On the Measurement of Disease Prevalence

Sotiris Georganas, Alina Velias and Sotiris Vandoros*

February 12, 2021

Abstract

For any infectious disease, including the Covid-19 pandemic, timely, accurate epidemic figures are necessary for informed policy. In the Covid-19 pandemic, mismeasurement can lead to tremendous waste, in health or economic output. "Random" testing is commonly used to estimate virus prevalence, reporting daily positivity rates. However, since testing is necessarily voluntary, all "random" tests done in the field suffer from *selection bias*. This bias, unlike standard polling biases, goes beyond demographical representativeness and cannot be corrected by oversampling (i.e. selecting people without symptoms to test). Using controlled, incentivized experiments on a sample of all ages, we show that people who feel symptoms are up to 42 times more likely to seek testing. The testing propensity bias leads to sizeable *prevalence bias*: even under costless testing, test positivity can inflate true prevalence fivefold. The inflation factor varies greatly across time and age groups, making intertemporal and between nation comparisons misleading. We validate our results using the largest population surveillance studies of Covid-19 in England, and indeed find that the bias varies intertemporally from 4 to 23 times. We present calculations to debias positivity, but importantly, suggest a parsimonious approach to sampling the population that bypasses the bias altogether. Estimates are both real-time and consistently close to true values. Our results are relevant to any epidemic, besides Covid-19, where carriers have informative beliefs about their own status.

Keywords: covid-19, prevalence, pandemic, bias, self-selection

JEL Codes: D90, C81, I18

*We thank Andrew Atkeson, Rebecca Thornton. William Havage, PJ Healy, Ichiro Kawachi and Roberto Weber for useful comments, Fay Angelidou for outstanding reserch support, and the Department of Economics, City University, for funding the experiments.

1 Introduction

How to measure prevalence for infectious diseases? In the Covid-19 pandemic, health agencies and lay citizens alike, closely watch two measures derived from daily testing, the absolute number of recorded cases and the percentage of positives in the tested population. These numbers influence individual decisions but also official measures against the pandemic, with a profound impact on public health and the economy.

In this paper we claim that such commonly used measures are fundamentally flawed, because they ignore the demand side for testing. In virtually all countries in the world, testing is voluntary, leading to *self-selection bias*. People are likelier to self-select into testing if they have reasons to believe they might be having Covid-19 (such as, e.g. if they have symptoms or if they are exposed to a high-risk environment). We experimentally show there is a substantial bias in testing, driven by self-selection and demonstrate how the testing bias translates into *biased prevalence estimates*. We then validate our results on how the accuracy of prevalence estimation is affected by the bias, using external data. Finally, we propose a novel, fast and relatively economical method to estimate prevalence in real time, using a combination of polling methods and characteristics of endogenously done virus testing.

Let us start with a simple illustration of the problem for both economic policy makers and health agencies, using a real example from a European country. During Christmas all shops and schools were closed. On January 18 2020 the government allowed elementary schools and the retail sector to open (for in-store buys). About a week later, recorded cases started to rise. On the 29th of January, 941 cases were recorded, almost double the cases a week before (506). Ignoring standard statistical questions of significance, two questions arise: is that rise in cases a clear sign of a worsening disease, and can we blame the retail sector or schools? Due to the selection bias, even the first question is hard to answer. At the same time as cases rose, testing rose too. The number of tests on 29 January was about double the amount of tests on the 22nd. Actually, *test positivity* is similar between these dates. But our self-selection argument implies that the number of tests is endogenous. Higher disease

prevalence leads to higher demand for testing. As we will demonstrate, the self-selection bias changes over time, making comparisons using test positivity data meaningless in many cases.

The self-selection bias is idiosyncratic and varies with age, which complicates answering the second question too, how to tell whether schools or shops are to blame. One would think to (and indeed, health agencies *do*) compare test positivity among school pupils and middle aged people who went shopping, to see what channel of infection was more important. But our experiments show that demand for testing differs strongly by age, and also, virus symptoms affect this demand differentially. This means we cannot compare positivity across age groups either.

