



## Assessing COVID-19 Prevalence in Austria with Infection Surveys and Case Count Data as Auxiliary Information

Stéphane Guerrier, Christoph Kuzmics & Maria-Pia Victoria-Feser

To cite this article: Stéphane Guerrier, Christoph Kuzmics & Maria-Pia Victoria-Feser (08 Mar 2024): Assessing COVID-19 Prevalence in Austria with Infection Surveys and Case Count Data as Auxiliary Information, Journal of the American Statistical Association, DOI: [10.1080/01621459.2024.2313790](https://doi.org/10.1080/01621459.2024.2313790)

To link to this article: <https://doi.org/10.1080/01621459.2024.2313790>



© 2024 The Author(s). Published with license by Taylor & Francis Group, LLC.



[View supplementary material](#)



Published online: 08 Mar 2024.



[Submit your article to this journal](#)



Article views: 579



[View related articles](#)



[View Crossmark data](#)

# Assessing COVID-19 Prevalence in Austria with Infection Surveys and Case Count Data as Auxiliary Information

Stéphane Guerrier<sup>a</sup>, Christoph Kuzmics<sup>b</sup>, and Maria-Pia Victoria-Feser<sup>c,d</sup>

<sup>a</sup>Faculty of Science and Geneva School of Economics and Management, University of Geneva, Geneva, Switzerland; <sup>b</sup>Department of Economics, University of Graz, Graz, Austria; <sup>c</sup>Geneva School of Economics and Management, University of Geneva, Geneva, Switzerland; <sup>d</sup>Department of Statistical Sciences, University of Bologna, Bologna, Italy

## ABSTRACT

Countries officially record the number of COVID-19 cases based on medical tests of a subset of the population. These case count data obviously suffer from participation bias, and for prevalence estimation, these data are typically discarded in favor of infection surveys, or possibly also completed with auxiliary information. One exception is the series of infection surveys recorded by the Statistics Austria Federal Institute to study the prevalence of COVID-19 in Austria in April, May, and November 2020. In these infection surveys, participants were additionally asked if they were simultaneously recorded as COVID-19 positive in the case count data. In this article, we analyze the benefits of properly combining the outcomes from the infection survey with the case count data, to analyze the prevalence of COVID-19 in Austria in 2020, from which the case ascertainment rate can be deduced. The results show that our approach leads to a significant efficiency gain. Indeed, considerably smaller infection survey samples suffice to obtain the same level of estimation accuracy. Our estimation method can also handle measurement errors due to the sensitivity and specificity of medical testing devices and to the nonrandom sample weighting scheme of the infection survey. The proposed estimators and associated confidence intervals are implemented in the companion open source R package `pemp_i` available on the Comprehensive R Archive Network (CRAN). Supplementary materials for this article are available online including a standardized description of the materials available for reproducing the work.

## ARTICLE HISTORY

Received November 2022  
Accepted December 2023

## KEYWORDS

Clopper-Pearson confidence interval; Case ascertainment rate; (Generalized) method of moments; Infectious disease; Measurement error; Maximum likelihood estimation; Sample proportion; Stratified sampling


## 1. Introduction


In the COVID-19 pandemic, governments faced a tradeoff between reducing the wealth or the health of citizens when choosing the degree of economic slowdown in their policy measures. The key to assess this tradeoff is an understanding of the number or proportion of cases in the population and their evolution. Acquiring this understanding, in turn, depends on reliable estimates of the number of cases (at different points in time).

The data collected by official national institutions are case count data with possibly some additional information such as gender, age and geographical location. These data can also be understood as a nonprobability sample and, as such, the number of recorded positive cases can only be seen as a lower bound of the actual number of cases. Hence, the direct analysis of these count data suffers from bias, due to participation bias, and possibly also from measurement error, due, in particular, to medical testing devices (see also Accorsi et al. 2021; Kahn et al. 2021; Dempsey 2023, for bias effects of case count data in epidemiological studies). Acknowledging this problem, for the case of COVID-19, some studies have proposed estimates for the prevalence among asymptomatic patients (see e.g., Mizumoto et al. 2020; Nishiura et al. 2020), or have attempted to infer the

population prevalence from the case count data (see e.g., Manski and Molinari 2021; Dempsey 2023).

One way of reducing the sampling bias in case count data, is to estimate selection propensities, using representative random samples that can provide necessary covariate information (see e.g., Elliott and Valliant 2017, and the references therein). These propensities are then used in an inverse-probability weighting scheme to construct estimators of disease prevalence that attempt to control for selection bias. Using a random sample as auxiliary information for reducing the sampling bias has also been proposed by Chen, Li, and Wu (2020) who consider a pseudo-likelihood approach that uses the random samples as a proxy for a term in the log-likelihood. Dempsey (2023) extends this approach to account for measurement error. Alternatively, the prevalence can be estimated using infection surveys, typically stratified with associated sampling weights. As will be explained in Section 2 for the Austrian data, the sampling weights are often calibrated for removing the potential sampling bias such as participation bias. Infection samples are obviously more costly to collect, so they cannot provide information on a daily basis. However, if the sampling is properly done and the sampling weights are properly calibrated, they allow to obtain consistent estimators of different population measures, such as the prevalence. Moreover, the resulting prevalence estimate can

**CONTACT** Maria-Pia Victoria-Feser  [maria-pia.victoriafeser@unige.ch](mailto:maria-pia.victoriafeser@unige.ch)  Geneva School of Economics and Management, University of Geneva, Geneva, Switzerland.

 Supplementary materials for this article are available online. Please go to [www.tandfonline.com/r/JASA](http://www.tandfonline.com/r/JASA).

© 2024 The Author(s). Published with license by Taylor & Francis Group, LLC.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

be used to estimate the ascertainment rate (see e.g., Gibbons et al. 2014), which is inversely proportional to the proportion of positive cases that are actually not in the case count dataset, that we qualify as the *asymptomatic cases* for simplicity. This information can then be used to monitor and correct the prevalence estimation between two consecutive surveys, using only the case count data.

In order to alleviate the burden of data collection with infection surveys, we here evaluate the potential statistical gain of appropriately including the information available from the case count data in the analysis of the infection survey data. The statistical gain is relative efficiency, and we show that by linking the information provided by the case count data to the infection survey for prevalence estimation with the Austrian data, for the same statistical accuracy, only half of the infection survey sample size is needed, thus, substantially reducing the costs and/or time for data acquisition. Alternatively said, from the same infection survey, finer analysis at sub-population levels (e.g., regions) could reasonably be done even if the number of participants in these levels is rather small. For that, the additional information that is needed is to also record, for at least each participant found positive in the infection survey, whether they are currently recorded positive in the (national) case count data. Since at the beginning of a pandemic the prevalence is expected to be rather small, collecting this information, using for example follow-up calls, is rather easy to implement. This is what has been done in Austria, where the Federal Ministry of Education, Science and Research commissioned Statistics Austria in cooperation with the Medical University of Vienna and the Red Cross Austria to study the prevalence of COVID-19 in Austria in April, May, and November 2020. For this purpose, a representative stratified sample was drawn based on the Austrian central register of residents, inviting addressed participants to participate for each study, respectively. We discuss these data in some detail in Section 2; for more information see Kowarik et al. (2022). Additionally, they also collected, for each participant found positive in the infection survey, whether they were also recorded positive at the time in the national case count, and we consider in this article, both the infection survey and this additional information.

Our framework supposes that the sampling weights are properly calibrated, up to possible measurement errors (misclassification) of the (medical) testing devices used to collect the data (see e.g., Kobokovich, West, and Gronvall 2020; Surkova, Nikolayevskyy, and Drobniowski 2020). As is the case with the Austrian infection surveys, this calibration is done before the infection survey data are used to estimate the prevalence. For this type of settings, we develop estimators that include both sources of information, for stratified samples, accounting as well for measurement error due to the medical testing device. The associated measurement errors are induced by their sensitivity, that is, the complement to the False Positive (FP) rate, and by their specificity, that is, the complement to the False Negative (FN) rate, and adjusting for these errors avoids biased estimates (see e.g., Diggle 2011; Lewis and Torgerson 2012, and the references therein). The proposed estimators are based on standard assumptions, and the only necessary and known quantities are the FP and FN rates of the medical test used to collect the data in the infection survey, as well as the sampling weights. For

computing the prevalence with associated Confidence Intervals (CI), one can use the Proportion Estimation with Marginal Proxy Information, or `pempi` R package, that we developed for that purpose. This software also allows to reproduce all simulation results and the data analysis presented in this article (see Section 6 for more details).

As is shown and illustrated in this article, the proposed estimators have a considerably reduced variability compared to the ones that ignore the information provided by the case count data. Moreover, we also find the following additional advantages. First, using a sensitivity analysis with the Austrian infection survey, we find that the proposed estimators are much less influenced by the value of the FN rate than the infection survey (weighted) proportion, allowing, in practice, to limit the impact of the potentially incorrect choice for the medical test specificity when estimating the prevalence. Second, within the proposed statistical framework, it is possible to obtain an estimate of the proportion of asymptomatic (or mild) cases, which is of major importance for controlling the spread of an infectious disease lacking severe disease manifestations (see e.g., Munster et al. 2020, in the case of the COVID-19). In particular, it can be directly used to adjust prevalence estimators computed from case count data between two consecutive infection surveys, as well as estimating ascertainment rates.

The article is organized as follows. We first present the collection process of the data we use in this article in Section 2. In Section 3, we present the methodological framework, while keeping the more technical aspects in the supplementary material. In particular, we present different estimators that include the information obtained from the case count data, with possibly missing information, and that are corrected for measurement errors due to the FP and FN rates. We also study the efficiencies of the different estimators. The COVID-19 prevalence analysis in Austria, November 2020, is presented in Section 4, and in Section 5 we provide the results of simulation studies to assess the comparative properties of the prevalence estimators. Finally, we provide some concluding remarks in Section 6.

## 2. Austrian Infection Surveys in April, May, and November 2020

In this section, we present the Austrian data collection process, for the case count data and the infection surveys. The official description is given in Kowarik et al. (2022).

The case count data are based on participants who had symptoms or thought they had symptoms calling a medical hotline number (i.e., 1450). Depending on the information provided by the participants, depending on their assessment by the medical staff, and depending on testing availability, participants were visited and tested with a Polymerase Chain Reaction (PCR) test using “the fully automatic Roche cobas® 6800 Test System using the Roche cobas® SARS-CoV-2 Test (CE/IVD)” (Kowarik et al. 2022, p. 32). A participant was officially recorded as a positive case of COVID-19 in the case count data, if the participant was then tested and found to be positive by means of the PCR test.

For the infection surveys, Statistics Austria took three stratified random samples from the Austrian population aged 16 or older living in private homes in April, May, and November 2020.

In this case study we focus on the November 2020 one, which was the only one with a substantial number of cases. It was also the only one in which the sample was drawn in one step from the entire population. Indeed, in the April and May infection surveys, the samples included a local clustering of participants to make the data collection (the PCR testing) easier for the Austrian Red Cross who performed this task.

