_{0r}|N_{Tr}), E(\hat{p}_{0s}|N_{Ts})] \\ &= 0 + 0 = 0 \end{aligned} \tag{89}$$

whenever $r \neq s$. That is, the elements of $\hat{\boldsymbol{p}}_0 = (\hat{p}_{00}, \ldots, \hat{p}_{S-1,S-1})$ are uncorrelated. This implies that

$$N^{1/2}(\hat{\boldsymbol{p}}_0 - \boldsymbol{p}_0) \longrightarrow_{\mathcal{L}} N(0, A), \tag{90}$$

asymptotically as $N \to \infty$, with

$$A = \mathrm{diag}(A_{00}, \ldots, A_{S-1,S-1})$$

a diagonal matrix, with diagonal entries as in (88). Next, in order to verify that

$$N^{1/2}(\hat{\boldsymbol{\rho}}_\pi - \boldsymbol{\rho}_\pi) \longrightarrow_{\mathcal{L}} N(0, B) \tag{91}$$

we will analyze the elements of $\hat{\boldsymbol{\rho}}_\pi - \boldsymbol{\rho}_\pi$. The number of tested individuals with symptoms s is binomially distributed, $N_{Ts} \sim \mathrm{Bin}(N\rho_s, \pi_s)$, for $s = 0, \ldots, S-1$. Writing $N_{Ts}/N = \rho_s\pi_s + \varepsilon_s$, (42) yields that

$$\begin{aligned} \hat{\rho}_{\pi s} &= \frac{\rho_s\pi_s + \varepsilon_s}{\sum_{r=0}^{S-1}(\rho_r\pi_r + \varepsilon_r)} \\ &= \rho_{\pi s} + \frac{\varepsilon_s}{\sum_{r=0}^{S-1}\rho_r\pi_r} - \frac{\rho_s\pi_s\sum_{r=0}^{S-1}\varepsilon_r}{\left(\sum_{r=0}^{S-1}\rho_r\pi_r\right)^2} \\ &\quad + \frac{\sum_{r=0}^{S-1}\varepsilon_r}{\sum_{r=0}^{S-1}\rho_r\pi_r}\left[\rho_{\pi s} - \frac{\rho_s\pi_s + \varepsilon_s}{\sum_{r=0}^{S-1}(\rho_r\pi_r + \varepsilon_r)}\right], \end{aligned}$$

and the last term on the right-hand side is $o_p\left(N^{-1/2}\right)$. Invoking the definition of π in (38), and rearranging terms, it is possible to rewrite the last displayed equation as

$$\begin{aligned} \pi^2(\hat{\rho}_{\pi s} - \rho_{\pi s}) &= (\pi - \rho_s\pi_s)\varepsilon_s - \rho_s\pi_s\sum_{r\neq s}\varepsilon_r \\ &\quad + o_p\left(N^{-1/2}\right). \end{aligned} \tag{92}$$

The random variables $\varepsilon_0, \ldots, \varepsilon_{S-1}$ are independent with binomial variances

$$\mathrm{Var}(\varepsilon_s) = \rho_s\pi_s(1-\pi_s)/N. \tag{93}$$

It therefore follows from (92) that (91) holds, with asymptotic variance matrix $B = (B_{rs})$ having elements

$$\begin{aligned} \pi^4 B_{ss} &= (\pi - \rho_s\pi_s)^2\rho_s\pi_s(1-\pi_s) \\ &\quad + (\rho_s\pi_s)^2\sum_{r\neq s}\rho_r\pi_r(1-\pi_r) \end{aligned}$$

and

$$\begin{aligned} \pi^4 B_{rs} &= -\pi\rho_r\pi_r(1-\pi_r)\rho_s\pi_s - \pi\rho_r\pi_r\rho_s\pi_s(1-\pi_s) \\ &\quad + \rho_r\pi_r\rho_s\pi_s\sum_t\rho_t\pi_t(1-\pi_t) \end{aligned}$$

when $r \neq s$. Making use of the definition of Σ_π in (50), it is easily seen that the last two displayed equations simplify to (49). The proof of (47) is finalized by making use of (90) and (91), and noticing that $\hat{\boldsymbol{p}}_0$ is asymptotically independent of $\{\varepsilon_s\}_{s=0}^{S-1}$, and hence of $\hat{\boldsymbol{\rho}}$. In order to prove (50), it follows from the definition of $\{\varepsilon_s\}_{s=0}^{S-1}$ and (38) that

$$\hat{\pi} = \frac{\sum_s N_{Ts}}{N} = \frac{\sum_s N(\pi_s\pi_s + \varepsilon_s)}{N} = \pi + \sum_s \varepsilon_s. \tag{94}$$