The use of standard test positivity or the number of recorded cases, to compare prevalence over time or across age groups, is rarely advisable. In the paper we use incentivised controlled experiments to estimate the size of the testing bias and calculate the corresponding prevalence bias. Interestingly the bias is estimated to be drastically different by age groups (as mentioned above) and to also rely greatly on two important characteristics of the testing procedure: waiting times and cost.

Our testing bias estimates can be used to calculate the prevalence bias, and debias the current prevalence estimates in the field (as derived from test positivity). As long as the characteristics of the testing procedures are known by the health agencies and published (unfortunately, the former is rare and the latter is most often not the case), we sketch the parameter estimations necessary for debiasing. Given that we have estimates by age, simulations can be done for countries with different demographic structures too. Of course accurately estimating all necessary parameters presents challenges of its own.

To fix all the problems with measurement, we suggest a novel method to *bypass* the self-selection bias altogether, with an estimation procedure that is at the same time faster, more accurate and more feasible than current methods. The idea is to poll a representative sample about their symptoms, and get the symptoms-to-virus conversion parameters from

existing tests.[1]

To give an overview of the experimental results, we find that the commonly used "test positivity" measure may inflate actual prevalence by up to 5 times, even if testing is provided at zero cost. If Covid-19 tests are costly for the testee (as is common), this inflation factor or *prevalence bias* can be much higher. To make estimation harder, the prevalence bias is *not constant*, but rather depends strongly on actual prevalence. This means that we cannot apply a fixed adjustment to test positivity measures, and such measures cannot be used to compare prevalence across countries, as is commonly done. To validate our results, we compare prevalence estimates from the REACT and ONS studies in the UK, to test positivity ratios at the same dates (Riley et al., 2021). As predicted by our calculations, the prevalence bias is indeed positive, very large and time varying, ranging from 3.8 to 23.6 in the different waves of the study. To say it another way, our estimates of the testing bias and calculations of how this translates to a prevalence bias, explain why test positivity rates always seem to be too high.

To understand the relevance of these results, start by noting that suggested policy responses and their implementation (e.g. social distancing rules) will inevitably be inefficient if we are not aware of the real number of active cases, and in which areas and age groups these occur. Observing mortality rates or the number of hospitalisations and patients in ICU are not real time measurements; they only provide an estimate of how many people caught Covid-19 *weeks earlier* (and estimating the fatality rate is also challenging, Atkeson, 2020). This time lag is very important when trying to evaluate interventions. Without real time data, measuring the effect of a vaccine will take months, on top of the time the vaccine takes to have a medical effect. Understanding the full effect of other events on the disease, like the Christmas holidays (which led to more interaction and possibly higher transmission)

[1]Replacing mass testing with polling may sound unusual, but it is in line with suggestions of using statistical sampling to replace exhaustive counting, when the latter can be biased, as in a census. In the case of the pandemic, it has even been argued that symptoms-based diagnosis should be used instead of PCR testing (Cadegiani et al., 2021), because it is more informative.

similarly takes months (see the influential cross country study on the effectiveness of pharmaceutical interventions (NPIs), which uses death counts, lagging by several weeks, Brauner et al., 2020). On the other hand, knowing the current number of actual cases, allows the design of optimal policy response, and also provides a forward-looking estimate of hospitalisations and mortality. Health systems get warning several weeks ahead, gaining invaluable time for necessary adjustments.

