# Estimating True Prevalence Through Questionnaire Data

Adam Mielke[1] and Lasse Engbo Christiansen[2]
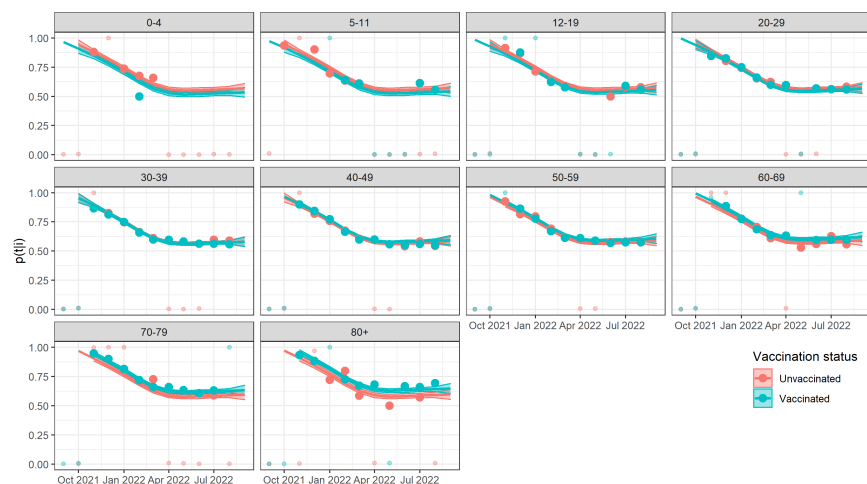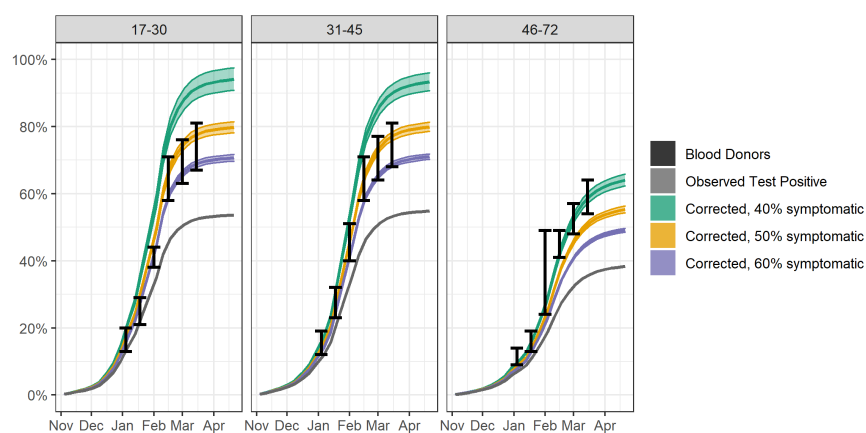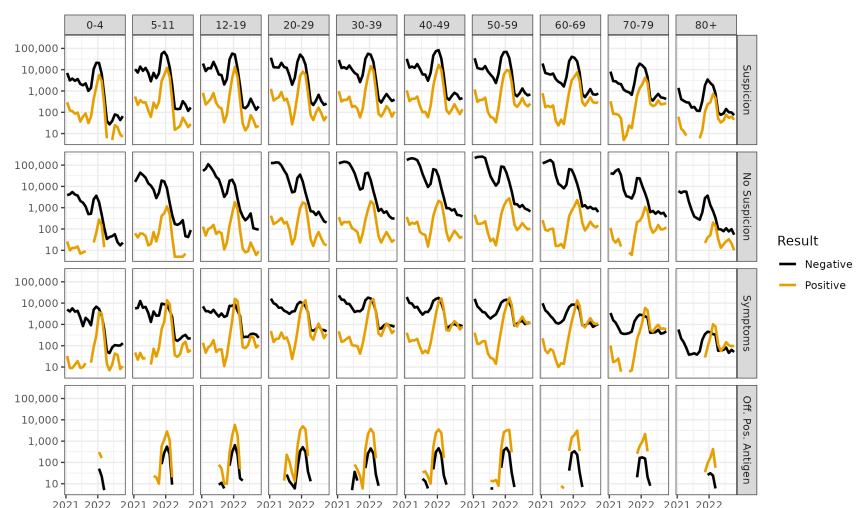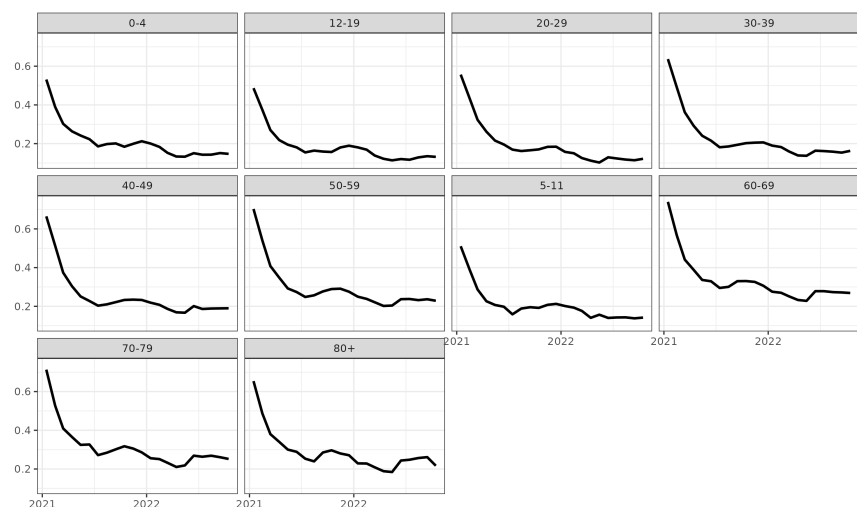
[1]Danmarks Tekniske Universitet
[2]Statens Serum Institut

February 3, 2023

## Abstract

A method using questionnaire data for estimating the level of under reporting during an outbreak is presented. It is based on rewriting the conditional probabilities for getting tested, being infected, and having symptoms. It shows very good agreement with seroprevalence studies of blood donors. On the one hand, this shows the strength of questionnaires when testing the general population during an outbreak as a means to find the true prevalence. On the other, applying it to covid-19 demonstrates that the asymptomatic cases likely make up around 50% of the infected.