Since $\{\varepsilon\}_{s=0}^{S-1}$ are independent with variances as in (93), formula (50) follows.

Proof of Theorem 1: We will start by proving (52). To this end, write

$$\begin{aligned} \hat{p} - p &= \sum_s \rho_{\pi s}(\hat{p}_{0s} - p_{0s}) + \sum_s (\hat{\rho}_{\pi s} - \rho_{\pi s})p_{0s} \\ &\quad + \sum_s (\hat{\rho}_{\pi s} - \rho_{\pi s})(\hat{p}_{0s} - p_{0s}). \end{aligned} \tag{95}$$

Each term on the right-hand side of (95) is now analyzed. As for the first term of (95) we invoke (88) and find that

$$N^{1/2}\sum_{s=0}^{S-1}\rho_{\pi s}(\hat{p}_{0s} - p_{0s}) \longrightarrow_{\mathcal{L}} N(0, V_1) \tag{96}$$

as $N \to \infty$. Since the left hand side of (96) is a linear combination of the elements of $\hat{\boldsymbol{p}}_0 - \boldsymbol{p}_0$, and since A is a diagonal matrix, it follows from (51) that the asymptotic variance in (96) equals $V_1 = \sum_{s=0}^{S-1} \rho_{\pi s}^2 A_{ss}$. This formula for V_1 is identical to the first line of (55). The expressions for V_1 in the next two lines of (55) follow from the definitions of A_{ss} and $\rho_{\pi s}$ in (48) and (40) respectively. As for the second term of (95),

$$N^{1/2} \sum_{s=0}^{S-1} (\hat{\rho}_{\pi s} - \rho_{\pi s}) p_{0s} \longrightarrow_{\mathcal{L}} N(0, V_2) \tag{97}$$

is also deduced from (51), with $V_2 = \sum_{r,s} p_{0r} p_{0s} B_{rs}$ as defined in (55). In order to simplify this expression of V_2 according to the subsequent line of (55), we make use of the definition of B_{rs} in (49) and find that

$$\begin{aligned}
\pi^4 V_2 &= \pi^2 \sum_s \rho_s \pi_s (1 - \pi_s) p_{0s} \\
&\quad - 2\pi \sum_{r,s} \rho_r \pi_r \rho_s \pi_s (1 - \pi_s) p_{0r} p_{0s} \\
&\quad + \Sigma_\pi \sum_{r,s} \rho_r \pi_r p_{0r} \rho_s \pi_s p_{0s} \\
&= \sum_s \rho_s \pi_s (1 - \pi_s) \left[\pi^2 - 2\pi p_{0s} \sum_r \rho_r \pi_r p_{0r} \right. \\
&\qquad \left. + \left(\sum_r \rho_r \pi_r p_{0r} \right)^2 \right] \\
&= \pi^2 \sum_s \rho_s \pi_s (1 - \pi_s) (p_{0s} - p)^2,
\end{aligned}$$

where in the last step we inserted the definition of p in (39). Because of (51), the first two terms on the right-hand side of (95) are asymptotically independent. Moreover, since $\hat{p}_{0s} - p_{0s} = O_p\left(N^{-1/2}\right)$ according to (88), and $\hat{\rho}_{\pi s} - \rho_{\pi s} = O_p\left(N^{-1/2}\right)$ according to (91), the last term on the right hand side of (95) is $o_p\left(N^{-1/2}\right)$. Equation (52) therefore follows from (95), (96), and (97), by summing the asymptotic variances of the latter two formulas.