Community testing, often conducted in the high street and in neighbourhoods, is widely considered a useful tool to monitor incidence and trends. The ECDC, 2020 listed "[to] reliably monitor SARS-CoV-2 transmission rates and severity" among five objectives of testing. It also published weekly testing data and "positivity rates" by EU State (ECDC, 2021). However, as we have argued, such testing cannot provide accurate estimates of Covid-19 prevalence, and the main problem is not related to typical issues that arise in population sampling, such as sampling representative age groups (contrary to what some studies suggest (Pearce et al., 2020)). We find in our data that the self-selection bias increases non-linearly with waiting times and any other cost associated with testing. To make prevalence estimation harder, the bias is time-varying, and also depends non-linearly on time varying parameters. For example, when cases rise steeply, people might be more likely to want to test out of fear. This leads to longer queues for testing, longer waiting times and a disproportionately larger testing bias.

To summarise, using standard self-selection calculations and results from a large scale, incentivised and controlled experiment, we formulate three main hypotheses:

1. Test-positivity is always inflated due to self-selection

2. The inflation factor is time-varying

3. As virus prevalence in the population increases, so does the bias in its measurement (for reasonable prevalence ranges in the Covid-19 pandemic)

We validate our results comparing prevalence estimates from the REACT study in the UK, to official test positivity figures. Our two predictions are strongly confirmed and we

also find support for the hypothesis that the bias rises with prevalence.

Finally, we present an application of the testing bias to the much debated policy question of school openings. We show that the testing bias can explain why the young do not show up in simple case counts, while they are very likely getting infected (and possibly transmitting) more than older people.

The possibility that infection rates in the untested population can be different than in the tested subsample, has been raised (Manski and Molinari, 2021). The issue is treated as a purely econometric inference problem however, with no reference to self-selection. In a somewhat similar vein, (Greene et al., 2021) propose statistical nowcasting, but the accuracy of both these methods is not as high as polling and detection of trend reversals is not possible in real time.

Experimental methods with incentives have been used on virus testing before, in a seminal paper to measure demand for HIV testing (Thornton, 2008). However prevalence estimation was not the goal of that paper, and of course the diseases are different in several ways.

More generally, the existing literature does not offer much guidance on personal incentives to test on a large scale. Should people be averse to learning they are infected, as information avoidance models suggest (Golman et al., 2017), prevalence figures would be deflated due to symptomatic people testing less that non-symptomatic ones. If, however, people do not test unless they experience symptoms, as is a known case in medical literature (Oster et al., 2013), this would lead to inflated prevalence figures due to non-symptomatic people testing less frequently than symptomatic ones.

Why care about test positivity rates? These are currently widely used to evaluate the effect of the mass testing *within* a country (Mahase, 2020), to compare the effect of government policies *between* countries (Haug et al., 2020), to build arguments about which age or socio-demographic groups are most affected (Elimian et al., 2020), and generally as a "baseline against which the impact of subsequent relaxation of lockdown can be assessed" (p2, Riley et al., 2020). A biased prevalence estimate makes these comparisons at best uninformative

(Middelburg and Rosendaal, 2020) - a problem to which we offer a solution.

Our approach is also relevant for past research based on historical data. For example, major studies of policy measures to prevent spread of viral diseases rely on prevalence estimates affected by the same type of bias (Adda, 2016).

Some studies rely on death rates instead of test positivity to evaluate effect of the policies aimed to contain the pandemics (Dergiades et al., 2020). This measure does not circumvent the problem of incomparability. Deaths are affected by harvesting and specifics of the health system, so do not fit as a perfect proxy of prevalence for cross country comparisons. Likewise, the infection fatality rates (IFR) are also subject to the testing bias. Whilst researchers already raise concerns about methodological and econometric issues affecting IFR (Shen et al., 2021), the bias we find cannot be addressed by the measures they propose.

The rest of this paper is organised as follows. Section 2 presents calculations of the self-selection testing bias. Section 3 describes the experimental procedures to measure this bias. Section 4 presents the experimental results and their implications regarding the prevalence bias. Section 5 compares our debiasing solutions, partly with parameters derived from the experiments, to field data. Section 6 presents an application to a common policy problem, the evaluation of school openings, while Section 7 concludes.