1

# Estimating True Prevalence Through Questionnaire Data

Adam Mielke*

*Dynamical Systems, Technical University of Denmark,*
*Asmussens Allé, 303B, 2800 Kgs. Lyngby, Denmark*

Lasse Engbo Christiansen[†]

*Statens Serum Institut, Artillerivej 5, 2300 Copenhagen S, Denmark*
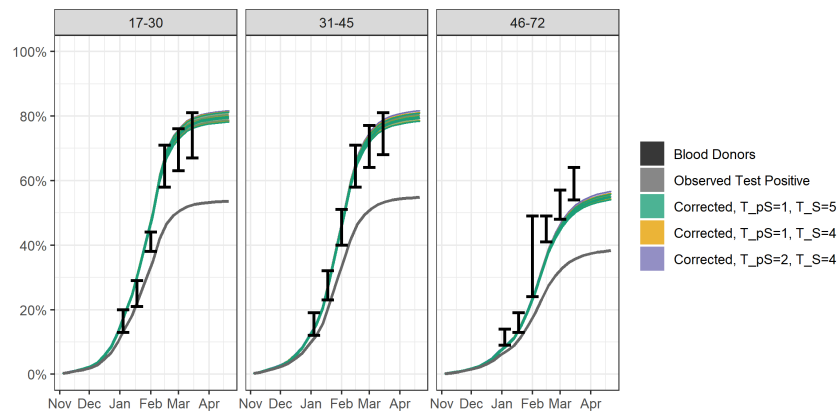( Dated: February 2, 2023)

A method using questionnaire data for estimating the level of under reporting during an outbreak is presented. It is based on rewriting the conditional probabilities for getting tested, being infected, and having symptoms. It shows very good agreement with seroprevalence studies of blood donors. On the one hand, this shows the strength of questionnaires when testing the general population during an outbreak as a means to find the true prevalence. On the other, applying it to covid-19 demonstrates that the asymptomatic cases likely make up around 50% of the infected.

* admi@dtu.dk
[†] lsec@ssi.dk

# I.   INTRODUCTION

Limited information is a fundamental problem in human epidemiology [1–8]. Lack of available tests, low compliance, and/or asymptomatic disease leave many cases undiscovered, and the proportion changes over time. Because of self-selection bias, the prevalence among the tested is rarely indicative of the prevalence in the population, which makes it very difficult to estimate severity of a given disease as well as predict when the number of infections peaks.

During most of 2021 and 2022, it has been possible for Danish citizens to get a free PCR-test for SARS-CoV-2 on www.coronaprover.dk. As either a negative test or vaccination was necessary for several activities during 2021, such as visiting restaurants, bars, hairdressers, and theaters, these test were taken often. However, as more people have received vaccination, and as COVID-19 fades from public memory, the question remains: "How large a proportion of the truly infected are actually detected at different points in the pandemic and to which extend could this information be used in future outbreaks?" To answer this, the present study delves into the role that detailed information about a subset of tests from a questionnaire can bring. An essential component is the risk of symptoms given infection and independent of test. With this, and dividing the screening tests into symptomatic and asymptomatic cases, the conditional probabilities can be inverted and compared, leading to an estimate of the probability of being tested given infection.

The paper is organized as follows. First the questionnaire data is presented in more detail along with the division of it. Second, the theoretic probability calculations are performed and the necessary comparison illustrated. Finally, the method is applied to test data from Denmark in the autumn and winter 2021-2022 and the merits of it discussed.

# II.   QUESTIONNAIRE DATA

When booking a PCR-test in Denmark, people had the option to fill out a questionnaire stating their reasons for being tested [9]. The possible answers are presented in Table I. While it is only a subgroup that fills out this questionnaire, there is still a significant amount. For the analysis in this paper, the answers are aggregated by month, age group, and risk level. See Figure 2 for the fraction of test over time with answered questionnaire and Figure 3 for the monthly number of positive and negative tests for all groups.
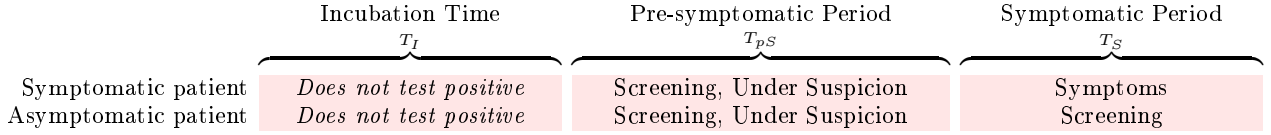
It becomes important for later analysis to note which stage of the disease each risk level can indicate, and whether a symptomatic or asymptomatic case is more likely. The primary answers are grouped into risk levels as follows:

- **Symptoms:** These have explicitly answered that they have symptoms, and positive tests must therefore be in the symptomatic part of a symptomatic case if they test positive.

- **Confirmatory PCR-Test:** Those that answer that they are getting a confirmatory PCR-test contain a mix of self-tests and official antigen tests. The self-tests are more difficult to determine the reason for, so only those who have a positive antigen from a test center less than 14 days before the PCR-test is booked are included, and the rest are discarded. As only one primary answer is allowed in the questionnaire, the positive PCR-tests in this category are assumed to be evenly distributed over the course of the disease.

- **Suspicion:** These have answered that they were in close contact with an infected or for some other reason believe they are at risk. It is assumed that these are taken relatively quickly after a potential infection, and positive tests are therefore assumed to be in the pre-symptomatic period.