The November infection survey was a stratified sample based on the Austrian central registry of residents (i.e., ZMR). Participants were invited by letter (almost one month before the PCR test, October 13 or 20—in two tranches) and reminded by postcard about six days after they received the letter. Ultimately of the 7823 invited participants 2263 were tested for COVID-19 by means of a PCR test. Stratification was based on information about federal state (there are nine in Austria), education, urbanization, and citizenship. In anticipation of higher nonresponse rates for participants with lower education (i.e., compulsory schooling or less), such participants were oversampled. Once the data were collected, the infection survey weights were calibrated by means of an iterated procedure based on information about participants’ age, gender, degree of urbanization, household size, federal state, risk category, nationality, education, underlying condition, and educational attainment, and some other information. Finally, nonresponse modeling was done based on, among others, information about participants’ age, nationality, education, quintile of household income, and the relative number of infected participants in the participants’ communal area.

Regarding the sampling weights, Kowarik et al. (2022) mention that their nonresponse modeling may not fully account for whether a person considers themselves at high risk or not. This, they argue, may generate a positive or negative bias, depending on whether high risk individuals are more or less likely to participate (because they are more interested to know their health status or they are more concerned of becoming infected when participating). Thus, the sampling weights may not fully correct for nonresponse related biases, however, we expected these potential biases to have a relatively small effect on prevalence estimation.

### 3. Methodology

In this section, we first define the required random variables, as well as the needed assumptions, which we discuss in detail in Appendix A. We also introduce the different forms of measurement errors, with some additional explanation provided in Appendix B. We then present the sufficient statistics as well as their associated success probabilities, which are formally derived in Appendix C. Based on these sufficient statistics, we then provide several estimators for the prevalence that take into account the information provided by the case count data, the measurement errors, and stratified sampling. Their formal derivations and properties are given in supplementary material A, that also includes the cases with non-stratified sampling. Finally, we present the inferential properties of the prevalence estimators and the estimator for the rate of asymptomatic cases, and assess the efficiency gain when using a prevalence estimator that includes the information provided by the case count data.

### 3.1. General Framework

Recall that the aim of this article is to augment infection surveys with case count data in order to increase the estimation efficiency. We first define the following unobserved random variable:

$$X_i := \begin{cases} 1 & \text{if participant (in the infection survey) } i \text{ is infected,} \\ 0 & \text{otherwise.} \end{cases}$$

The variable  $X_i$  is not observable because we allow the possibility that the measurement can be subject to error due to the, assumed known, FN and FP rates of the PCR testing device. The objective is to provide an estimator for the unknown population proportion, that is, the prevalence  $\pi := \mathbb{P}(X_i = 1)$ . To do so, we consider the following two (observed) random variables:

$$\begin{aligned} Y_i &:= \begin{cases} 1 & \text{if participant } i \text{ is tested positive in the infection survey,} \\ 0 & \text{otherwise;} \end{cases} \\ Z_i &:= \begin{cases} 1 & \text{if participant } i \text{ is declared positive in the case count data,} \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \tag{1}$$

For the proposed general framework to hold, we need to make the following assumptions.

**Assumption 1.** Conditional on  $X_i$ , the random variables  $Z_i$  and  $Y_i$  are stochastically independent.

**Assumption 1** means that for an infected participant  $i$  ( $X_i = 1$ ), or indeed also for an uninfected participant ( $X_i = 0$ ), whether or not they were declared positive in the case count data ( $Z_i = 1$ ), is an independent event from whether or not they were found positive with the PCR test in the infection survey ( $Y_i = 1$ ). This (indirectly) implies that the occurrence of a FP test is independent of the participants’ characteristics (other than  $X_i$ ), or, in other words, that the occurrence of a FP test is purely a characteristic of the PCR test itself. A more detailed justification for **Assumption 1** to hold (for our case study) is provided in Appendix A. In practical settings, this assumption appears to be plausible and is not required, in particular, when medical tests without measurement error are employed. Moreover, the effect on inference of potential deviations from **Assumption 1** are assessed by simulations in Section 5.2, which suggest that this effect is negligible.

**Remark 1.** **Assumption 1** implies that  $\mathbb{P}(Z_i = 1|X_i = 1, Y_i = 1) = \mathbb{P}(Z_i = 1|X_i = 1)$ , and  $\mathbb{P}(Z_i = 1|X_i = 0, Y_i = 1) = \mathbb{P}(Z_i = 1|X_i = 0)$ , which are used to derive the success probabilities provided in Appendix C.

We also allow for the possibility that the outcome of (PCR) tests can be subject to measurement error. Hence, we define  $\alpha := \mathbb{P}(Y_i = 1|X_i = 0)$  and  $\beta := \mathbb{P}(Y_i = 0|X_i = 1)$ . The probabilities  $\alpha$  and  $\beta$ , are the (assumed known) FP rate ( $\alpha = 1 - \text{specificity}$ ) and FN rate ( $\beta = 1 - \text{sensitivity}$ ) of the particular PCR test employed in the infection survey. Moreover, we will make use of the *case prevalence rate*  $\pi_0 := \mathbb{P}(Z_i = 1)$  obtained from the case count data. As explained in Appendix C,  $\pi_0$  is the joint probability of being selected for the case count data and declared positive. It can be considered as the proportion of positive rates among the whole population, obtained from

**Table 1.** Definition of the most important parameters.

Symbol	Definition	Interpretation
$\pi$	$\mathbb{P}(X_i = 1)$	Prevalence
$\alpha$	$\mathbb{P}(Y_i = 1 X_i = 0)$	PCR false positive rate
$\beta$	$\mathbb{P}(Y_i = 0 X_i = 1)$	PCR false negative rate
$\Delta$	$1 - (\alpha + \beta)$	A measure of PCR test precision
$\pi_0$	$\mathbb{P}(Z_i = 1)$	Case prevalence rate
$\alpha_0$	$\mathbb{P}(Z_i = 1 X_i = 0)$	False case positive rate
$\beta_0$	$\mathbb{P}(Z_i = 0 X_i = 1)$	False case negative rate

a census with participation bias, so that we assume it is fixed. Moreover, we expect  $\pi_0 \leq \pi$  (as justified in [Remark 4](#)).

Since the objective is to take advantage of the information obtained from the case count data in estimating the prevalence  $\pi$ , we also need to take into account the possible biases in the case count data. Therefore, we define  $\alpha_0 := \mathbb{P}(Z_i = 1|X_i = 0)$  and  $\beta_0 := \mathbb{P}(Z_i = 0|X_i = 1)$ . It turns out, as explained in [Remarks 2](#) and [3](#), that  $\alpha_0$  is a negligible quantity (for our case study) and hence can be set to 0, and that  $\beta_0$  can be deduced from the other available or estimable quantities and can be seen as the rate of asymptomatic cases (and of other cases overlooked by the case count data). For convenience, the definition of the most relevant parameters is provided in [Table 1](#).

*Remark 2.* As explained in [Appendix B](#),  $\alpha_0$  is the probability that participant  $i$ , who is uninfected ( $X_i = 0$ ), has been first selected for assessment and then tested positive and, thus, recorded in the case count data. Obviously,  $\alpha_0$  can be considered as almost nil, as it is essentially the product of the probability of a person being falsely selected for assessment and the probability of this person being falsely declared positive through the medical test.

*Remark 3.* The rate  $\beta_0$  is the proportion of infected cases that were either not tested and hence not reported in the case count data or tested and nevertheless found to be negative (with the latter probably much less likely than the former). It therefore provides, approximately, the proportion of asymptomatic or mild cases and other cases that could not undergo a testing for the case count data. It is given by

$$\beta_0 = 1 - \frac{\pi_0 - \alpha_0(1 - \pi)}{\pi}, \quad (2)$$

as shown in [Appendix B](#).

*Remark 4.* The unknown population prevalence  $\pi$  is bounded from below by  $\underline{\pi} := \frac{\pi_0 - \alpha_0}{1 - \alpha_0}$ . Indeed, from  $\pi_0 = (1 - \pi)\alpha_0 + \pi(1 - \beta_0)$  (with both  $\pi$  and  $\beta_0$  unknown parameters), see [Appendix B](#), we deduce that the lowest admissible value for  $\pi$  is achieved when  $\beta_0 = 0$ , in which case we get the lower bound  $\frac{\pi_0 - \alpha_0}{1 - \alpha_0}$ . Given the assumptions and the form of  $\underline{\pi}$ , we have  $0 \leq \underline{\pi} \leq \pi_0$ . In practice as discussed in [Remark 2](#), we expect that  $\alpha_0 \approx 0$ , implying that  $\pi \geq \underline{\pi} \approx \pi_0$ .

Additionally, it is useful to make the following two mild assumptions.

*Assumption 2.*  $\alpha + \beta < 1$ .

*Assumption 3.*  $\alpha_0 + \beta_0 < 1$ .

[Assumptions 2](#) and [3](#) are very mild, since they only rule out the cases  $\alpha + \beta = 1$  and  $\alpha_0 + \beta_0 = 1$  for which the PCR tests and the case count data are completely uninformative. Indeed, if  $\alpha + \beta = 1$ , one can show that the updated probability of a participant being COVID-19 infected would not change after we condition on the outcome of the PCR test. If  $\alpha + \beta > 1$  then the PCR test is so bad that a finding of a negative test result is actually an indication of a COVID-19 infection. Not only is this not a very likely situation for PCR tests, but one could also then just reformulate a positive test result as a negative one and vice versa, and the induced FP and FN rates would satisfy our assumptions. More details on the validity of these assumptions can be found in [Appendix A](#).

### 3.2. Sufficient Statistics and Associated Probabilities

For building prevalence estimators, we make use of the following random variables:

$$\begin{aligned} R_{i11} &:= Y_i Z_i, \\ R_{i10} &:= (1 - Y_i) Z_i, \\ R_{i01} &:= Y_i (1 - Z_i), \\ R_{i00} &:= (1 - Y_i) (1 - Z_i) = 1 - R_{i11} - R_{i10} - R_{i01}. \end{aligned} \quad (3)$$

Namely,  $R_{i11} = 1$ , if participant  $i$  is tested positive in the infection survey and in the case count data;  $R_{i10} = 1$ , if participant  $i$  is tested negative in the infection survey, but is declared positive in the case count data;  $R_{i01} = 1$ , if participant  $i$  is tested positive in the infection survey, but is declared negative in the case count data;  $R_{i00} = 1$ , if participant  $i$  is tested negative in the infection survey and is also declared negative in the case count data.

The associated probabilities are given by

$$\begin{aligned} \tau_{11}(\pi) &:= \mathbb{P}(R_{i11} = 1) \\ &= \pi \Delta \alpha_0 + (\pi_0 - \alpha_0)(1 - \beta) + \alpha \alpha_0, \\ \tau_{10}(\pi) &:= \mathbb{P}(R_{i10} = 1) \\ &= -\pi \Delta \alpha_0 + (\pi_0 - \alpha_0)\beta + (1 - \alpha)\alpha_0, \\ \tau_{01}(\pi) &:= \mathbb{P}(R_{i01} = 1) \\ &= \pi \Delta (1 - \alpha_0) - (\pi_0 - \alpha_0)(1 - \beta) + \alpha(1 - \alpha_0), \\ \tau_{00}(\pi) &:= \mathbb{P}(R_{i00} = 1) \\ &= -\pi \Delta (1 - \alpha_0) - (\pi_0 - \alpha_0)\beta + (1 - \alpha)(1 - \alpha_0), \end{aligned} \quad (4)$$

where  $\Delta := 1 - (\alpha + \beta)$  (see [Appendix C](#) for their derivation). Without measurement error, we would have  $\tau_{11}(\pi) = \pi_0$ ,  $\tau_{10}(\pi) = 0$ ,  $\tau_{01}(\pi) = \pi - \pi_0$ ,  $\tau_{00}(\pi) = 1 - \pi$ . Moreover, it is easy to verify that given our assumptions, we have that all probabilities defined in [\(4\)](#) are nonnegative and sum up to 1.