2 Bias calculations

The aim of the calculations is to infer the percentage of sick people in the population from the "random" testing in the field figures, as released by Health Agencies worldwide. The problem is that testing is voluntary, which leads to selection bias. How large is this bias?

To start, some people believe they have symptoms, some do not: call them $S(ymptomatic)$ and $H(ealthy)$. Note that the discussion below has to do with what people believe, not what they actually have. Also, we distinguish between people believing they have symptoms and those who do not, but the analysis readily extends to people having strong beliefs that they

might be carrying the virus and those who do not.

Let the frequency of people who believe they have symptoms be p_s, or just p, with $(1-p)$ being the frequency of people who do not think they have symptoms.

Of each group, some percentage turns out having the virus. Let v_s be the virus prevalence for those who believe they have symptoms, v_h for those who do not.

Of each group, some percentage are willing to take the test (for a given waiting time to take the test). Assume this only depends on symptoms, but not on actually having the virus (this assumption is mostly innocuous, unless there is a very large number of people in hospital). Let then t_s be the percentage of people who believe they have symptoms who actually take the test, and t_h for those who do not.

True prevalence is then

$$\tau = p_s v_s + (1 - p_s) v_h \tag{1}$$

The sample prevalence, also called test positivity throughout the paper (i.e. the virus frequency in the sample population) ϕ, however, is given by the positive rate in the sample (assuming that the test itself is perfect).

$$\pi = p_s t_s v_s + (1 - p_s) t_h v_h \tag{2}$$

Divided by the total sampling rate

$$m = p_s t_s + (1 - p_s) t_h \tag{3}$$

Note that if $t_s = t_h = t$, then $\pi = t(p_s v_s + (1 - p_s) v_h)$ and $\phi = t(p_s v_s + (1 - p_s) v_h)/t = p_s v_s + (1 - p_s) v_h = \tau$ which makes sense; if testing propensities are equal, there is no bias.

If on the other hand the testing propensities t are not the same, then the sample is selected leading to bias. Before we calculate the bias, express the propensities to test and be virus positive, for the people who believe they have symptoms, as a multiple of the propensities of those who do not: $v_s = a v_h, t_s = b t_h$ vs = a vh, ts = b th . Then, using these equations,

rewrite (1), (2) and (3).

$$\tau = p_s v_s + (1 - p_s)v_h = a p_s v_h + (1 - p_s)v_h = v_h(a p_s + 1 - p_s)$$

$$\pi = p_s t_s v_s + (1 - p_s)t_h v_h = ab p_s t_h v_h + (1 - p_s)t_h v_h = t_h v_h(ab p_s + 1 - p_s)$$

$$m = p_s t_s + (1 - p_s)t_h = b p_s t_h + (1 - p_s)t_h = t_h(b p_s + 1 - p_s)$$

Simplify the notation by writing p for p_s and calculate

$$\phi = \frac{\pi}{m} = \frac{t_n v_h(abp + 1 - p)}{t_h(bp + 1 - p)} = \frac{v_h(abp + 1 - p)}{(bp + 1 - p)}$$

Now, divide $\frac{\phi}{\tau}$ which yields the bias in the estimates

$$\beta = \frac{abp + 1 - p}{(ap + 1 - p)(bp + 1 - p)}$$

For example, suppose the true symptoms prevalence is 10%, $p = 0.10$. Then $\beta = (0.1ab + 0.9)/(0.1a + 0.9)/(0.1b + 0.9)$. Figure 1 illustrates the size of the prevalence bias for different values of a and b. For instance, if $a = b = 20$, street testing is overestimating the virus prevalence by about 5 times.

In order to debias the test positivity in the field, on simply has to deflate the field figures by the estimated β, as long as p is known. If it is not, calculations are available upon demand to get p from the data.