- **Screening/no suspicion:** These are the remaining tests that are not at increased risk of being infected. It is therefore assumed that the positive screening tests are taken with equal probability either in the pre-symptomatic period for a symptomatic patient or at any point for asymptomatic patients.

| Primary Answer | Risk Level |
|---|---|
| COVID-19 symptoms | Symptomatic |
| Tested before treatment at hospital or similar (planned operation, dentist, etc.) | No Suspicion |
| Has been in contact with an infected person | Suspicion |
| Tested because of school, work, or education | No Suspicion |
| Tested because of activities during spare time | No Suspicion |
| Tested because of traveling in or out of Denmark | No Suspicion |
| Confirmatory PCR-test after positive antigen test | Confirmatory PCR-Test |
| Suspicion of infection without known contact to an infected | Suspicion |
| Has previously been ill with corona-like illness | Suspicion |
| Other/Participating in population study | No Suspicion |

**Table I:** The possible primary answers to the questionnaire along with the risk level assigned in the present analysis. Note that a number of secondary sub-answers also are available, such as the specific symptom or the type of work/contact, but these are not used here. See the Section II for elaboration on the risk levels.

See Figure 1 for a graphic illustration of the division of tests.

| | Incubation Time $T_I$ | Pre-symptomatic Period $T_{pS}$ | Symptomatic Period $T_S$ |
|---|---|---|---|
| Symptomatic patient | *Does not test positive* | Screening, Under Suspicion | Symptoms |
| Asymptomatic patient | *Does not test positive* | Screening, Under Suspicion | Screening |

**Figure 1:** Illustration of the different patients and times as well as the assumed possible answers to the questionnaire for positive test. Of course the asymptomatic patients do not have a symptomatic period, but as they are assumed to test positive for as long as the symptomatic patients, and as the division between early and late tests is important for the analysis, this divide is meaningful regardless.

## III. METHOD AND ASSUMPTIONS

The method is based on rewriting of conditional probabilities within a certain period of time. In this paper, a month is chosen in order to have enough questionnaire answers. This means that daily fluctuations are not captured, but the level of information during different stages of the epidemic is. The following quantities are relevant.

- **p(t)**: Probability of being tested for SARS-CoV-2 irrespective of disease. This is simply the fraction of the population that are tested. As self-tests are not considered here, $p(t)$ is known perfectly.

- **p(i)**: Probability of being infectious and positive if tested for SARS-CoV-2. This is the true prevalence and the goal of the method. The observed quantity is instead $p(i|t)$ (under assumption of 100% sensitivity) and lacks the untested population $p(i|\bar{t})$. An estimate of this will be made via the prevalence in the screened population who answered the questionnaire.

- **p(s)**: Probability of COVID-19-like symptoms. Note that $p(s|i)$ is the symptomatic COVID-19 cases, whereas $p(s)$ also includes symptoms from for instance the common cold, influenza, or RS virus.

From here two different starting points are made and compared. These two setups have physical interpretations, namely a population and an individual perspective, but should mostly be seen as mathematical tools.
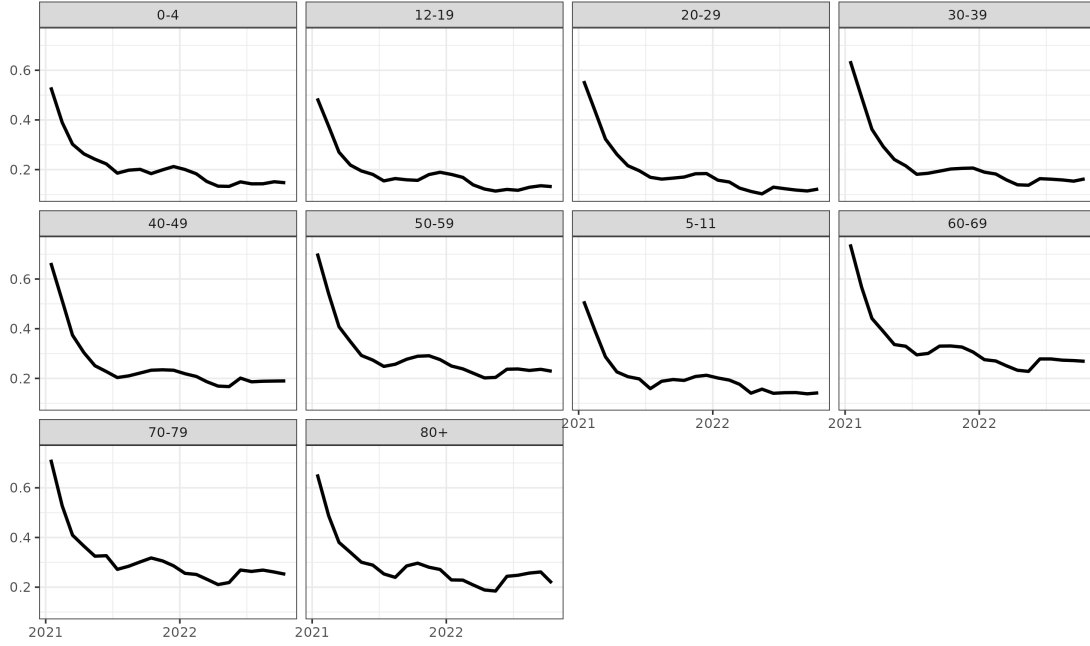
**Figure 2:** Fraction of tests where the questionnaire was answered.

### A.   Setup 1: From a Population Perspective

Start with the true prevalence in the population

$$p(i) = \underbrace{p(i|t)p(t)}_{\text{Detected cases}} + \underbrace{p(i|\bar{t})}_{\text{Prevalence among non-tested}} \underbrace{p(\bar{t})}_{\text{Non-tested population}} \tag{1}$$

The first term is simply the detected positive cases, and the second term is the amount that would be found if the rest of the population were tested. It is the prevalence among the untested $p(i|\bar{t})$ that is the unknown variable here.