In order to prove (53), we proceed similarly as for (52) and split the estimation error

$$\begin{aligned}
\hat{p}_0 - \bar{p}_0 &= \sum_s \bar{\rho}_s (\hat{p}_{0s} - p_{0s}) \\
&\quad + \sum_s p_{0s} (\hat{\rho}_s - \bar{\rho}_s) + o_p\left(N^{-1/2}\right)
\end{aligned} \tag{98}$$

into a sum of three terms. By an argument similar to the one that led to (96) and (97), we find that

$$\begin{aligned}
N^{1/2} \sum_s \bar{\rho}_s (\hat{p}_{0s} - p_{0s}) &\longrightarrow_{\mathcal{L}} N(0, V_3), \\
N^{1/2} \sum_s p_{0s} (\hat{\rho}_s - \bar{\rho}_s) &\longrightarrow_{\mathcal{L}} N(0, V_4),
\end{aligned} \tag{99}$$

with asymptotic variances $V_3 = \sum_s \bar{\rho}_s^2 A_{ss}$ and $V_4 = \sum_{r,s} p_{0r} p_{0s} C_{rs}$, in agreement with (55). Formula (53) follows from (99) and the fact that the two main terms on the right hand side of (98) are asymptotically independent (which is a consequence of (51)).

Only (54) remains to be proven. To this end, write

$$\hat{I}_T^+ = I_T^+ - \log \frac{\bar{p}_0}{p_0} + \log \frac{\hat{p}}{p} - \log \frac{\hat{p}_0}{\bar{p}_0}. \tag{100}$$

Consequently, by a Taylor expansion of the logarithmic function around 1,

$$\begin{aligned}
\hat{I}_T^+ &= I_T^+ - \log(\bar{p}_0 / p_0) + (R_1 + R_2)/p \\
&\quad - (R_3 + R_4)/\bar{p}_0 + o_p\left(N^{-1/2}\right),
\end{aligned} \tag{101}$$

where $R_1 = \sum_s \rho_{\pi s} (\hat{p}_{0s} - p_{0s})$ and $R_2 = \sum_s (\hat{\rho}_{\pi s} - \rho_{\pi s}) p_{0s}$ are the first two terms on the right hand side of (95), whereas $R_3 = \sum_s \bar{\rho}_s (\hat{p}_{0s} - p_{0s})$ and $R_4 = \sum_s p_{0s} (\hat{\rho}_s - \bar{\rho}_s)$ denote the first two terms on the right hand side of (98).

In analogy with (96), (97) and (99), it can be shown that

$$\begin{aligned}
&N^{1/2}(R_1, R_2, R_3, R_4) \\
&\longrightarrow_{\mathcal{L}} N\left((0,0,0,0), \begin{pmatrix} V_1 & 0 & V_5 & 0 \\ 0 & V_2 & 0 & V_6 \\ V_5 & 0 & V_3 & 0 \\ 0 & V_6 & 0 & V_4 \end{pmatrix} \right),
\end{aligned} \tag{102}$$

as $N \to \infty$, with $V_1, V_2, V_3, V_4, V_5, V_6$ as defined in (55). The proof of (54) is finalized by combining (101) and (102).

Proof of Lemma 2: There are N_{Ts0} tested individuals with symptoms s and no disease, and each one of them is independently classified as diseased with probability α_s. From this it follows that the total number of subjects with symptoms s and no disease, that are reported as diseased, is $N_{Ts0}\breve{\alpha}_s$, with a binomial distribution $N_{Ts0}\breve{\alpha}_s | N_{Ts0} \sim \text{Bin}(N_{Ts0}, \alpha_s)$ conditionally on N_{Ts0}. Since $N_{Ts0} = N\rho_{\pi s}(1 - p_{0s}) + o_p(N)$, and $E(\breve{\alpha}_s | N_{Ts0}) = \alpha_s$, by first moment properties of the binomial distribution it follows that

$$\begin{aligned}
N\text{Var}(\breve{\alpha}_s - \alpha_s) &= NE[\text{Var}(\breve{\alpha}_s - \alpha_s | N_{Ts0})] \\
&= NE[\alpha_s(1 - \alpha_s)/N_{Ts0}] \\
&\to \alpha_s(1 - \alpha_s)/[\rho_{\pi s}(1 - p_{0s})] \\
&= \Sigma_{\alpha\alpha s}
\end{aligned} \tag{103}$$

as $N \to \infty$, where in the second step we made use of the formula for the variance of a binomial distribution. Weak convergence of $\breve{\alpha}_s$ in (75) follows from (103) and the Central Limit Theorem, applied to the binomial distribution. Weak converge of $\breve{\beta}_s$ in (76) is proved in the same way.