3 Experiment Design

To find the testing propensity parameters t_h and t_s, we design an incentivised experiment where we

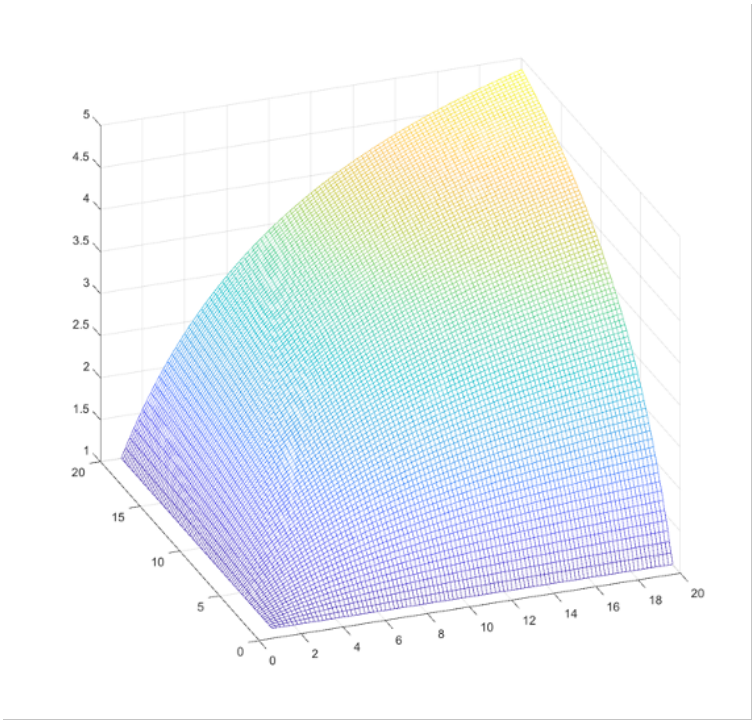

Figure 1: Bias in estimates for the true prevalence of 10%.(z-axis). X-axes: a (propensity-to-be virus positive ratio between those who have symptoms and not). Y-axis: b (propensity-to-test ratio between those who have symptoms and not).

1. Elicit hypothetical willingness to wait (WTW) to take a rapid test for Covid-19, conditional on (i) feeling healthy, (ii) having flu-like symptoms, (iii) having Covid-19 like symptoms.

2. Elicit real WTW to gain a voucher for a free rapid test for Covid-19.

3.1 Experiment data

Data collection took place over a week, from 11 till 18 December 2020. The majority of the responses were collected online, via the *QualtricsTX* platform. To enable greater representativeness of the sample, 94 responses (16%) from elder people (median age = 63) were collected using phone interviews. Out of 608 participants starting the online study, 24 (4.7%) dropped out mostly after the first few questions, 3 did not report age, resulting in the final sample of 575 observations. Median age for the sample was 39 years (median for Greece

45.6), and the age distribution is shown in the appendix.

Subjects were recruited from the database paignia.net and invited to participate in a study, answering a few question on behavior. Upon signing up for the experiment, they and signing a consent form, the participant was first asked about general and Covid-19-related health. We then elicited hypothetical willingness to wait (WTW) to take a rapid test for Covid-19, conditional on (i) feeling healthy, (ii) having flu-like symptoms, (iii) having Covid-19 like symptoms. For all three hypothetical scenarios, the test was being offered by the national health authority (EODY) while the participant was walking down the street (this is a procedure actually happening and discussed on popular media, so they should be well familiar with it). The hypothetical location of the participant was chosen to eliminate the (hypothetical) travel costs and reliability-related concerns. After eliciting the hypothetical WTW, we asked the subjects several control questions, including exposure to Covid-19 risky environments (e.g. taking public transport or working fate-to-face with many people) and socio-demographics. After completing the compulsory part of the study, the participants were randomly allocated to one of the two treatments. In treatment *Test*, the participant would be offered a 1/30 chance lottery for a voucher for a home-administered Covid-19 test, worth €80 at the time of the study[2]. In the baseline treatment *Book*, the participant would be offered the same 1/30 chance lottery for a voucher for the local large-scale bookshop chain ("Public"), which we also set to €80 value, for comparability[3]. Crucially, the participant had to complete a real-effort task to enter the lottery, and we made it clear that the part was optional and would only need to do it if they wanted to enter the lottery for the prize. Participants were also reminded that they could stop the waiting task and leave at any moment.