Even if there is no measurement error, infection surveys can provide biased estimates of, say, the population prevalence, especially when the level of nonresponse is rather high (see e.g., [Bethlehem and Schouten 2017](#)). A popular method consists in calibrating sampling weights, as has been done for the Austrian data ([Kowarik et al. 2022](#)). Therefore, we let  $\gamma_i$  denote the known (fixed) sampling weight associated to participant  $i$  with  $i = 1, \dots, n$ , which is proportional to the reciprocal of the sampling probability for participant  $i$ , and adjusted, for convenience, such that  $\sum_{i=1}^n \gamma_i = n$ . The sufficient statistics for the estimators we

propose in [Section 3.3](#) (see also supplementary material A) are given by

$$\begin{aligned}\bar{R}_{11} &:= \sum_{i=1}^n \gamma_i Y_i Z_i = \sum_{i=1}^n \gamma_i R_{i11}, \\ \bar{R}_{10} &:= \sum_{i=1}^n \gamma_i (1 - Y_i) Z_i = \sum_{i=1}^n \gamma_i R_{i10}, \\ \bar{R}_{01} &:= \sum_{i=1}^n \gamma_i Y_i (1 - Z_i) = \sum_{i=1}^n \gamma_i R_{i01}, \\ \bar{R}_{00} &:= \sum_{i=1}^n \gamma_i (1 - Y_i) (1 - Z_i) = \sum_{i=1}^n \gamma_i R_{i00}.\end{aligned}\quad (5)$$

We also make use of  $\bar{R}_{*1} := \sum_{i=1}^n \gamma_i Y_i = \bar{R}_{11} + \bar{R}_{01}$ , the weighted number of participants that are tested positive in the infection survey. As previously mentioned, our framework supposes that the sampling weights are properly calibrated, up to possible measurement errors of the testing devices used to collect the data.

### 3.3. Prevalence Estimators

In this section, we present several closed-form estimators for the prevalence  $\pi$  that take into account the information provided by the case count data. The formal derivations and properties are provided in supplementary material A, and all estimators (and associated CI) can be computed using the R package `pempi`.

First, we consider a Method of Moments Estimator (MME) based on  $\bar{R}_{01}$  (with expectation  $n\tau_{01}(\pi)$ ), which is given by (see also supplementary material A.5)

$$\tilde{\pi} = \frac{1}{\Delta(1 - \alpha_0)} \left[ \frac{\bar{R}_{01}}{n} + \pi_0(1 - \beta) - \alpha_0\Delta - \alpha \right]. \quad (6)$$

Since  $\mathbb{E}[\bar{R}_{01}] = n\mathbb{E}[R_{i01}] = n\tau_{01}(\pi)$ , it is easy to see that the MME is unbiased. When  $\alpha_0 = 0$  (see [Remark 2](#)), this reduces to

$$\tilde{\pi} = \frac{1}{\Delta} \left[ \frac{\bar{R}_{01}}{n} + \pi_0(1 - \beta) - \alpha \right]. \quad (7)$$

When  $\alpha_0 = \alpha = \beta = 0$ , this further reduces to

$$\tilde{\pi} = \pi_0 + \frac{\bar{R}_{01}}{n}. \quad (8)$$

One can also consider a Weighted  $M$ -Estimator (WME)  $\hat{\pi}$  as proposed for example by Wooldridge (2001), which is based on the conditional log-likelihood function (see supplementary material A.1). Generally, this estimator has no closed-form solution but can be computed numerically (see supplementary material A.6). However, in the case when  $\alpha_0 = 0$ , we obtain a closed-form solution given by

$$\hat{\pi} = \frac{1}{\Delta} \left( \frac{\pi_0 \bar{R}_{00} + \bar{R}_{01}}{\bar{R}_{01} + \bar{R}_{00}} - \pi_0 \beta - \alpha \right). \quad (9)$$

When  $\alpha_0 = \alpha = \beta = 0$ , then  $\bar{R}_{10} = 0$ , so that (9) further reduces to

$$\hat{\pi} = \pi_0 \frac{n - \bar{R}_{*1}}{n - \bar{R}_{11}} + \frac{\bar{R}_{01}}{(n - \bar{R}_{11})}. \quad (10)$$

*Remark 5.* Interestingly, in the case of no measurement error ( $\alpha_0 = \alpha = \beta = 0$ ),  $\tilde{\pi}$  in (8) can also be seen as an approximation to the WME in (10) for small values of  $\pi_0$  and  $\pi$ . Namely, we can approximate  $(n - \bar{R}_{*1})/(n - \bar{R}_{11}) \approx 1$  and  $\pi_0(n - \bar{R}_{11}) \approx \pi_0 n$ . The MME has the advantage of being an unbiased estimator at the cost of being (slightly) less efficient than the WME (see [Section 3.4](#)).

In some cases, it might be that the information in  $\bar{R}_{00}$  in (9) (and  $\bar{R}_{10}$ ) is not easily available, for example, when additional data are collected using follow-up procedures. Although the MME can still be used in these instances, one can alternatively proceed with the marginalization of the conditional likelihood function and use a WME based on the latter, to obtain a Marginal WME (MWME). The MWME has generally no closed-form, but can however be easily computed using a numerical optimization method (see supplementary material A.7 for more details). In our simulation studies (not presented here), we found that the behavior of the MWME is comparable to the one of the WME.

In [Sections 3.4, 4, and 5.1](#), we compare, in terms of efficiency gain, the estimators that make use of the information obtained from the case count data, to the (weighted) infection survey proportion, that is, a Survey Maximum Likelihood Estimator (SMLE). It is based only on  $\bar{R}_{*1}$  (we recall  $\bar{R}_{*1} = \bar{R}_{11} + \bar{R}_{01}$ ), the number of positive cases in the infection survey. It is given by

$$\bar{\pi} = \frac{\bar{R}_{*1}/n - \alpha}{\Delta}, \quad (11)$$

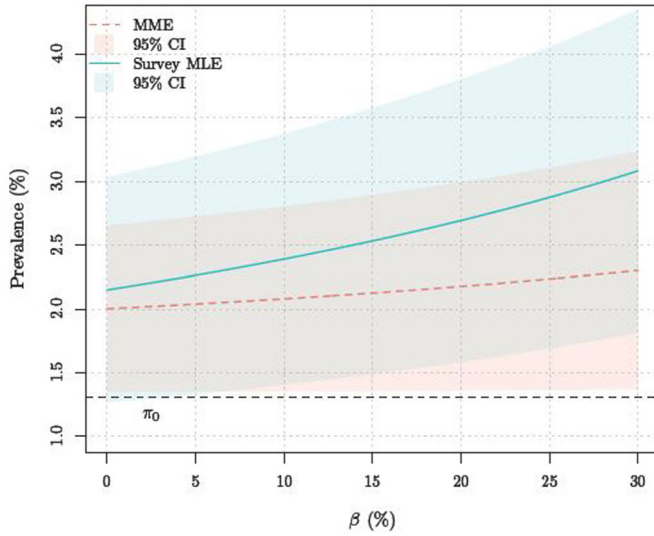
which reduces to  $\bar{\pi} = \bar{R}_{*1}/n$  when  $\alpha = \beta = 0$ .

*Remark 6.* Note that for the SMLE given in (11), an increase in the FP rate  $\beta$  induces a decrease in  $\Delta = 1 - (\alpha + \beta)$ , which directly induces an increased value for  $\bar{\pi}$ . However, the influence of an increase in  $\beta$  on the MME is milder. Indeed, we have

$$\begin{aligned}\frac{\partial(\bar{\pi} - \tilde{\pi})}{\partial\beta} &= \frac{\bar{R}_{11}/n - \alpha\pi_0}{\Delta^2} \\ &= \frac{\pi(1 - \alpha_0 - \beta_0)(1 + \alpha_0)}{\Delta} + \mathcal{O}_p(n^{-1/2}),\end{aligned}$$

which implies, by [Assumption 3](#), that this quantity is positive for sufficiently large  $n$ , hence, it explains the result found in the sensitivity analysis presented in [Figure 1](#) of [Section 4](#).

Finally, the MME, WME and MWME, as well as the SMLE, can be defined in a straightforward manner in the non-stratified sample cases, by setting  $\gamma_i = 1, \forall i = 1, \dots, n$ . However, in these cases, the properties of the estimators are different (see supplementary material A.1, A.3, and A.4 for the MME, WME, and MWME). In particular, for the MME and the SMLE with  $\gamma_i = 1, \forall i = 1, \dots, n$ , their finite sample distribution is known, and can be used to construct (exact, and thus possibly conservative) CI using the (fiducial) approach put forward in Clopper and Pearson (1934) (CP) (see also e.g., Fisher 1935; Brown, Cai, and DasGupta 2001). These are provided in supplementary material A.8.



**Figure 1.** Sensitivity analysis of the SMLE and MME prevalence estimators, with the (stratified) Austrian data (November 2020), as a function of the FN rate  $\beta$ . The FP rate of the PCR test is set to  $\alpha = 1\%$ .

### 3.4. Efficiency and Inference for Prevalence Estimators

As shown in supplementary material A, all the estimators presented previously are consistent and asymptotically normally distributed. Hence, the potential differences are due to their respective variances, especially the relative efficiency gain between estimators not ignoring the information obtained from the case count data, with respect to the one ignoring this information. Obviously, properly including additional information in the analysis should increase the resulting efficiency. The question then lies on the effective gain in efficiency, given that the needed additional information is quite easy to obtain while conducting the infection survey.

We can study analytically the efficiency gain of an estimator that takes into account the information obtained from the case count data, relative to the SMLE  $\bar{\pi}$  in (11). As shown in supplementary material A, for all considered estimators, the (asymptotic) variances are equal to the (asymptotic) variances of their counterpart in the non-stratified sampling case, up to a multiplicative constant given by  $V := \frac{1}{n} \sum_{i=1}^n \gamma_i^2$ . Hence, without loss of generality, we can study the relative efficiencies using the variances in the non-stratified sampling case. In this case, letting  $R_{jk} = \sum_{i=1}^n R_{ijk}$  for  $j, k \in \{0, 1\}$  and  $R_{*1} = R_{11} + R_{01}$ , where the  $R_{ijk}$  are provided in (3), we have, for example, that the SMLE (with  $\gamma_i = 1, \forall i = 1, \dots, n$ ) is given by  $\bar{\pi} = \frac{R_{*1}/n - \alpha}{\Delta}$ , which reduces to  $\bar{\pi} = R_{*1}/n$ , as expected, when  $\alpha = \beta = 0$ . Its variance is given by

$$\begin{aligned} \text{var}(\bar{\pi}) &= \frac{(\tau_{11}(\pi) + \tau_{01}(\pi))(1 - \tau_{11}(\pi) - \tau_{01}(\pi))}{n\Delta^2} \\ &= \frac{(\pi\Delta + \alpha)(1 - \pi\Delta - \alpha)}{n\Delta^2}, \end{aligned}$$

which, without measurement errors (i.e.,  $\alpha = \beta = 0$ ), reduces to  $\text{var}(\bar{\pi}) = \frac{1}{n}\pi(1 - \pi)$ , as expected. The variance of the MME  $\tilde{\pi}$  in (6) with  $\gamma_i = 1, \forall i = 1, \dots, n$ , is easily determined to be

$$\text{var}(\tilde{\pi}) = \frac{1}{\Delta^2(1 - \alpha_0)^2} \text{var}\left(\frac{R_{01}}{n}\right)$$

$$\begin{aligned} &= \frac{\tau_{01}(\pi)(1 - \tau_{01}(\pi))}{n\Delta^2(1 - \alpha_0)^2} \\ &= \frac{(\pi - \pi_0)(1 - (\pi - \pi_0))}{n\Delta^2(1 - \alpha_0)^2}, \end{aligned} \quad (12)$$

which, assuming  $\alpha_0 = 0$  (see Remark 2), reduces to  $\text{var}(\tilde{\pi}) = \frac{(\pi - \pi_0)(1 - (\pi - \pi_0))}{n\Delta^2}$ . Without measurement errors, it further reduces to  $\text{var}(\tilde{\pi}) = \frac{1}{n}(\pi - \pi_0)(1 - (\pi - \pi_0))$ .