It is tempting to assume that the prevalence is the same among the tested with no increased risk and the untested population, but self-selection bias is a big problem here. A more active subgroup is for instance more likely to be tested when tests are required for social events.

### B.   Setup 2: From an Individual's Perspective

How does this look from the individual's perspective? That is, how likely is an infected individual to get tested? Start with this probability

$$\underbrace{p(t|i)}_{\substack{\text{Getting tested when infected}}} = \underbrace{p(t|si)}_{\substack{\text{The symptoms} \\ \text{make you get tested}}} \underbrace{p(s|i)}_{\substack{\text{Symptomatic cases}}} + \underbrace{p(t|\bar{s}i)}_{\substack{\text{You get tested despite} \\ \text{being asymptomatic}}} \underbrace{p(\bar{s}|i)}_{\substack{\text{Asymptomatic cases}}} \tag{2}$$

The factors $p(s|i)$ and $p(\bar{s}|i) = 1 - p(s|i)$ are biological and denote the probability of getting a symptomatic illness. Note that to reconcile these probabilities, the pre-symptomatic cases are counted among the symp-
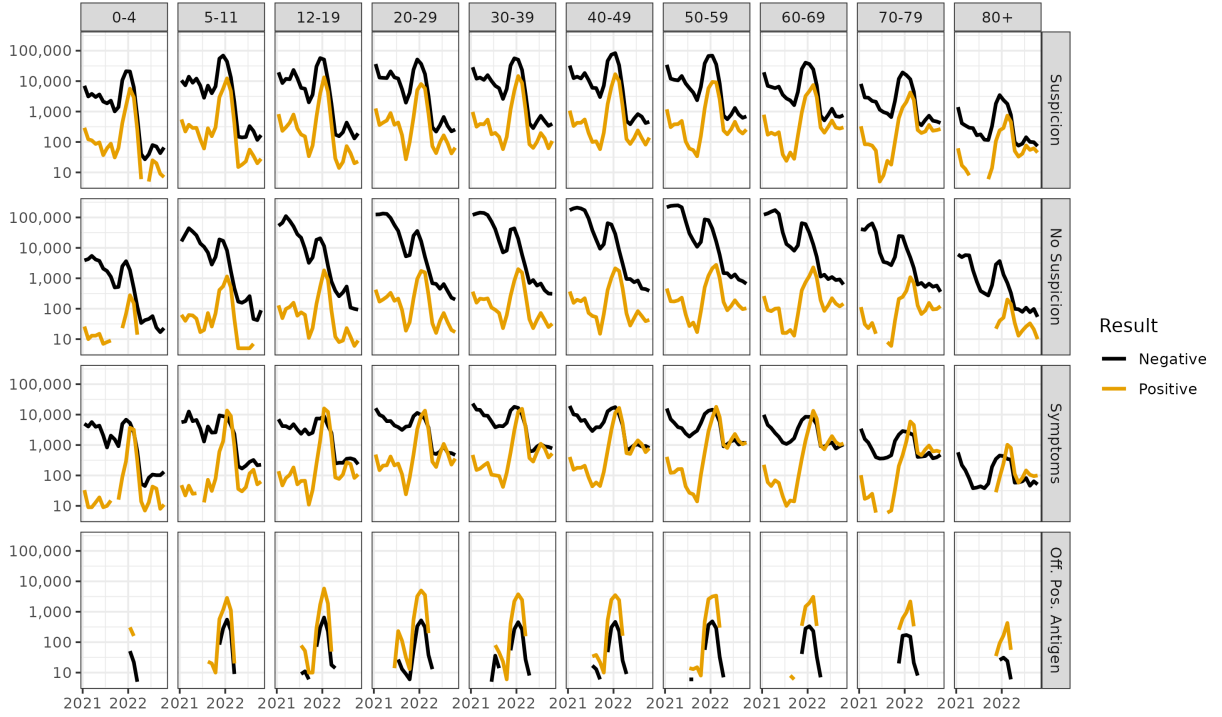
**Figure 3:** Number of tests with answered questionnaire taken per month in 2021 and 2022 for each risk level and age group. Numbers below 5 have been omitted in the plot to comply with GDPR-rules, but have been included in the analysis. Note the logarithmic scale to show both positive and negative tests on the same plots.

tomatic cases, even though the patient does not exhibit symptoms yet.

This setup allows for a number of rewritings. First, Bayes' Theorem is applied to $p(t|\bar{s}i)$ to find known quantities

$$p(t|\bar{s}i) = \frac{p(i|\bar{s}t)}{p(i|\bar{s})}p(t|\bar{s}) \ . \tag{3}$$

Both $p(i|\bar{s}t)$ and $p(t|\bar{s})$ can be estimated through the questionnaire data, see Section III C. So the challenge here is $p(i|\bar{s})$, i.e., the prevalence in the asymptomatic population. It can be subdivided according to whether they are tested as

$$p(i|\bar{s}) = p(i|\bar{s}t)p(t|\bar{s}) + p(i|\bar{s}\bar{t})p(\bar{t}|\bar{s}) \ . \tag{4}$$

The factor $p(\bar{t}|\bar{s}) = 1 - p(t|\bar{s})$ can again be estimated from the questionnaire data, and $p(t|si)$ can be rewritten the same way as $p(t|\bar{s}i)$.