All 575 participants completed the hypothetical elicitation and the control questions (left part of Figure 2).

[2]For both prizes, the delivery was guaranteed within next 36 hours.

[3]Evidence shows that people tend to value a high stakes lottery much higher than a certainty equivalent of its expected value (Kachelmeier and Shehata, 1992)

As was partly expected, a substantial part of the sample (n=174) did not continue to the optional task. A major part of it (n=78) was the elder people subsample. We are not very concerned that the inconvenience of the waiting task over the phone was the issue, since the participants came from the sample that had participated about a month ago in an unrelated study involving a real effort task over the phone. For n=38 participants, a software glitch in Qualtrics, in the first five hours of the study resulted in missing recording of the treatment allocation, so we had to drop their data despite completion of the optional task.

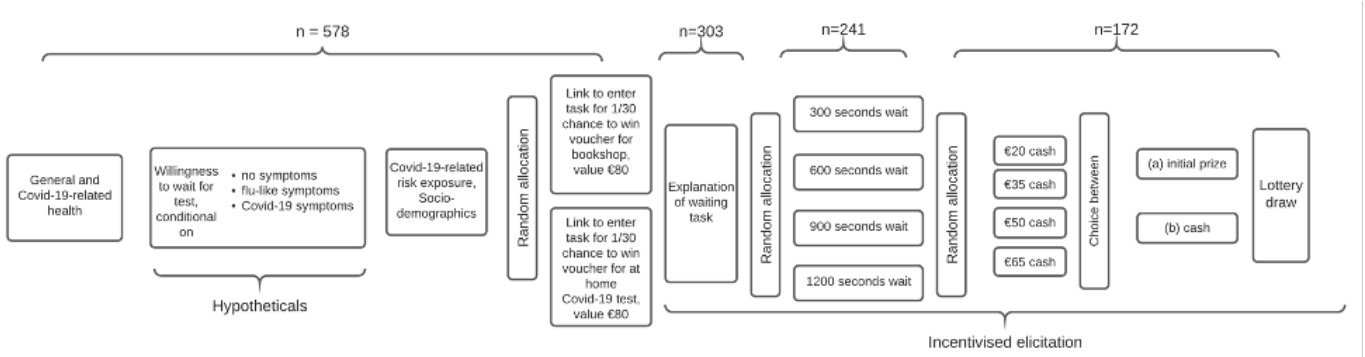

Figure 2: Experimental Flow.

The participants then read the description of the optional task. They learned that it involved waiting in front of their screen for some time (target) that would be revealed in the next screen, and the lottery draw for the prize would take place right after the wait. They also learned that to ensure that they are waiting, a button would appear at random times and they would need to press it within 4 seconds to avoid being disqualified. Among the 303 participants who read the description of the optional task, 241 continued to the next screen which revealed the waiting target. At this stage, they were randomly allocated to one of the four *Wait* target conditions $\{300, 600, 900, 1200\}$ seconds. Upon learning the *Wait* time, further 59 participants dropped out instantly (median target time 900 seconds). Among the 241 waiting, 69 dropped out before completing the target (median $Wait = 900$ seconds). In total, 172 participants completed the waiting target (median $Wait = 600$ seconds).

Upon completing the waiting task, each participant was randomly allocated to one of the four Cash conditions, {€20,€35,€50,€65}. The participant was offered a choice to enter the

lottery for: (a) the original prize ($Book, Test$), or (b) the displayed Cash amount. Out of the 172 participants, 112 chose to swap the original prize for the cash amount, whilst 60 chose to stay with the original prize (median cash value €35 for both). A total of 7 participants won the lottery.