Proof of Theorem 2: It follows from (74), (78), and a Taylor expansion of the difference between these two equations, that

$$\begin{aligned}
\hat{p}_{0s} - \bar{p}_{0s} &= K_{s1}(N_{Ts1}/N_{Ts} - p_{0s}) \\
&\quad + K_{s2}(\breve{\alpha}_s - \alpha_s) + K_{s3}(\breve{\beta}_s - \beta_s) \\
&\quad + K_{s4}(\hat{\alpha}_s - \bar{\alpha}_s) + K_{s5}(\hat{\beta}_s - \bar{\beta}_s) \\
&\quad + o_p(N^{-1/2}),
\end{aligned} \tag{104}$$

with

$$\begin{aligned}
K_{s1} &= (1 - \alpha_s - \beta_s)/K_s \\
K_{s2} &= (1 - p_{0s})/K_s, \\
K_{s3} &= -p_{0s}/K_s, \\
K_{s4} &= \left[\alpha_s + \bar{\beta}_s - 1 + p_{0s}(1 - \alpha_s - \beta_s)\right]/K_s^2, \\
K_{s5} &= \left[\alpha_s - \bar{\alpha}_s + p_{0s}(1 - \alpha_s - \beta_s)\right]/K_s^2,
\end{aligned}$$

$$K_s = 1 - \bar{\alpha}_s - \bar{\beta}_s. \tag{105}$$

Making use of (75)-(77) and (88), it follows from (104) that

$$N^{1/2}\left(\hat{\boldsymbol{p}}_0 - \bar{\boldsymbol{p}}_0\right) \longrightarrow_{\mathcal{L}} N\left(0, \bar{A}\right) \tag{106}$$

as $N \to \infty$, with $\bar{A} = \left(\bar{A}_{rs}\right)$ having components

$$\begin{aligned}\bar{A}_{ss} = {} & K_{s1}^2 A_{ss} + K_{s2}^2 \Sigma_{\alpha\alpha s} + K_{s3}^2 \Sigma_{\beta\beta s} + K_{s4}^2 \Omega_{\alpha\alpha ss} \\ & + K_{s5}^2 \Omega_{\beta\beta ss} + 2K_{s4}K_{s5}\Omega_{\alpha\beta ss},\end{aligned} \tag{107}$$

and

$$\begin{aligned}\bar{A}_{rs} = {} & K_{r4}K_{s4}\Omega_{\alpha\alpha rs} + K_{r5}K_{s5}\Omega_{\beta\beta rs} \\ & + K_{r4}K_{s5}\Omega_{\alpha\beta rs} + K_{r5}K_{s4}\Omega_{\alpha\beta sr}\end{aligned} \tag{108}$$

when $r \neq s$. The rest of the proof of Theorem 2 is analogous to the proof of Theorem 1.

Verifying formulas (67)-(69): We will show that the MNAR model of Example 4 satisfies the assumptions of Theorem 1, i.e. that formulas (67)-(69) of the follow-up Example 5 hold. Since there are only $S = 2$ symptom classes, their estimated sizes, before and after correction for testing bias, satisfy

$$\begin{aligned}\hat{\rho}_{\pi 0} + \hat{\rho}_{\pi 1} &= 1, \\ \hat{\rho}_0 + \hat{\rho}_1 &= 1.\end{aligned} \tag{109}$$