In the case without measurement errors, the asymptotic variance of the WME with  $\gamma_i = 1, \forall i = 1, \dots, n$  (called the Conditional Maximum Likelihood Estimator in supplementary material A), using (S.6), is given by

$$\begin{aligned} \text{var}(\hat{\pi}) &= \frac{1}{n} \left( \frac{\tau_{01}(\pi)}{(\pi - \pi_0)^2} + \frac{\tau_{00}(\pi)}{(1 - \pi)^2} \right)^{-1} \\ &= \frac{1}{n} \frac{(1 - \pi)(\pi - \pi_0)}{(1 - \pi_0)}, \end{aligned}$$

so that the efficiency of the SMLE relative to the WME is given by

$$e(\bar{\pi}) = \frac{\text{var}(\hat{\pi})}{\text{var}(\bar{\pi})} = \frac{\pi - \pi_0}{\pi(1 - \pi_0)} < 1, \quad (13)$$

since  $\pi_0 \leq \pi < 1$ . If considering the ratio of the variance of the SMLE relative to the one of the MME, we obtain

$$\begin{aligned} \frac{\text{var}(\bar{\pi})}{\text{var}(\tilde{\pi})} &= \frac{\pi(1 - \pi)}{(\pi - \pi_0)(1 + \pi_0 - \pi)} \\ &= \frac{\pi(1 - \pi)}{\pi(1 - \pi) - \pi_0(1 + \pi_0 - 2\pi)}. \end{aligned} \quad (14)$$

Therefore, when  $2\pi > 1 + \pi_0$  we have  $\text{var}(\bar{\pi}) < \text{var}(\tilde{\pi})$ , while when  $2\pi < 1 + \pi_0$  we have  $\text{var}(\bar{\pi}) > \text{var}(\tilde{\pi})$ . A sufficient condition for the variance of the MME to be lower than the variance of the SMLE is, therefore, that the true population prevalence  $\pi$  is below  $1/2$ . Moreover, since the variance of the WME (in the non-stratified sampling case) attains the Cramer-Rao lower bound for the variance of any unbiased estimator of  $\pi$ , the MME, being unbiased, must have a higher variance. Indeed, the relative efficiency of  $\tilde{\pi}$  versus the WME (for sufficiently large  $n$ ) is  $e(\tilde{\pi}) = \frac{(1 - (\pi - \pi_0))(1 - \pi_0)}{1 - \pi} \leq 1$ , since  $\pi \geq \pi_0$ .

The efficiency loss of  $\bar{\pi}$  relative, for example, to the MME  $\tilde{\pi}$  in (14), is slightly different from (13), but can also be expressed in terms of the increase in infection survey sample size needed when using  $\bar{\pi}$  rather than  $\tilde{\pi}$ . Let  $n^*$  denote the infection survey sample size that is needed to obtain a variance for the SMLE that is equal to the one of the MME using an infection survey sample size of  $n$ . We obtain

$$\frac{n^*}{n} = \frac{1 - \pi_0}{1 - \pi_0/\pi},$$

which, for small  $\pi_0$ , is approximately equal to  $\frac{1}{1 - \pi_0/\pi}$ . If, for instance,  $\pi = 2\pi_0$  then  $\frac{n^*}{n} \approx 2$ . The added value in using the additional information provided in  $R_{11}$ , therefore, is equivalent to using the SMLE with a sample with twice the size.

In Section 5.1, we present a simulation study to assess the efficiency gain and the coverages of the CIs for  $\pi$  in a wider range of settings, in particular when introducing measurement errors. The results and conclusions also apply to the cases of stratified sampling.

### 3.5. Inference for the Rate of Asymptomatic Cases $\beta_0$

An estimator of the rate of asymptomatic cases  $\beta_0$  of the case count data is given by  $\hat{\beta}_0 = 1 - (\pi_0 - \alpha_0(1 - \hat{\pi}))/\hat{\pi}$ , and can be used to perform statistical inference for  $\beta_0$ , using the properties of the prevalence estimator. More details are provided in supplementary material B, and the CI for  $\beta_0$  are implemented in the R package `pempi`.

Assuming that the rate of asymptomatic cases  $\beta_0$  is independent of the case prevalence rate, its estimate obtained using infection survey data at some point in time, can be used to monitor the case prevalence rate from the case count data, since assuming  $\alpha_0 = 0$ , we have

$$\pi = \frac{\pi_0}{1 - \beta_0}. \tag{15}$$

It can also be used, for example, to estimate the ascertainment rate (see e.g., Gibbons et al. 2014), that is, the ratio of detected cases to the true number of cases given by  $\pi_0/\pi$ . Namely, assuming again  $\alpha_0 = 0$ , we have that

$$\frac{\pi_0}{\pi} = \frac{\pi(1 - \beta_0)}{\pi} = 1 - \beta_0. \tag{16}$$

Inference on the two quantities in (15) and (16) can be obtained using  $\hat{\beta}_0$  and its statistical properties, and for the updated case prevalence in (15), confidence intervals can be computed using the `pempi` R package.

### 4. COVID-19 Prevalence Estimation in Austria, November 2020

Making use of the data collected in Austria in November 2020 as described in Section 2 (see Kowarik et al. 2022, for more details), we propose different estimates of the COVID-19 prevalence  $\pi$ , the MME  $\tilde{\pi}$ , which takes into account the information obtained from the case count data, and the SMLE  $\bar{\pi}$  that is solely based on the infection survey data. We compare the estimates in different data settings, namely with or without measurement errors. The estimates are given with their respective CI at the 95% confidence level, to ease the comparisons and their potential effective differences.

The November 2020 infection survey consists of  $n = 2290$  participants who went through a testing procedure for COVID-19 using PCR tests. Seventy-two participants ( $R_{*1} = 72$ ) were tested positive, and among these ones, thirty-five ( $R_{11} = 35$ ) had been declared to have been tested positive in the case count data during the same month. An important information needed to compute the proposed estimators is the (known) case prevalence rate  $\pi_0$ . In November, there were 93,914 declared cases among the official (approximately) 7,166,167 inhabitants in Austria (above 16 years old), so that  $\pi_0 \approx 1.3105\%$ . The sensitivity  $(1 - \alpha)$  and the specificity  $(1 - \beta)$  of the PCR test are not known with precision, so that we present estimates of the prevalence  $\pi$  without measurement errors as well as for values for the FP and FN rates, that are plausible, given the data and according to the sensitivity and specificity reported, for example, in Kobokovich, West, and Gronvall (2020) or Surkova, Nikolayevskyy, and Drobniowski (2020). Moreover, as justified in Remark 2, we consider  $\alpha_0 = 0$ .

**Table 2.** Values for  $\bar{R}_{jk}$  with  $j, k \in \{0, 1\}$  in (5), with  $\frac{1}{n} \sum_{i=1}^n \gamma_i^2 = 1.51$ , and  $R_{jk}$  with  $j, k \in \{0, 1\}$  in (3), for the Austrian data, November 2020.

Case	$j = 0, k = 0$	$j = 0, k = 1$	$j = 1, k = 0$	$j = 1, k = 1$
$\bar{R}_{jk}$	2218.3698	38.2712	0	33.3589
$R_{jk}$	2218	37	0	35

**Table 3.** Prevalence estimation (Est.) from the Austrian data (November 2020), based on the  $\bar{R}_{jk}$  in Table 2, with associated 95% CI.

	$\alpha = \beta = 0$		$\alpha = 1\%, \beta = 10\%$	
	Est. (%)	95% CI (%)	Est. (%)	95% CI (%)
SMLE	3.128	(2.252–4.004)	2.391	(1.406–3.375)
MME	2.982	(2.336–3.627)	2.079	(1.354–2.804)
$\hat{\beta}_0$	56.0	(46.5–65.6)	37.0	(15.0–59.0)

NOTE: The first two columns are under the assumption of no measurement errors. The second two columns assume  $\alpha = 1\%$  and  $\beta = 10\%$ .

To compute the point estimates and associated CI for the prevalence, we use the `pempi` R package which requires, as input data, the infection survey sample size  $n$ , the values for  $\alpha$ ,  $\beta$  and  $\alpha_0$ ,  $\pi_0$ , and the sample values for  $\bar{R}_{jk}$ , see (5). For the Austrian data, the latter are given in Table 2, together with the sample values for  $R_{jk}$  computed using  $\gamma_i = 1, \forall i = 1, \dots, n$  for comparison. As one can see, while  $\frac{1}{n} \sum_{i=1}^n \gamma_i^2 = 1.51$ , compared to one in the non-stratified case, the differences in the sufficient statistics are very small, and have no significant impact on the prevalence estimation (results not shown here). This indicates that the sensitivity of the prevalence estimator to the sampling weights is rather mild, so that even if the calibration for nonresponse does not totally remove the sampling bias, its impact on the final estimates is limited.

Table 3 provides various estimates of  $\pi$ , the COVID-19 prevalence in Austria in November 2020, and of the rate of asymptomatic cases  $\beta_0$  of the case count data, based on the  $\bar{R}_{jk}$  in Table 2. These are computed with or without measurement errors, with, for the former,  $\alpha = 1\%$  (FP rate for the PCR test in the infection survey) and  $\beta = 10\%$  (FN rate for the PCR test in the infection survey). We choose a small  $\alpha$ , because we only observe 71 positive cases out of 2287 participants. If  $\alpha$  were larger, say  $\alpha = 5\%$ , we would also expect a larger number of positive cases, that is, 114 positive cases just because of false positives. The prevalence estimates are given with their respective CI at the 95% confidence level.

From Table 3, one can derive a series of insights. First, we note that, without measurement errors, the estimates are larger than with measurement errors, although the differences still remain in the CI bounds. Second, and most importantly, by comparing the CI lengths, one can observe that they are substantially smaller for the MME which takes into account the information obtained from the case count data. Actually, the (estimated) variance of the latter, with this dataset, is about half of the one of the SMLE that ignores the information obtained from the case count data. Third, although the point estimates are rather close when comparing the MME and the SMLE without measurement errors, one can observe that their difference is larger with measurement errors.

The measurement error (the FN rate  $\beta$ ) has more impact on the prevalence estimator when ignoring the information obtained from the case count data. To illustrate this point, and



since the FP rate  $\alpha$  has a limited range of possible values, given the data, we present in [Figure 1](#) a sensitivity analysis of the prevalence estimation by the SMLE and the MME (stratified sampling case), when the FN rate  $\beta$  varies from 0% to 30%. It is clear that the SMLE is much more influenced by the value of the FN rate  $\beta$  compared to the MME, which shows a far better stability. In [Remark 6](#), we provide an explanation for this result.