4 Experiment Results and Prevalence Bias

4.1 Impact of self-selection on the bias in prevalence measurement

4.1.1 Hypothetical

We find heterogeneity of waiting times between the age groups, driven by the self-assessed symptoms (Table 1). Younger people tend to behave similarly to elder people, while people between 30 and 50 are willing to wait the least time. This can be reconciled with the fact that this age group has the highest employments rates and possibly family obligations, leading most probably to the least available free time.

Age group	WTW— No Symptoms	WTW — Flu Symptoms	WTW — Covid-19 Symptoms	N
Under 30	0.156	0.391	0.641	192
30-50	0.094	0.279	0.558	222
50+	0.161	0.373	0.596	161

Table 1: Proportions of respondents reporting willingness to wait for a Covid-19 PCR test for over half an hour

Table 2 shows the testing bias, as calculated by the ratio of willingness to test between people with symptoms and those without. The figure ranges between 1.5 and 42, depending on the age group and waiting times. People under 30 with symptoms are 1.5 times more likely to test when there is no waiting time, compared to those without symptoms. This figure increases to 2.74 when there is a short wait of 5-15 minutes; 4.10 with a 15-30 minute wait; 11.67 with a 30-60 minute wait and 42 with a 1-2 hour wait. The ratio for 30-50 year-olds ranges between 1.50 for no wait and 17.33 for a 1-2 hour wait. For over 50-year-olds,

the ratio ranges between 1.66 and 9.4. The symptom-conditional difference is significant at $p < 0.01$, see appendix for details.

Age group	No wait/ Immediate	5-15 min	15-30 min	30-60 min	1-2h	over 2h	N
Under 30	1.50	2.74	4.10	11.67	42.00	42+	192
30-50	1.50	3.29	5.91	9.57	17.33	17.333 +	222
50+	1.66	2.67	3.69	7.62	9.4	9.4 +	161
Total	1.54	2.91	4.46	9.43	15.67	15.67 +	575

Table 2: Bias (ratio of people with Covid-19 symptoms to people with no symptoms) by hypothetical waiting time for rapid test, N=575

The propensity to test bias, translates to a biased virus prevalence estimate β, according to the calculations in Section 2. The prevalence bias is also time varying, even with no changes in testing strategies. It depends, crucially, on symptom prevalence, which, given the exponential spread of Covid-19, can change drastically in a short period of time. This means that the estimate depends on symptom prevalence, but the bias itself also depends on it – so the bias is time variant.

4.1.2 Incentivized Elicitation

Apart from waiting times, self-selecting into testing also depends on the cost associated with it (if applicable – costs can vary from time to monetary value, travel etc). We test whether the hypothetical willingness to wait to take a Covid-19 test correlates with the incentivized real waiting time for the 1/30 lottery and find a significant positive relationship between the two ($p < 0.05$).

Also, we measure willingness to pay for the test (see Table 5 in the appendix). Of those who won a test voucher, 83.8% swapped it for cash, as opposed to 48.9% of those who won the book voucher, indicating that the majority of subjects would not be willing to pay to receive a test.

Note that there were too few people reporting no symptoms to be able to compare the willingness to pay of people with symptoms, to those without. The scope of this study is

to measure and correct the bias for free tests subject to different waiting times, and further experiments are needed to explore the effect of other monetary and non-monetary costs.

4.2 Impact of the Bias on Prevalence Estimates - Calculations and Demographic Simulation

We have launched an online calculator that provides estimates on the testing bias (available at http://georgana.net/sotiris/task/atten/covid.php), as described in Section 2. The estimates on the testing bias depend on (a) the percentage of tests yielding positive results; (b) the percentage of the general population that reports symptoms; (c) the relative likelihood of having Covid-19 for those with symptoms compared to those without symptoms; and (d) how more likely are people with symptoms to self-select into testing than those without symptoms. According to our methodology, it is possible to calculate these figures and thus estimate the bias. Parameter (a) is provided by the results of community testing; (b) is provided by surveying; (c) can be obtained by asking people a simple question before testing them for Covid-19; and (d) is provided by surveying.