From this it follows that

$$\begin{aligned}B_{00} = B_{11} = -B_{01} = -B_{10}, \\ C_{00} = C_{11} = -C_{01} = -C_{10}, \\ D_{00} = D_{11} = -D_{01} = -D_{10},\end{aligned} \tag{110}$$

and the upper equation is also a direct consequence of the explicit formula for B_{rs} in (49). It suffices, in view of (110), to establish (67) and (68) for C_{11} and D_{11}, with $\Sigma_{\pi\rho s}$ as in (69). It is possible, because of (10) and (42), to rewrite the expression for $\hat{\rho}_1$ in (66) as

$$\hat{\rho}_1 = \frac{1}{2}\hat{\rho}_{\pi 1}(1 + \hat{\pi}), \tag{111}$$

with asymptotic limit

$$\bar{\rho}_1 = \frac{1}{2}\rho_{\pi 1}(1 + \pi). \tag{112}$$

Taking the difference between (111) and (112), we find that

$$\begin{aligned}\hat{\rho}_1 - \bar{\rho}_1 = {} & \frac{1}{2}(1 + \pi)(\hat{\rho}_{\pi 1} - \rho_{\pi 1}) \\ & + \frac{1}{2}\rho_{\pi 1}(\hat{\pi} - \pi) + o_p\left(N^{-1/2}\right).\end{aligned} \tag{113}$$

Similarly as in the proof of Theorem 1, it can be shown that

$$\begin{aligned}& N^{1/2}\left(\hat{\rho}_1 - \bar{\rho}_1, \hat{\pi} - \pi\right) \\ & \longrightarrow_{\mathcal{L}} N\left((0,0), \begin{pmatrix} B_{11} & \Sigma_{\pi\rho 1} \\ \Sigma_{\pi\rho 1} & \Sigma_{\pi}\end{pmatrix}\right)\end{aligned} \tag{114}$$

as $N \to \infty$, where

$$B_{11} = \rho_0\pi_0\rho_1\pi_1(\pi - \pi_0\pi_1)$$

is deduced from (49) when $S = 2$,

$$\Sigma_{\pi} = N\mathrm{Var}(\hat{\pi}) = \rho_0\pi_0(1 - \pi_0) + \rho_1\pi_1(1 - \pi_1)$$

is taken from (50), and

$$\Sigma_{\pi\rho 1} = \lim_{N\to\infty} \mathrm{Cov}(\hat{\pi}, \hat{\pi}_{\pi 1}).$$

Because of (113) and (114) we have that

$$\begin{aligned}C_{11} &= \lim_{N\to\infty} N\mathrm{Var}(\hat{\rho}_1) \\ &= \frac{(1+\pi)^2 B_{11} + \rho_{\pi 1}^2 \Sigma_{\pi} + 2(1+\pi)\rho_{\pi 1}\Sigma_{\pi\rho 1}}{4},\end{aligned}$$

and

$$\begin{aligned}D_{11} &= \lim_{N\to\infty} N\mathrm{Cov}(\hat{\rho}_{\pi 1}, \hat{\rho}) \\ &= (1+\pi)B_{11}/2 + \rho_{\pi 1}\Sigma_{\pi\rho 1}/2,\end{aligned}$$

in agreement with (67) and (68) respectively. It remains to establish the formula for $\Sigma_{\pi\rho s}$ in (69). To this end, we make use of (92) and (94) in order to write

$$\begin{aligned}\Sigma_{\pi\rho s} &= \lim_{N\to\infty} N\mathrm{Cov}(\hat{\pi}, \hat{\rho}_{\pi 1}) \\ &= \lim_{N\to\infty} N\mathrm{Cov}\left(\sum_r \varepsilon_r, \varepsilon_s/\pi - \rho_s\pi_s \sum_r \varepsilon_r/\pi^2\right) \\ &= \rho_s\pi_s(1 - \pi_s)/\pi - \rho_s\pi_s\Sigma_{\pi}/\pi^2,\end{aligned}$$

where in the last step we utilized (50), (93), and the independence of $\{\varepsilon_r\}_{r=0}^{S-1}$. □

REFERENCES

[1] D. H. Wolpert and W. G. Macready, "No free lunch theorems for optimization," *IEEE Trans. Evol. Comput.*, vol. 1, no. 1, pp. 67–82, Apr. 1997.

[2] W. A. Dembski and R. J. Marks, "Bernoulli's principle of insufficient reason and conservation of information in computer search," in *Proc. IEEE Int. Conf. Syst., Man Cybern.*, San Antonio, TX, USA, Oct. 2009, pp. 2647–2652.