Finally, the estimated rate of asymptomatic cases  $\hat{\beta}_0$ , reported in [Table 3](#), indicates that, supposing the PCR test does not produce any FN, 55% of the positive cases were not reported in the case count data (at that time). If one considers some measurement error for the PCR test, this rate is approximately 35%, which is lower, as expected. This result is in line, for example and among others, with [Li et al. \(2021\)](#) who reported a fraction of “undocumented infections” in Henan Province of 53% with CI of (50–68), during January and February 2020.

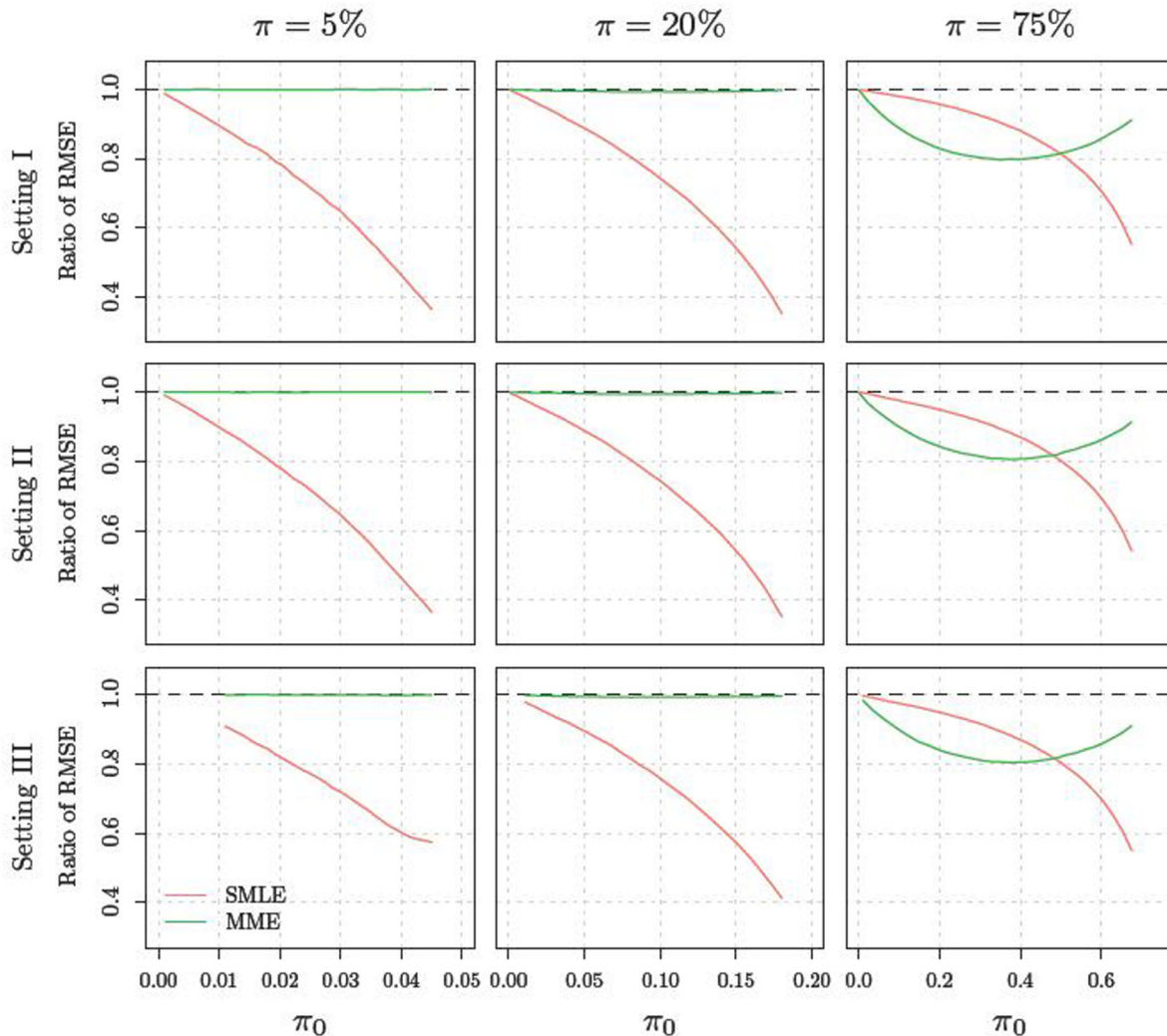
## 5. Simulation Studies

In this section, we present the results of two simulation studies, one on the relative efficiencies and CI coverage for  $\pi$ , for the different estimators, and one to assess the effect on the Root

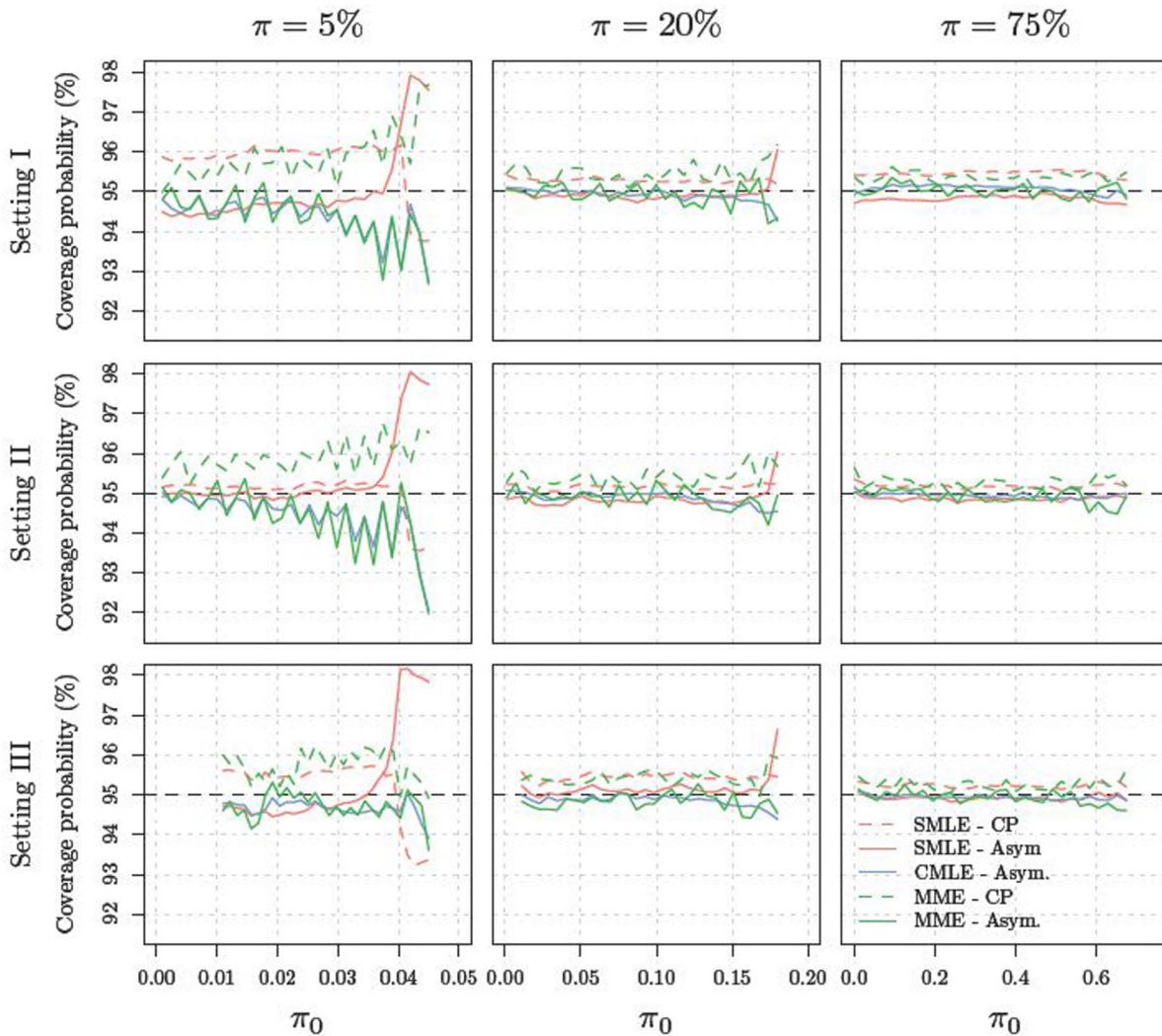
Mean Squared Error (RMSE) and CI coverage for  $\pi$  of violations of [Assumption 1](#).

### 5.1. Simulation Study for Relative Efficiencies and Coverage

In order to gain more insights on the properties of the different estimators, we perform a simulation study to evaluate, in finite samples, the efficiencies, CI coverage for  $\pi$  and CI lengths of the different methods. Without loss of generality, the simulation study is based on the estimators for the non-stratified sampling case, since the (asymptotic) variance is proportional to the stratified sampling case, up to a constant (see also [Section 3.4](#)). Throughout, we choose  $\alpha_0 = 0$  (see [Remark 2](#)). We consider three settings. Setting I is without measurement errors, that is, with  $\alpha = \beta = 0$ . Setting II has only a FN rate, that is,  $\alpha = 0$ ,  $\beta = 2\%$ , and Setting III, finally, has both types of measurement errors, that is,  $\alpha = 1\%$ ,  $\beta = 2\%$ . We consider an infection survey sample size of  $n = 2000$  which leads to the same conclusions (not presented here) as a somewhat smaller infection survey sample size (e.g.,  $n = 1500$ ). For the (true) prevalence  $\pi$ , we consider three rather different values, that is, 5%, 20%, and 75%



**Figure 2.** Relative efficiencies, as measured by the relative empirical RMSE, for the MME  $\tilde{\pi}$  (green) and the SMLE  $\bar{\pi}$  (red) relative to the CMLE  $\hat{\pi}$ . First row with no measurement error, middle row with a positive FN rate ( $\alpha = 0$ ,  $\beta = 2\%$ ), bottom row with both types of measurement errors ( $\alpha = 1\%$ ,  $\beta = 2\%$ ). The infection survey sample size is  $n = 2000$  and the number of Monte Carlo simulations is 50,000.