Combining Equations (2), (3), and (4) yields

$$p(t|i) = \frac{p(s|i)}{1 + \frac{p(i|s\bar{t})(1-p(t|s))}{p(i|st)p(t|s)}} + \frac{p(\bar{s}|i)}{1 + \frac{p(i|\bar{s}\bar{t})(1-p(t|\bar{s}))}{p(i|\bar{s}t)p(t|\bar{s})}} \ . \tag{5}$$

**C. Combining the Setups**

Denote the contribution from self-selection bias in testing in the following way

$$\gamma := \frac{p(i|\bar{s}\bar{t})}{p(i|\bar{s}t)} \text{ for asymptomatic,} \tag{6}$$

$$\xi := \frac{p(i|s\bar{t})}{p(i|st)} \text{ for symptomatic.} \tag{7}$$

In other words, the ratio of positive percentage between untested and tested. These will be optimized later on. It is reasonable to expect $\gamma < 1$, since the asymptomatic people that are tested probably have some endeavor, which in turn makes them more likely to be infected. (That is, they are a more active population.) A similar argument may be made for $\xi$, but note that symptomatic people that are not tested during an outbreak is an unclear group. They are either very sure what they are infected with, or have low compliance. In any case, $\gamma$ and $\xi$ need not be bounded and may vary over time.

Equations (1) and (5) are rewritten in term of $\xi$ and $\gamma$

$$p(i) = p(i|t)p(t) + (p(i|\bar{t}s)p(s|\bar{t}) + p(i|\bar{t}\bar{s})p(\bar{s}|\bar{t}))\, p(\bar{t})$$
$$= p(i|t)p(t) + (\xi p(i|ts)p(s|\bar{t}) + \gamma p(i|t\bar{s})p(\bar{s}|\bar{t}))\, p(\bar{t}) \tag{8}$$

$$p(t|i) = \frac{p(s|i)}{1+\xi \frac{(1-p(t|s))}{p(t|s)}} + \frac{p(\bar{s}|i)}{1+\gamma \frac{(1-p(t|\bar{s}))}{p(t|\bar{s})}}\ . \tag{9}$$

Using the relations

$$p(t|\bar{s}) = \frac{p(\bar{s}|t)p(t)}{p(\bar{s})} = \frac{p(\bar{s}|t)p(t)}{p(\bar{s}|t)p(t)+p(\bar{s}|\bar{t})p(\bar{t})}\ ,$$
$$p(t|s) = \frac{p(s|t)p(t)}{p(s)} = \frac{p(s|t)p(t)}{p(s|t)p(t)+p(s|\bar{t})p(\bar{t})}\ , \tag{10}$$

where

$$p(\bar{s}|t) = p(\bar{s}|it)p(i|t) + p(\bar{s}|\bar{i}t)p(\bar{i}|t)\ ,$$
$$p(s|t) = p(s|it)p(i|t) + p(s|\bar{i}t)p(\bar{i}|t)\ ,$$
$$p(\bar{s}|\bar{t}) = p(\bar{s}|\bar{t}i)p(i|\bar{t}) + p(\bar{s}|\bar{t}\bar{i})p(\bar{i}|\bar{t})\ ,$$
$$p(s|\bar{t}) = p(s|\bar{t}i)p(i|\bar{t}) + p(s|\bar{t}\bar{i})p(\bar{i}|\bar{t})\ , \tag{11}$$

and

$$p(\bar{s}|i) = p(\bar{s}|\bar{t}i)(1-p(t|i)) + p(\bar{s}|ti)p(t|i)\ ,$$
$$p(s|i) = p(s|\bar{t}i)(1-p(t|i)) + p(s|ti)p(t|i)\ , \tag{12}$$

the following relations may be derived from Equation (5)

$$p(t|i) = \frac{p(s|i)}{1+\xi\left(1+\frac{p(s|\bar{t})p(\bar{t})}{p(s|t)p(t)}\right)\big/\left(1+\frac{p(s|t)p(t)}{p(s|\bar{t})p(\bar{t})}\right)} + \frac{p(\bar{s}|i)}{1+\gamma\left(1+\frac{p(\bar{s}|\bar{t})p(\bar{t})}{p(\bar{s}|t)p(t)}\right)\big/\left(1+\frac{p(\bar{s}|t)p(t)}{p(\bar{s}|\bar{t})p(\bar{t})}\right)}\ , \tag{13}$$

$$p(s|\bar{t}) = \frac{\gamma p(i|t\bar{s})}{\frac{1-p(t|i)}{p(s|i)-p(s|ti)p(t|i)} - (\xi p(i|ts) - \gamma p(i|t\bar{s}))} \tag{14}$$

$$p(\bar{s}|\bar{t}) = \frac{1-\left(1-\frac{p(\bar{s}|i)-p(\bar{s}|ti)p(t|i)}{1-p(t|i)}\right)\xi p(i|ts)}{1+\left(1-\frac{p(\bar{s}|i)-p(\bar{s}|ti)p(t|i)}{1-p(t|i)}\right)(\xi p(i|ts)-\gamma p(i|t\bar{s}))}\ . \tag{15}$$

The assumption $1 - p(s|\bar{t}i) = p(\bar{s}|\bar{t}i) \approx 1$ has also been used. That is, the non-infected, non-tested people will largely be asymptomatic. This breaks down if two parallel outbreaks with different diseases are occurring, but this is assumed not to be the case. Note that simply approximating $p(\bar{s}|\bar{t}) = 1 - p(s|\bar{t}) \approx 1$ implies that all symptomatic cases are detected, which we consider unrealistic, especially in times with low degree of testing.

Once all other quantities have been determined, Equation (13) is solved for $p(t|i)$ numerically. Using the relation

$$p(i)p(t|i) = p(t)p(i|t) , \tag{16}$$

Equations (8) and (13) may be compared. The parameters $\xi$ and $\gamma$ are optimized based on the relative difference between $p(t|i)$ in the two setups. The most obvious point where this occurs is of course $\xi = \gamma = 0$, which corresponds to $p(t|i) = 1$. This is unrealistic, and results $p(t|i) > 0.95$ and $p(t|i) < 0.05$ are therefore ignored. So rather unusually, the optimizer should avoid the global minimum at $\xi = \gamma = 0$ and instead find a local minimum.