[3] W. A. Dembski and R. J. Marks II, "Conservation of information in search: Measuring the cost of success," *IEEE Trans. Syst., Man, Cybern., A, Syst. Hum.*, vol. 39, no. 5, pp. 1051–1061, Sep. 2009.

[4] D. A. Díaz-Pachón, J. P. Sáenz, J. S. Rao, and J. Dazard, "Mode hunting through active information," *Appl. Stochastic Models Bus. Ind.*, vol. 35, no. 2, pp. 376–393, Mar. 2019.

[5] T. Liu, D. A. Díaz-Pachón, J. S. Rao, and J.-E. Dazard, "High dimensional mode hunting using pettiest components analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 4637–4649, Apr. 2023.

[6] C. Hom, A. Maina-Kilaas, K. Ginta, C. Lay, and G. Montañez, "The Gopher's gambit: Survival advantages of artifact-based intention perception," in *Proc. 13th Int. Conf. Agents Artif. Intell.*, 2021, pp. 205–215.

[7] D. A. Díaz-Pachón and O. Hössjer, "Assessing, testing and estimating the amount of fine-tuning by means of active information," *Entropy*, vol. 24, no. 10, p. 1323, Sep. 2022.

[8] R. Marks and D. A. D. Pachón, "Active information requirements for fixation on the Wright–Fisher model of population genetics," *BioComplex.*, vol. 2020, no. 4, pp. 1–6, Mar. 2020.

[9] E. Hargittai, "Potential biases in big data: Omitted voices on social media," *Social Sci. Comput. Rev.*, vol. 38, no. 1, pp. 10–24, Feb. 2020.

[10] Y. Zhao, P. Yin, and Y. Li, "Data and model biases in social media analyses: A case study of COVID-10 tweets," in *Proc. AMIA Annu. Symp.*, 2022, pp. 1264–1273.

[11] E. Hargittai and G. Karaoglu, "Biases of online political polls: Who participates?" *Socius*, vol. 4, Aug. 2018, Art. no. 2378023118791080.

[12] D. A. Díaz-Pachón and J. S. Rao, "A simple correction for COVID-19 sampling bias," *J. Theor. Biol.*, vol. 512, Mar. 2021, Art. no. 110556.

[13] J. Schafer, *Analysis of Incomplete Multivariate Data*. London, U.K.: CRC Press, 1997.

[14] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data*, 2nd ed. Hoboken, NJ, USA: Wiley, 2002.