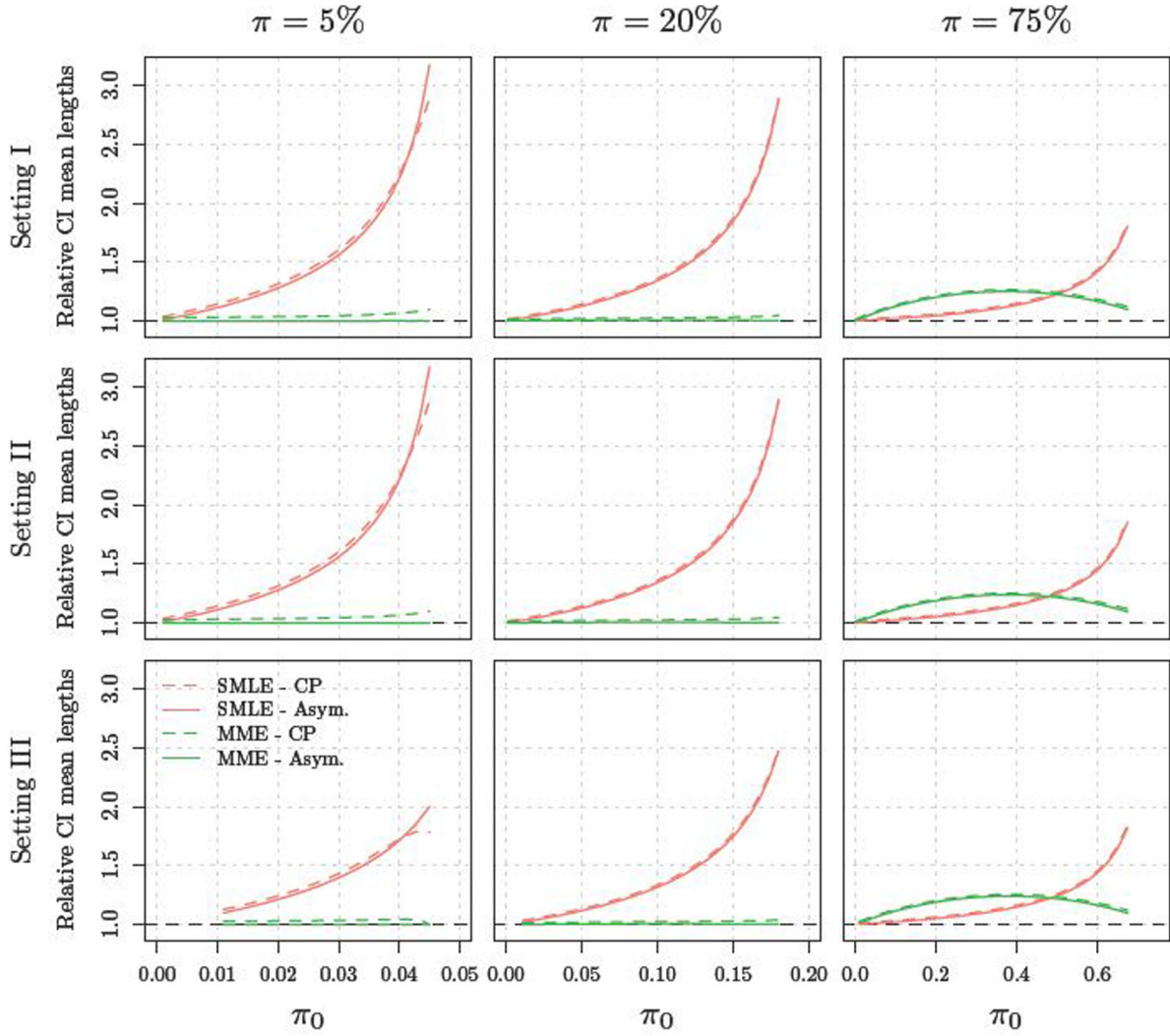
**Figure 3.** Empirical coverage of the CIs for  $\pi$  (at the 95% level) as a function of  $\pi_0$  for (true) prevalence  $\pi = 5\%$ ,  $10\%$ ,  $75\%$ , of CI based on the CP method (CP) and asymptotic variance (Asym) with the SMLE and the MME, and based on the asymptotic variance with the CMLE. Setting I:  $\alpha = \beta = 0$ . Setting II:  $\alpha = 0, \beta = 2\%$ . Setting III:  $\alpha = 1\%, \beta = 2\%$ . The infection survey sample size is 2000 and the number of Monte Carlo simulations is 50,000.

in order to cover a wide range of possible prevalence rates. For  $\pi_0$ , we consider, for each value of  $\pi$ , 30 equally spaced values between 0 and  $0.975\pi$ , so that, conditionally on the information brought in by  $Z_i$ , one can appreciate the efficiency and accuracy gain when including the information from the case count data. As estimators, we consider the SMLE  $\bar{\pi}$ , the Conditional MLE (CMLE) (see supplementary material A.1)  $\hat{\pi}$ , which is actually the WME when  $\gamma_i = 1, \forall i = 1, \dots, n$ , as well as the MME  $\tilde{\pi}$ .

Figure 2 presents the relative efficiencies, as measured by the relative empirical Root Mean Squared Error (RMSE), for the MME  $\tilde{\pi}$  and the SMLE  $\bar{\pi}$  relative to the CMLE  $\hat{\pi}$ . The main messages are the following. First, there is a substantial efficiency loss for the SMLE  $\bar{\pi}$  that increases drastically as  $\pi_0$  approaches  $\pi$ , with or without measurement errors. This is in line with the fact that the information brought in by considering  $Z_i$  obtained from the case count data, is more important as  $\pi_0$  is near  $\pi$ , and ignoring it, lowers the efficiency of the SMLE. Second, for the MME, the efficiency loss is negligible for  $\pi = 5\%$  and  $\pi = 20\%$  when  $\pi_0$  is not too near to  $\pi$ , while the efficiency loss is rather important for small values of  $\pi_0$  (relative to  $\pi$ ), compared to the one of the SMLE when  $\pi = 75\%$ .

Figure 3 presents the empirical coverage of the CIs for  $\pi$  (at the 95% level), computed using simulations, for different values of  $\pi_0$  and  $\pi$ , of CI based on the CP method and asymptotic variance (asymptotic method) with the SMLE and the MME, and based on the asymptotic variance with the CMLE. Overall, as expected, the CP method provides slightly conservative coverage of the CIs for  $\pi$  across settings, while the asymptotic method based on the SMLE is slightly liberal, especially for  $\pi = 5\%$ . Moreover, for both the CP method based on the SMLE and the asymptotic method based on the CMLE, for  $\pi = 5\%$  and  $\pi = 20\%$ , the coverage of the CIs for  $\pi$  worsens (even if they remain quite accurate) as  $\pi_0$  approaches  $\pi$ . For the asymptotic method, this can be explained by the fact that CIs might have bounds falling outside the domain of  $\pi$  (e.g., below  $\pi_0$ ), especially when  $\pi$  is near  $\pi_0$  and in settings such as Setting II.

Given that the coverage of the CIs for  $\pi$  is reasonable across methods, it is worth comparing the CI lengths. Figure 4 presents the relative CI (at the 95% level) lengths, computed using simulations, for the CP method based on  $R_{*1}$  in (3) (associated to the SMLE  $\bar{\pi}$ ) and the CP method based on  $R_{01}$  in (3) (associated to the MME  $\tilde{\pi}$ ), relative to the CI (at the 95% level) lengths



**Figure 4.** Relative empirical CI (at the 95% level) mean lengths as a function of  $\pi_0$  for (true) prevalence  $\pi = 5\%$ ,  $10\%$ ,  $75\%$ , of CI based on the CP (CP) method and asymptotic variance (Asym) with the SMLE and the MME, relative to the empirical CI (at the 95% level) based on the asymptotic variance with the Conditional MLE (CMLE). Setting I:  $\alpha_0 = \alpha = \beta = 0$ . Setting II:  $\alpha_0 = \alpha = 0, \beta = 2\%$ . Setting III:  $\alpha_0 = 0, \alpha = 1\%, \beta = 2\%$ . The infection survey sample size is 2000 and the number of Monte Carlo simulations is 50,000.

for the asymptotic method based on the CMLE  $\hat{\pi}$ . One can observe, as expected, that the (mean) CI lengths can be a lot larger when ignoring the information provided by  $Z_i$  from the case count data, especially as the information increases, that is, as  $\pi_0$  approaches  $\pi$ . An interesting feature appears, however, for a small population prevalence ( $\pi = 5\%$ ) when  $\pi_0$  approaches  $\pi$ , in that the mean CI length for the CP based on  $R_{01}$  (associated to the MME) is smaller than the one of the asymptotic method based on the CMLE. However, for a large population prevalence ( $\pi = 75\%$ ), the mean CI length for the CP based on  $R_{*1}$  are relatively smaller than the ones based on  $R_{01}$ , while remaining larger than the mean CI length for the asymptotic method based on the CMLE. This is especially the case for small values of  $\pi_0$  relative to  $\pi$ , and is in line with the study of the efficiencies provided in Figure 2.

## 5.2. Simulation Study to Assess Violations of Assumption 1

In this section, we conduct a simulation study by generating samples which violates Assumption 1, or more precisely, which

violates its direct implication, namely that  $\mathbb{P}(Z = 1|X = 1, Y = 1) = \mathbb{P}(Z = 1|X = 1)$  and  $\mathbb{P}(Z = 1|X = 0, Y = 1) = \mathbb{P}(Z = 1|X = 0)$ .

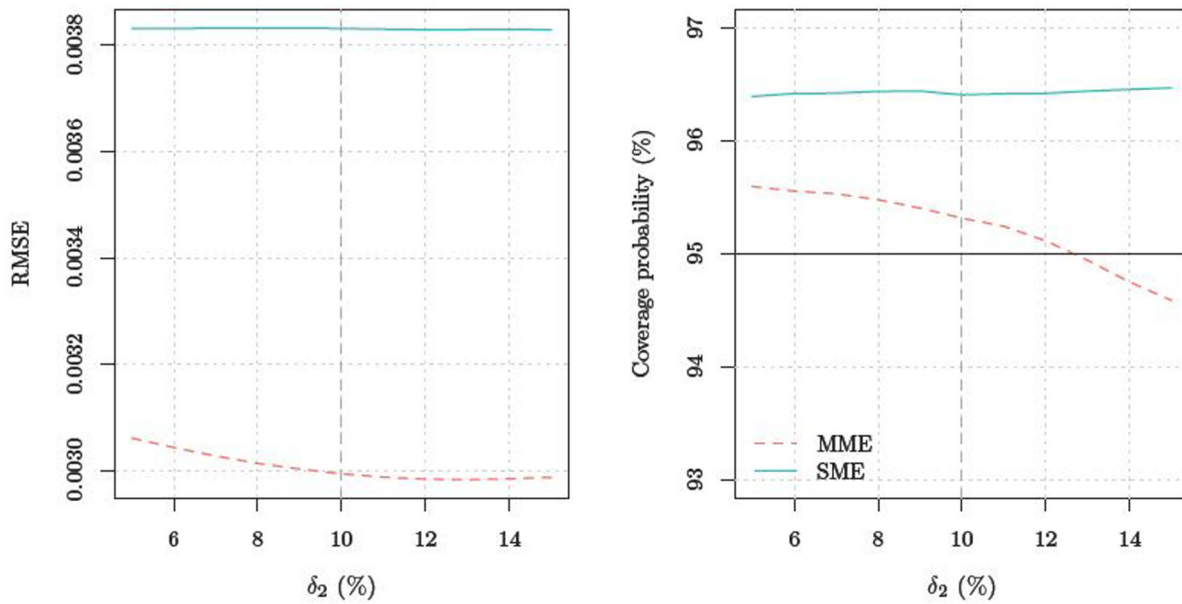
Recall that  $\mathbb{P}(X_i = 1) = \pi$ ,  $\mathbb{P}(Z_i = 1|X_i = 1) = 1 - \beta_0$ ,  $\mathbb{P}(Z_i = 1|X_i = 0) = \alpha_0$  and  $\mathbb{P}(Y_i = 1|X_i = 0) = \alpha$ . We now suppose that, in violation of Assumption 1,  $\mathbb{P}(Y_i = 1|X_i = Z_i = 1) = 1 - \delta_1$  (and  $\mathbb{P}(Y_i = 0|X_i = Z_i = 1) = \delta_1$ ) and  $\mathbb{P}(Y_i = 1|X_i = 1, Z_i = 0) = 1 - \delta_2$  (and  $\mathbb{P}(Y_i = 0|X_i = 1, Z_i = 0) = \delta_2$ ) with  $\delta_1$  possibly different from  $\delta_2$ . We then obtain

$$\begin{aligned} 1 - \beta &= \mathbb{P}(Y_i = 1|X_i = 1) \\ &= \mathbb{P}(Z_i = 1|X_i = 1) \mathbb{P}(Y_i = 1|X_i = Z_i = 1) \\ &\quad + \mathbb{P}(Z_i = 0|X_i = 1) \mathbb{P}(Y_i = 1|X_i = 1, Z_i = 0) \\ &= (1 - \beta_0)(1 - \delta_1) + \beta_0(1 - \delta_2), \end{aligned}$$

and therefore, we have

$$\delta_1 = 1 - \frac{(1 - \beta) - \beta_0(1 - \delta_2)}{1 - \beta_0}. \quad (17)$$

Note that Assumption 1 is violated for values of  $\delta_1 \neq \beta$  and  $\delta_2 \neq \beta$ .