## D.    Estimating Probabilities with Questionnaire Data

The main part of this section is correct distribution of the test categories described in Figure 1. Denote the duration of symptoms, pre-symptomatic period, and incubation time by $T_S$, $T_{pS}$, and $T_I$ respectively. It is assumed that symptomatic and asymptomatic patients test positive in the same duration of time, namely $T_S + T_{pS}$. The probabilities are then estimated in the following way

- $p(\bar{s}|it)$: This is illustrated in Figure 1. Making a weighted sum of the positive tests, these consists of $p(\bar{s}|i)$ of the positive under suspicion, $p(\bar{s}|i)$ of the positive confirmatory tests, and $\frac{(T_{pS}+T_S)p(\bar{s}|i)}{T_{pS}p(s|i)+(T_{pS}+T_S)p(\bar{s}|i)}$ of the screening tests. This is normalized by the total number of positive who filled out the questionnaire.

- $p(\bar{s}|\bar{i}t)$: The best estimate here is simply the fraction of negative tests with a questionnaire answer other than "symptomatic". This assumes that there is no particularly high probability of developing symptoms if you are a non-infected person under suspicion. (This is of course not completely correct as someone under suspicion will probably have been in contact with more people and thus be more likely to also catch another disease, but this effect is assumed to be negligible.)

This of course also indirectly estimates $p(s|it) = 1 - p(\bar{s}|it)$ and $p(s|\bar{i}t) = 1 - p(\bar{s}|\bar{i}t)$.

## IV.    RESULTS AND DISCUSSION

There is some uncertainty regarding the probability of getting a symptomatic illness [10–14], but most estimates, with the exception of [11], are around $p(s|i) = 0.5$. The main estimate used in the following will be $p(s|i) = 0.5$, but $p(s|i) = 0.4$ and $p(s|i) = 0.6$ will also be investigated.

The age groups 0-4, 5-11, 12-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, and 80+ are used. These are relevant because they coincide with those used by the Danish model group for prediction of COVID-19 [15]. They approximately divide the population into both vaccination and behavioral groups at the same time. Again to conform with these models, the time parameters $T_S = 4$ days and $T_{pS} = 1$ day are used. The length of the incubation time is not used. Figure 4 present estimates of $p(t|i)$ for each month among unvaccinated and vaccinated in all age groups.

Additionally, in order to compare with estimates from seroprevalence of blood donors [16], $p(t|i)$ is estimated in the age groups 17-30, 31-45, and 46-72 as well. The observed number of test positive is corrected,
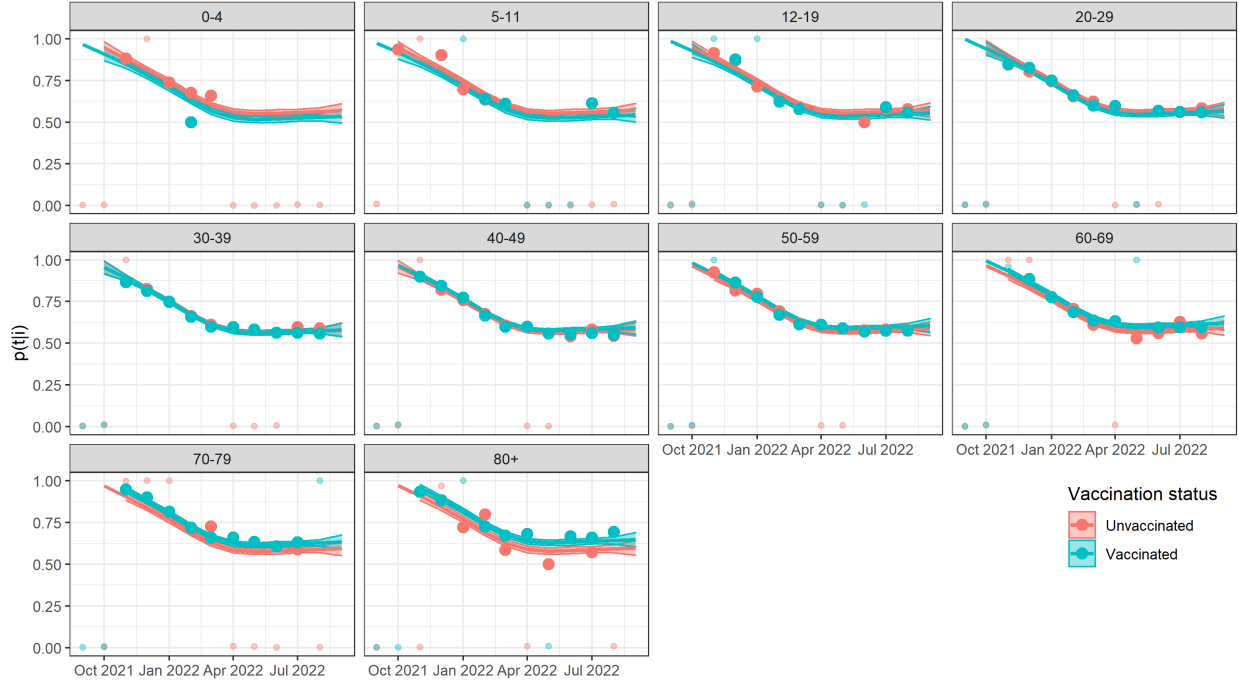
**Figure 4:** Result of the $p(t|i)$-fit for each age group and vaccine status. Each point is an independent fit to the tests and answers in that month. Points $p(t|i) > 0.95$ and $p(t|i) < 0.05$ are ignored. The fact that the curves for the successful fits are more or less continuous lends great merit to the method. A linear model with quadratic b-splines and 4 degrees of freedom has been used to smoothen out the points such that the estimate can be used in continuous models. The ribbons indicate this smoothened fit including $\pm 1$ standard deviation.

stratified on age group and vaccine status, before being summed up, see Figure 5. This comparison may be interpreted in two different ways: If the estimate $p(s|i) = 0.5$ is trusted, it highlights the precision of the method with the given assumptions. If the $p(i|s)$-fit is trusted, it places a very tight bound on the number of asymptomatic there are.