[15] J. Qin, J. Shao, and B. Zhang, "Efficient and doubly robust imputation for covariate dependent missing response," *J. Amer. Stat. Assoc.*, vol. 103, pp. 793–810, Jan. 2008.
[16] M. Sverchkov and D. Pfeffermann, "Prediction of finite population totals based on sample distribution," *Surv. Methodol.*, vol. 30, pp. 79–92, Jan. 2004.
[17] J. Kim and J. Shao, *Statistical Methods for Handling Incomplete Data*, 2nd ed. Boca Raton, NJ, USA: Chapman & Hall, 2021.
[18] R. M. Groves, F. J. Fowler Jr., M. P. Couper, J. M. Lepkowski, E. Singer, and R. Tourangeau, *Survey Methodology*, 2nd ed. Hoboken, NJ, USA: Wiley, 2009.
[19] J. Beaumont, "An estimation method for nonignorable nonresponse," *Surv. Methodol.*, vol. 26, no. 2, pp. 131–136, 2000.
[20] J. S. Greenlees, W. S. Reece, and K. D. Zieschang, "Imputation of missing values when the probability of response depends on the variable being imputed," *J. Amer. Stat. Assoc.*, vol. 77, no. 378, pp. 251–261, Jun. 1982.
[21] R. J. A. Little, "Pattern-mixture models for multivariate incomplete data," *J. Amer. Stat. Assoc.*, vol. 88, no. 421, pp. 125–134, Mar. 1993.
[22] D. Rubin, *Multiple Imputation for Nonresponse in Surveys*. New York, NY, USA: Wiley, 1987.
[23] J. Qin, D. Leung, and J. Shao, "Estimation with survey data under nonignorable nonresponse or informative sampling," *J. Amer. Stat. Assoc.*, vol. 97, no. 457, pp. 193–200, Mar. 2002.
[24] M. Sverchkov, "A new approach to estimation of response probabilities when missing data are not missing at random," in *Proceedings of the Survey Research Methods Section*. Alexandria, VA, USA: American Statistical Association, 2008, pp. 867–874.
[25] M. Sverchkov and D. Pfeffermann, "Small area estimation under informative sampling and not missing at random non-response," *J. Roy. Stat. Soc. A, Statist. Soc.*, vol. 181, no. 4, pp. 981–1008, Oct. 2018.
[26] M. K. Riddles, J. K. Kim, and J. Im, "A propensity-score-adjustment method for nonignorable nonresponse," *J. Surv. Statist. Methodol.*, vol. 4, no. 2, pp. 215–245, Jun. 2016.
[27] J. N. K. Rao, "On double sampling for stratification and analytical surveys," *Biometrika*, vol. 60, no. 1, pp. 125–133, Apr. 1973.
[28] J. K. Kim and J. N. K. Rao, "Combining data from two independent surveys: A model-assisted approach," *Biometrika*, vol. 99, no. 1, pp. 85–100, Mar. 2012.
[29] D. Pfeffermann and A. Sikov, "Imputation and estimation under nonignorable nonresponse in household services with missing covariate information," *J. Off. Statist.*, vol. 27, no. 2, pp. 181–209, 2011.
[30] L. Zhou, D. A. Díaz-Pachón, C. Zhao, J. S. Rao, and O. Hössjer, "Correcting prevalence estimation for biased sampling with testing errors," *Statist. Med.*, vol. 42, no. 26, pp. 4713–4737, Nov. 2023.
[31] R. Marks and D. A. Díaz-Pachón, "Generalized active information: Extensions to unbounded domains," *Bio-Complex.*, vol. 2020, no. 3, pp. 1–6, Mar. 2020.
[32] E. T. Jaynes, "Prior probabilities," *IEEE Trans. Syst. Sci. Cybern.*, vol. SSC-4, no. 3, pp. 227–241, Sep. 1968.
[33] A. Agresti, *Categorical Data Analysis*, 3rd ed. Hoboken, NJ, USA: Wiley, 2013.
[34] E. L. Lehmann and G. Casella, *Theory of Point Estimation*, 2nd ed. Cham, Switzerland: Springer, 1998.
[35] G. Patil and C. Rao, "Weighted distributions and size-biased sampling with applications to wildlife populations and human families," *Biometrics*, vol. 34, pp. 179–189, Jan. 1978.
[36] A. Gut, *An Intermediate Course in Probability Theory*, 2nd ed. Cham, Switzerland: Springer, 1995.

Ola Hössjer has been a Professor of mathematical statistics with Stockholm University, Sweden, since 2002. He has done research in statistics and probability theory, with applications in population genetics, epidemiology, and insurance mathematics. He is the author of about 110 peer-reviewed publications. He has supervised 14 Ph.D. students. In 2009, he received the Gustafsson Prize in Mathematics.

Daniel Andrés Díaz-Pachón (Member, IEEE) received the B.S. degree in mathematical statistics from Universidad Nacional de Colombia, Colombia, in 2005, and the Ph.D. degree in probability theory from Universidade de São Paulo, Brazil, in 2009. In 2011, he moved to the University of Miami, FL, USA, where he was a Post-Doctoral Associate in biostatistics (2011–2015) and then became a Research Assistant Professor. His research is focused on the intersection of probability theory, statistics, machine learning, and information theory.

Chen Zhao received the B.S. degree in statistics from Beijing Technology and Business University, China, in 2018, and the M.S. degree in actuarial science from The Ohio State University in 2020. He is currently pursuing the Ph.D. degree in biostatistics with the University of Miami. His research interests are small area estimation, linear mixed models, and information theory.

J. Sunil Rao received the Ph.D. degree. He is a Professor with the Division of Biostatistics, University of Minnesota, Twin Cities, where he is also the Director of biostatistics with the Masonic Cancer Center. His research interests include mixed model prediction and selection, Bayesian model selection, small area estimation, machine learning, and applied biostatistics, with a focus on cancer and health disparities.