**Figure 5.** Sensitivity analysis of the MME prevalence estimators to violations of Assumption 1, compared to the same analysis for the SMLE which is not affected by violations of Assumption 1. The sensitivity is measured by the RMSE (left panel) and associated coverage for  $\pi$  (right panel), as a function of  $\delta_2 = \mathbb{P}(Y_i = 0|X_i = 1, Z_i = 0)$ . The model’s parameters values are the ones obtained from the analysis of the Austrian data, November 2020. The vertical line at  $\delta_2 = \beta = 10\%$  corresponds to the case of no violation of Assumption B. The probability  $\delta_2$  varies in a grid of 20 evenly spaced values going from 5% to 15%, corresponding to values of  $\delta_1 = \mathbb{P}(Y_i = 0|X_i = Z_i = 1)$  ranging from 7.1% to 13%.

We can study the effect on the RMSE of the prevalence estimators as well as their associated coverage for  $\pi$ , for different values of, say,  $\delta_2$ , which imply different values for  $\delta_1$ . We chose  $\beta = 10\%$  and using (17), with  $\beta_0 = 36.977\%$  (obtained from the Austrian data), a grid of 20 evenly spaced values going from 5% to 15% for  $\delta_2$  corresponds to a range from 7.1% to 13% for  $\delta_1$ , which represents a non-negligible dependence. From the Austrian data, we also have  $n = 2290$ ,  $\tilde{\pi} = 2.079\%$ ,  $\pi_0 = 1.311\%$ , and we set  $\alpha = 1\%$  and  $\alpha_0 = 0\%$ . We consider  $B = 10^5$  Monte Carlo replications of simulated data using the values for the different parameters as set above, and we compute the RMSE of the MME and the SMLE for comparison since it is not affected by the violation of Assumption 1, as well as their associated coverage for the population prevalence (i.e.,  $\pi = 2.079\%$ ). The simulation results are presented in Figure 5. The conclusions are as follows: for the considered ranges of the parameters  $\delta_1$  and  $\delta_2$ , the RMSE of the MME and its associated coverage for  $\pi$ , are quite stable in that the relative differences with the RMSE and associated coverage for  $\pi$ , using the SMLE (which is not affected by violations of Assumption 1) remain quite stable. It is therefore reasonable to make use of Assumption 1, which allows to construct estimators in a more concise manner.

### 6. Discussion

We propose a method to estimate the prevalence of a disease based on information both from a survey sample and from case count data. We show that our approach provides estimates that are substantially more accurate than the simple sample proportion of participants in the survey sample. As an important consequence, our approach can provide a given level of desired accuracy, with a substantially smaller survey sample size. For the case of the November 2020 survey of Austrian COVID-19

cases (Kowarik et al. 2022), using our approach one could have achieved the same level of accuracy that the sample proportion achieves with roughly half the data. This is useful when data collection is costly or when medical tests (or lab spaces to evaluate tests) are in limited supply. We study this problem with and without the possibility of measurement errors for both the case count data and infection survey. Additionally, we also find that estimators that use the information provided by the case count data, are far less sensitive to the value of the FN rate of the testing device, which limits the estimation bias of potential incorrect choices for the latter.

While we have cast this article in the language of prevalence estimation, the method we propose has a more general range of applications. It can be applied whenever we are interested in estimating the proportion of a characteristic A in a population, when there is a characteristic B that can be seen as a possibly fairly imprecise indicator of characteristic A, but with the advantage that the proportion of characteristic B in the population is known.

To give but one other example, consider a case in production quality control, where a cheap test (that is not very accurate) is routinely used to assess if a product (participant) is faulty or not (this is characteristic B), producing the proportion  $\pi_0$  of products declared faulty. To accurately estimate the proportion  $\pi$  of faulty products, one can use a survey using an expensive test on a limited number of products chosen randomly (to measure characteristic A), possibly producing FP (with known probability  $\alpha$ ) and FN (with known probability  $\beta$ ). For better accuracy of estimation of the proportion  $\pi$ , it is then advisable to also measure the quality of the products in the survey with the cheap test (i.e., characteristic B), which allows the computation of the sufficient statistics in (3). If  $\alpha_0$  (the FP of the cheap test), now not necessarily close to zero, is known, one can use our approach as it is developed here. If  $\alpha_0$  is unknown, one

can easily adapt our approach to estimate  $\alpha_0$  along with the proportion  $\pi$ .

Other adjustments are also necessary with binary outcomes in logistic regression (see e.g., Ni et al. 2019; Meyer and Mittag 2017, and the references therein). Our framework could be extended to the case of logistic regression, but this extension is left for further research. Moreover, while the data from the November 2020 infection survey collected by Statistics Austria (2020) is suitable for prevalence estimation, the same approach can be used to estimate other proportions such as the incidence of the COVID-19 (see e.g., Woodward 2014).

Finally, all computations presented in this article were done using the `pempi` R package that is directly available on the Comprehensive R Archive Network (CRAN), and can be installed using `install.packages("pempi")`. As previously mentioned, all simulation results and the data analysis, can be reproduced using the associated functions in the `pempi` package.

## Appendices

### Appendix A: Discussion of Assumptions 1, 2, and 3

For the case of  $X_i = 0$ , [Assumption 1](#) assumes the stochastic independence of FP results. This is empirically justified for two reasons: one, FP rates for PCR tests are very low, see also [Remark 2](#), and two, FP results of PCR tests are mostly contamination problems that are independent of the tested participant (see e.g., Braunstein et al. 2021).

For the case of  $X_i = 1$ , [Assumption 1](#) assumes the stochastic independence of FN results, the main source of which is a low viral load (see e.g., Kanji et al. 2021). For COVID-19 infected participants whose viral load increases during the few days between the two tests, [Assumption 1](#) is then valid. For COVID-19 infected participants, whose viral load remains low throughout their infection, [Assumption 1](#) is valid if one associates such PCR undetectable infected participants as being not infected. In other words,  $X_i$  should be defined as ‘‘PCR detectable’’ infected participant, and in this case, the results would need to be interpreted accordingly.

Having said that, if [Assumption 1](#) would be invalid, it would, presumably, be so by a relatively low level of dependence. Therefore, in [Section 5.2](#), we present a simulation study with data generated under violations of this assumption, and the conclusion is, for a reasonable range of deviations from [Assumption 1](#), that the proposed estimator and associated CI remain very stable.

With [Assumption 2](#), we rule out the uninteresting case  $\alpha + \beta = 1$ . If  $\alpha + \beta = 1$ ,  $Y_i$  is completely uninformative about the random variable of interest  $X_i$ , as  $\mathbb{P}(X_i = 1|Y_i = 1) = \mathbb{P}(X_i = 1|Y_i = 0) = \pi$ . Indeed, by Bayes’ law

$$\begin{aligned} \mathbb{P}(X_i = 1|Y_i = 1) &= \frac{\mathbb{P}(X_i = 1)\mathbb{P}(Y_i = 1|X_i = 1)}{\mathbb{P}(X_i = 1)\mathbb{P}(Y_i = 1|X_i = 1) + \mathbb{P}(X_i = 0)\mathbb{P}(Y_i = 1|X_i = 0)}, \end{aligned}$$

which is equal to

$$\mathbb{P}(X_i = 1|Y_i = 1) = \frac{\pi(1 - \beta)}{\pi(1 - \beta) + (1 - \pi)\alpha}.$$

If  $\alpha + \beta = 1$ , we have  $1 - \beta = \alpha$ . We can then divide both numerator and denominator by  $\alpha$  and get  $\mathbb{P}(X_i = 1|Y_i = 1) = \frac{\pi}{\pi + (1 - \pi)} = \pi$ . Similarly,

$$\mathbb{P}(X_i = 1|Y_i = 0)$$

$$= \frac{\mathbb{P}(X_i = 1)\mathbb{P}(Y_i = 0|X_i = 1)}{\mathbb{P}(X_i = 1)\mathbb{P}(Y_i = 0|X_i = 1) + \mathbb{P}(X_i = 0)\mathbb{P}(Y_i = 0|X_i = 0)},$$

which is equal to

$$\mathbb{P}(X_i = 1|Y_i = 0) = \frac{\pi\beta}{\pi\beta + (1 - \pi)(1 - \alpha)}.$$

If  $\alpha + \beta = 1$ , we have  $1 - \alpha = \beta$ . We can divide both numerator and denominator by  $\beta$  and get  $\mathbb{P}(X_i = 1|Y_i = 1) = \frac{\pi}{\pi + (1 - \pi)} = \pi$ . Otherwise, [Assumption 2](#) is without loss of generality in the following sense. If  $\alpha + \beta > 1$ , we could just use  $Y'_i = 1 - Y_i$  instead of  $Y_i$ , which would have FP and FN rates of  $\alpha' = 1 - \alpha$  and  $\beta' = 1 - \beta$ , with  $\alpha' + \beta' < 1$ .

[Assumption 3](#) is similarly without loss of generality. It also implies that  $\alpha_0 \leq \pi_0$ . To see this suppose that  $\alpha_0 > \pi_0 = (1 - \pi)\alpha_0 + \pi(1 - \beta_0)$ . This is equivalent to  $0 > -\pi\alpha_0 + \pi(1 - \beta_0)$ , which in turn, is equivalent to  $0 > 1 - \alpha_0 - \beta_0$ , a contradiction.

### Appendix B: Measurement Errors in the Case Count Data

In order to understand the role played by  $\alpha_0$ , we introduce, for each participant  $i = 1, \dots, n$  in the infection survey, two additional unobserved random variables of interest, namely

$$W_{1i} := \begin{cases} 1 & \text{if participant } i \text{ has been tested} \\ & \text{for the case count data,} \\ 0 & \text{otherwise,} \end{cases} \quad (\text{B1})$$

and

$$W_{2i} := \begin{cases} 1 & \text{if participant } i \text{ who has been tested for the case} \\ & \text{count data has tested positive using the PCR} \\ & \text{test for the case count data,} \\ 0 & \text{otherwise.} \end{cases}$$

Therefore, we can express the random variable  $Z_i$  in (1) as

$$Z_i := \begin{cases} 1 & \text{if } W_{1i} = 1 \text{ and } W_{2i} = 1, \\ 0 & \text{otherwise.} \end{cases}$$

Hence, the variable  $Z_i$  separates the participants that were tested for the case count data and had a positive PCR test there ( $Z_i = 1$ ), and the participants that were either not tested at all for the case count data or were tested with the PCR test there but that test was negative ( $Z_i = 0$ ). Then, using the definition of  $W_{1i}$  in (B1), we have

$$\begin{aligned} \alpha_0 &:= \mathbb{P}(Z_i = 1|X_i = 0) = \mathbb{P}(W_{1i} = 1, W_{2i} = 1|X_i = 0) \\ &= \mathbb{P}(W_{2i} = 1|W_{1i} = 1, X_i = 0)\mathbb{P}(W_{1i} = 1|X_i = 0), \end{aligned}$$

which is the product of the probabilities of participant  $i$ , that is actually uninfected ( $X_i = 0$ ), is tested positive for the case count data ( $W_{2i} = 1$ ) and is selected for the case count data ( $W_{1i} = 1$ ). This probability,  $\alpha_0$ , will, therefore, typically be much smaller than the probability of a FP due to the PCR test used for the case count data.