The questionaire was part of the booking of tests and during some periods there was a waiting time of a couple of days before one could get a screening test. This impacts the probability of being positive in the presymptomatic period and this was investigated by reestimating with $T_{pS} = 2$ days. The time parameters have very little impact on the estimate, see Figure 6 for $T_S = 5$ days and $T_{pS} = 2$ days.

Note the advantages of this approach. It can be extended to include outbreaks of several diseases with similar symptoms at the same time. It can also be generalized to any property (and not just general symptoms) that can be determined independently of tests, as long as it can be expected to be constant over the course of the relevant time period. Speaking in terms of a general disease, this can for instance be a specific symptom, or it could be a biological or social marker such as gender or age that in some known way correlates with the risk of infection. Distinguishing between symptoms is likely to add extra robustness to the method. However, parameters such as risk of serious illness or hospitalization are poorly suited, as they are difficult to get independent of tests. These risks may of course be derived much more accurately for future use once the true prevalence is known.
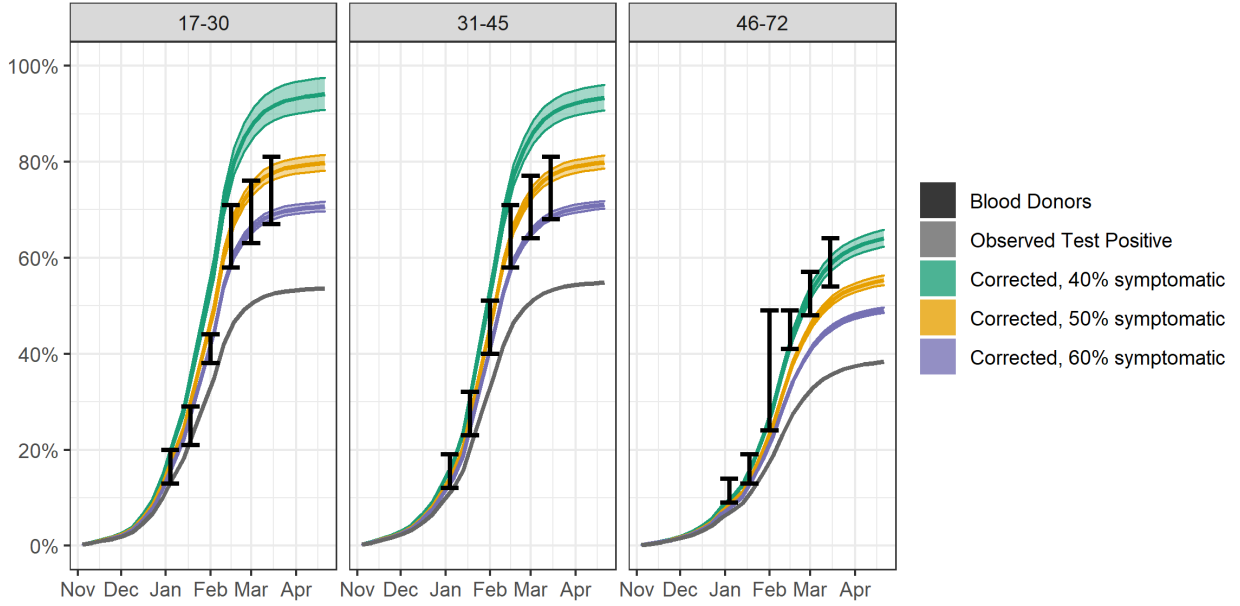
**Figure 5:** Cumulative test positive per population for different age groups in Denmark in the November 2021 to April 2022. The observed test positive ($p(i|t)p(t)$, [9], gray line) and corrected numbers ($p(i) = p(i|t)p(t)/p(t|i)$, ribbons) for different values of $p(s|i)$ are compared to seroprevalence in blood donors ([16], black error bars). The results from the spline fit in Figure 4, including $\pm 1$ standard deviation from the spline fit, have been used for $p(t|i)$ to compare with weekly numbers of test positive. The corrected numbers for $p(s|i) = 0.5$ show remarkable agreement with the blood donor study. The discrepancy for the oldest age group most likely comes from selection bias, and not a lower risk of symptoms. Blood donors represent the healthier part of the older population, which presumably also makes them more active and social, and thus more likely to be exposed.

## V.    CONCLUSION

Using answers given when taking a test, estimates of the true prevalence is made. The key point is that the conditional probabilities can be approached in two different ways, and requiring the two setups to give the same probabilities gives a measure to optimize, which helps combat the selection bias in the tested population.

When applying this method to the Danish test data over the autumn and winter 2021-2022, it shows very good agreement with seroprevalence studies in blood donors when 50% are asymptomatic cases. This illustrates the advantages of getting additional information for each test when providing accessible and free testing in a population. A consistent estimate of the true prevalence can be made based on an initially determined biological marker, such as the risk of symptoms. Or if the true prevalence can be obtained elsewhere, the method can give a very clear estimate of the biological marker.

## ACKNOWLEDGMENTS
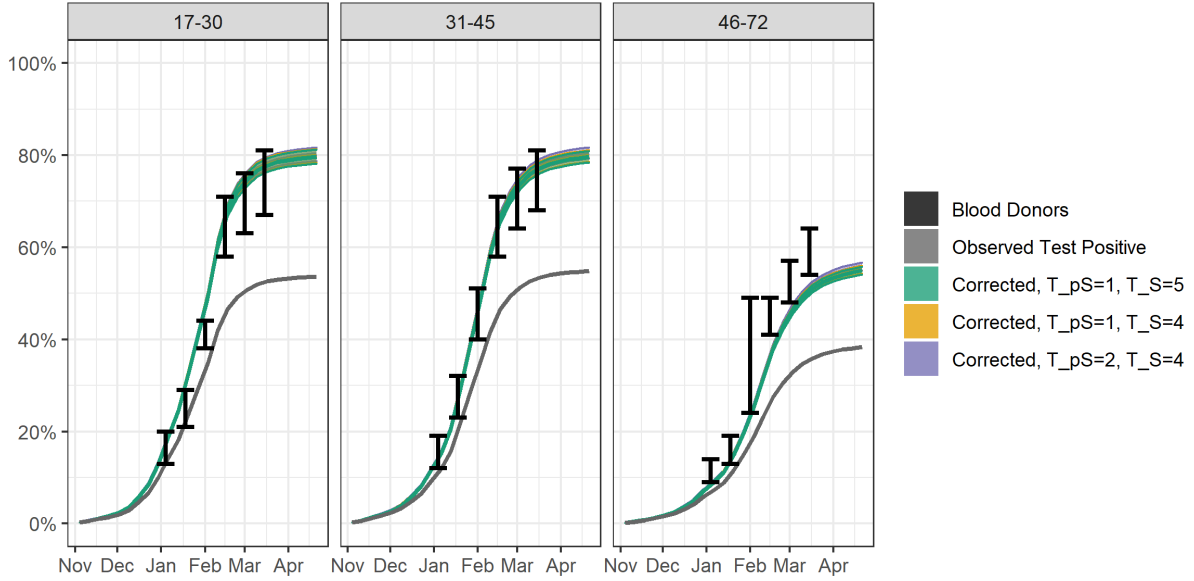
**Figure 6:** Cumulative test positive per population for different age groups in Denmark in the November 2021 to April 2022. The observed test positive ($p(i|t)p(t)$, [9], gray line) and corrected numbers ($p(i) = p(i|t)p(t)/p(t|i)$, ribbons) for different values of of $T$ are compared to seroprevalence in blood donors ([16], black error bars). It is clear that the duration of the symptomatic and pre-symptomatic periods makes very little difference to the estimates.

## VI.   BIBLIOGRAPHY

[1] R. Marier, *The reporting of communicable diseases*, Am J Epidemiol 105:587–90 (1977).

[2] P. M. Konowitz, G. A. Petrossian, D. N. Rose, *The underreporting of disease and physicians' knowledge of reporting requirements*, Public Health Rep 99:31–5 (1984).

[3] T. J. Doyle, M. K. Glynn, and S. L. Groseclose, *Completeness of notifiable infectious disease reporting in the United States: an analytical literature review*, Am J Epidemiol. 155(9):866-74 (2002).

[4] R. S. Hopkins, *Design and operation of state and local infectious disease surveillance systems*, J Public Health Management Practice 11(3):184—90 (2005).

[5] C. L. Gibbons, M.-J. J. Mangen, D. Plass, A. H. Havelaar, R. J. Brooke, P. Kramarz, K. L. Peterson, A. L. Stu-urman, A. Cassini, E. M. Fèvre, and M. E. E. Kretzschmar, *Measuring underreporting and under-ascertainment in infectious disease datasets: a comparison of methods*, BMC Public Health. 14, 147 (2014).

[6] J. Wilting, and V. Priesemann, *Inferring collective dynamical states from widely unobserved systems*, Nat Commun 9, 2325 (2018).

[7] H. Laua, T. Khosrawipour, P. Kocbachc, H. Ichiia, J. Baniad, and V. Khosrawipour, *Evaluating the massive underreporting and undertesting of COVID-19 cases in multiple global epicenters*, BMC Infect Dis. 27, 2 (2021).

[8] V. Albani, J. Loria, E. Massad, and J. Zubelli, *COVID-19 underreporting and its impact on vaccination strategies*, BMC Infect Dis. 21, 1 (2021).

[9] Stratified data provided directly by Statens Serum Institut Denmark. Aggregated numbers available at https://covid19.ssi.dk/overvagningsdata/download-fil-med-overvaagningdata (In Danish, visited 2022-12-12).

[10] D. P. Oran and E. J. Topol, *Prevalence of Asymptomatic SARS-CoV-2 Infection: A Narrative Review*, Annals of Internal Medicine 173, 5, 362-367 (2020).

[11] O. Byambasuren, M. Cardona, K. Bell, J. Clark, M.-L. McLaws, and P. Glasziou, *Estimating the extent of asymptomatic COVID-19 and its potential for community transmission: Systematic review and meta-analysis*, J.

Assoc. Med. Microbiol. Infect. Dis. Can. 5, 4 (2020).

[12] G. Syangtan, S. Bista, P. Dawadi, B. Rayamajhee, L. B. Shrestha, R. Tuladhar, and D. R. Joshi, *Asymptomatic SARS-CoV-2 Carriers: A Systematic Review and Meta-Analysis*, Front. Public Health 8, (2021).

[13] P. Sah, M. C. Fitzpatrick, C. F. Zimmer, E. Abdollahi, L. Juden-Kelly, S. M. Moghadas, B. H. Singer, and A. P. Galvani , *Asymptomatic SARS-CoV-2 infection: A systematic review and meta-analysis* PNAS 118, 34 (2021).

[14] Q. Ma, J. Liu, Q. Liu, L. Kang, R. Liu, W. Jing, Y. Wu, M. Liu, *Global Percentage of Asymptomatic SARS-CoV-2 Infections Among the Tested Population and Individuals With Confirmed COVID-19 Diagnosis: A Systematic Review and Meta-analysis*, JAMA Netw Open 4, 12 (2021)

[15] https://covid19.ssi.dk/analyser-og-prognoser/modelberegninger (In Danish, visited 2022-12-21).

[16] Statens Serum Institut Denmark, *Seroprævalensundersøgelse af bloddonorer*, https://covid19.ssi.dk/overvagningsdata/ undersoegelser/moerketal-for-covid-19 (In Danish, visited 2022-11-21).