To obtain  $\beta_0$ , we use

$$\begin{aligned} \pi_0 &= \mathbb{P}(Z_i = 1) \\ &= \mathbb{P}(X_i = 1)\mathbb{P}(Z_i = 1|X_i = 1) + \mathbb{P}(X_i = 0)\mathbb{P}(Z_i = 1|X_i = 0) \\ &= \pi(1 - \beta_0) + (1 - \pi)\alpha_0 \end{aligned}$$

to deduce (2).

### Appendix C: Success Probabilities

Recall that the objective is to provide an estimator for the unknown population proportion, that is, the prevalence, given by

$$\pi := \mathbb{P}(X_i = 1).$$

We also consider the information given by  $\pi_0 := \mathbb{P}(Z_i = 1)$ , the case prevalence rate.

The success probabilities  $\tau_{ij}(\pi)$  for  $R_{ij}$  with  $i, j \in \{0, 1\}$ , in (3), can be deduced from the following table:

$X_i$	$prob$	$\mathbb{P}(Y_i = 1)$	$\mathbb{P}(Z_i = 1)$
1	$\pi$	$1 - \beta$	$1 - \beta_0$
0	$1 - \pi$	$\alpha$	$\alpha_0$

with  $\beta_0$  given in (2). There are two fundamental cases  $X_i = 0$  and  $X_i = 1$ , and conditionally on each of these cases, errors are independently and identically distributed by Assumption 1. We, thus, have

$$\begin{aligned} \tau_{11} &= \mathbb{P}(X_i = 1)\mathbb{P}(Y_i = 1|X_i = 1)\mathbb{P}(Z_i = 1|X_i = 1) + \\ &\quad \mathbb{P}(X_i = 0)\mathbb{P}(Y_i = 1|X_i = 0)\mathbb{P}(Z_i = 1|X_i = 0) \\ &= \pi(1 - \beta)(1 - \beta_0) + (1 - \pi)\alpha\alpha_0. \end{aligned}$$

Plugging in  $\beta_0$  and using  $\Delta := 1 - (\alpha + \beta)$  we obtain

$$\tau_{11}(\pi) = \pi \Delta \alpha_0 + (\pi_0 - \alpha_0)(1 - \beta) + \alpha \alpha_0.$$

The remaining probabilities  $\tau_{10}$ ,  $\tau_{01}$ , and  $\tau_{00}$  can be similarly obtained.

### Supplementary Materials

The supplementary materials provides the mathematical derivation and statistical properties of the different estimators mentioned in the main text, that take into account the information provided in the case count data. Both the non-stratified and stratified sampling case are presented. The statistical properties concern consistency or unbiasedness as well as finite sample or asymptotic distributions. Using these later results, expressions for the CIs are provided, which, in some cases, can be exact, using the CP method. The estimators are the Conditional Maximum Likelihood Estimator (CMLE), its weighted version the Weighted M-Estimator (WME), the Method of Moments Estimator (MME) and its weighted version, and the Marginal Maximum Likelihood Estimator (MMLE) for missing information and its weighted version the MWME. The inferential properties of the rate of asymptomatic cases  $\beta_0$  are also provided.

### Acknowledgments

All authors are grateful to the editor, associated editor and two anonymous referees, as well as to Dominique-Laurent Couturier, Michael Greinecker, Helmut Kuzmics, Hans Manner, Michael Richter, Michael Scholz and Yuming Zhang for helpful comments and suggestions.

### Disclosure Statement

The authors report there are no competing interests to declare.

### Funding

Stephane Guerrier gratefully acknowledges the financial support of the Swiss National Science Foundation grants #176843 and #211007 as well as the Innosuisse grants #37308.1 IP-ENG and #53622.1 IP-ENG. L. Maria-Pia Victoria-Feser gratefully acknowledges the financial support of the Swiss National Science Foundation grant #182684.

### References

Accorsi, E., Qiu, X., Rumpler, E., Kennedy-Shaffer, L., Kahn, R., Joshi, K., Goldstein, E., Stensrud, M., Niehus, R., Cevik, M., and Lipsitch, M. (2021), “How to Detect and Reduce Potential Sources of Biases in Studies of SARS-COV-2 and COVID-19,” *European Journal of Epidemiology*, 36, 179–196. [1]

Bethlehem, J., and Schouten, B. (2017), “Nonresponse Error: Detection and Correction,” in *The SAGE Handbook of Survey Methodology*, eds. C. Wolf, D. Joye, T. W. Smith, and Y.-C. Fu, London: SAGE Publications Ltd. [4]

Braunstein, G. D., Schwartz, L., Hymel, P., and Fielding, J. (2021), “False Positive Results with SARS-CoV-2 RT-PCR Tests and How to Evaluate a RT-PCR-Positive Test for the Possibility of a False Positive Result,” *Journal of Occupational and Environmental Medicine*, 63, e159. [12]

Brown, L. D., Cai, T., and DasGupta, A. (2001), “Interval Estimation for a Binomial Proportion,” *Statistical Science*, 16, 101–133. [5]

Chen, Y., Li, P., and Wu, C. (2020), “Doubly Robust Inference with Nonprobability Survey Samples,” *Journal of the American Statistical Association*, 115, 2011–2021. [1]

Clopper, C. J., and Pearson, E. S. (1934), “The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial,” *Biometrika*, 26, 404–413. [5]

Dempsey, W. (2023), “Statistical Paradoxes in Coronavirus Case-Counts: Selection Bias, Measurement Error, and the COVID-19 Pandemic,” *Annals of Applied Statistics*, to appear. [1]

Diggle, P. J. (2011), “Estimating Prevalence Using an Imperfect Test,” *Epidemiology Research International*, 2011, 608719. [2]

Elliott, M., and Valliant, R. (2017), “Inference for Nonprobability Samples,” *Statistical Science*, 32, 249–264. [1]

Fisher, R. A. (1935), “The Fiducial Argument in Statistical Inference,” *Annals of Eugenics*, 6, 391–398. [5]

Gibbons, C. L., Mangen, M.-J. J., Plass, D., Havelaar, A. H., Brooke, R. J., Kramarz, P., Peterson, K. L., Stuurman, A. L., Cassini, A., Fèvre, E. M., and Kretzschmar, M. E. E. (2014), “Measuring Underreporting and Underascertainment in Infectious Disease Datasets: A Comparison of Methods,” *BMC Public Health*, 14, 147. [2,7]

Kahn, R., Kennedy-Shaffer, L., Grad, Y., Robins, J., and Lipsitch, M. (2021), “Potential Biases Arising from Epidemic Dynamics in Observational Seroprotection Studies,” *American Journal of Epidemiology*, 192, 328–335. [1]

Kanji, J. N., Zelyas, N., MacDonald, C., Pabbaraju, K., Khan, M. N., Prasad, A., Hu, J., Diggle, M., Berenger, B. M., and Tipples, G. (2021), “False Negative Rate of COVID-19 PCR Testing: A Discordant Testing Analysis,” *Virology Journal*, 18, 1–6. [12]

Kobokovich, A., West, R., and Gronvall, G. (2020), “Serology-based Tests for COVID-19,” Technical Report, Center for Health Security, Bloomberg School of Public Health, John Hopkins University. [2,7]

Kowarik, A., Paskvan, M., Weinauer, M., Till, M., Schrittwieser, K., Göllner, T., Hartleib, S., Klimont, J., Plate, M., Baumgartner, I., Edelhofer-Lielacher, E., Grasser, A., Kytir, J., Weseslindtner, L., and Strassl, R. (2022), “Assessing Coronavirus SARS-CoV-2 Prevalence in Austria with Sample Surveys in 2020,” *Austrian Journal of Statistics*, 51, 27–44. [2,3,4,7,11]

Lewis, F. I., and Torgerson, P. R. (2012), “A Tutorial in Estimating the Prevalence of Disease in Humans and Animals in the Absence of a Gold Standard Diagnostic,” *Emerging Themes in Epidemiology*, 9, 9. [2]

Li, C., Zhu, Y., Qi, C., Liu, L., Zhang, D., Wang, X., She, K., Jia, Y., Liu, T., He, D., Xiong, M., and Li, X. (2021), “Estimating the Prevalence of Asymptomatic COVID-19 Cases and their Contribution in Transmission - Using Henan Province, China, as an Example,” *Frontiers in Medicine*, 8, 773. [8]

Manski, C. F., and Molinari, F. (2021), “Estimating the COVID-19 Infection Rate: Anatomy of an Inference Problem,” *Journal of Econometrics*, 220, 181–192. [1]

Meyer, B. D., and Mittag, N. (2017), “Misclassification in Binary Choice Models,” *Journal of Econometrics*, 200, 295–311. [12]

- Mizumoto, K., Kagaya, K., Zarebski, A., and Chowell, G. (2020), “Estimating the Asymptomatic Proportion of Coronavirus Disease 2019 (COVID-19) Cases on board the Diamond Princess Cruise Ship,” *Euro Surveillance*, 25, 2000180. [1]
- Munster, V. J., Koopmans, M., van Doremalen, N., van Riel, D., and de Wit, E. (2020), “A Novel Coronavirus Emerging in China — Key Questions for Impact Assessment,” *New England Journal of Medicine*, 382, 692–694. [2]
- Ni, J., Dasgupta, K., Kahn, S. R., Talbot, D., Lefebvre, G., Lix, L. M., Berry, G., Burman, M., Dimentberg, R., Laflamme, Y., Cirkovic, A., and Rahme, E. (2019), “Comparing External and Internal Validation Methods in Correcting Outcome Misclassification Bias in Logistic Regression: A Simulation Study and Application to the Case of Postsurgical Venous Thromboembolism Following Total Hip and Knee Arthroplasty,” *Pharmacoepidemiology and Drug Safety*, 28, 217–226. [12]
- Nishiura, H., Kobayashi, T., Miyama, T., Suzuki, A., Jung, S.-M., Hayashi, K., Kinoshita, R., Yang, Y., Yuan, B., Akhmetzhanov, A. R., and Linton, N. M. (2020), “Estimation of the Asymptomatic Ratio of Novel Coronavirus Infections (COVID-19),” *International Journal of Infectious Diseases*, 94, 154–155. [1]
- Statistics Austria. (2020), “Prävalenz von SARS-CoV-2-Infektionen liegt bei 3,1%,” Technical Report. [12]
- Surkova, E., Nikolayevskyy, V., and Drobniewski, F. (2020), “False-Positive COVID-19 Results: Hidden Problems and Costs,” *The Lancet Respiratory Medicine*, 8, 1167–1168. [2,7]
- Woodward, M. (2014), *Epidemiology: Study Design and Data Analysis* (3rd ed.), London: Chapman and Hall/CRC. [12]
- Wooldridge, J. M. (2001), “Asymptotic Properties of Weighted M-estimators for Standard Stratified Samples,” *Econometric Theory*, 17, 451–470. [5]