A simple example is the following: Assume community testing led to 10% positive results, and 5% of the population reported symptoms. Without waiting time, if those with symptoms are 5 times more likely to have the virus than those without symptoms, then the results of community testing exaggerate by 27.71%, and the true prevalence in the population is 7.83% (instead of the reported 10%). At a 30-60 minute waiting time, the bias increases to 106.95%, meaning that the true prevalence in the population is 4.83%.

To further illustrate our results, Figure 3 depicts our best estimate of the virus prevalence bias, i.e. the ratio between reported prevalence and actual, depending on symptoms prevalence and waiting time, for the three age groups.

Based on these estimates, we can simulate how different demographic structures would affect the prevalence bias. In Figure 4 we depict the results from 3 million draws from the plausible parameter space (we assume symptoms prevalence of 5%, and allow the testing bias

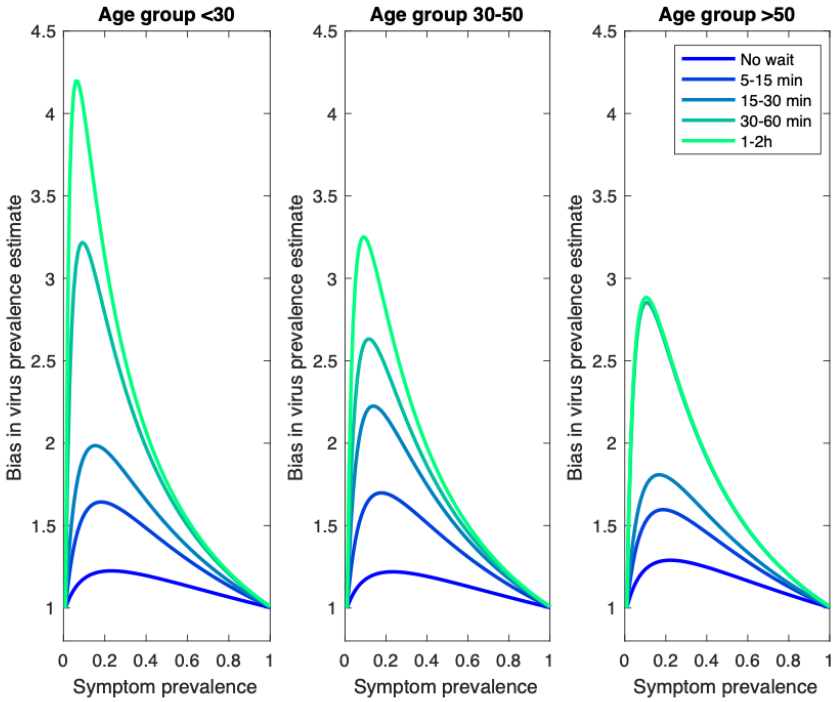

Figure 3: Best estimate of the virus prevalence bias: The ratio between reported prevalence and actual, depending on symptoms prevalence and waiting time, for the three age groups.

parameter to vary uniformly within the 95% confidence interval gained from the experiments) applied to three countries, with different demographic structures: Nigeria (with one of the youngest populations globally), Italy (heavily ageing population) and the USA (between the two extremes). The simulation shows that demography matters: a young country like Nigeria could have a substantially higher prevalence bias than Italy. However, it is also clear that the waiting times are more important than demographics. Lowering waiting times would result in a low bias for all countries.

5 Debiasing vs Polling for Prevalence Estimation: Validation using Existing Data

Debiasing the field prevalence numbers can be performed using our methodology, as long as there are good estimates for four parameters, namely (a) the percentage of tests in the field

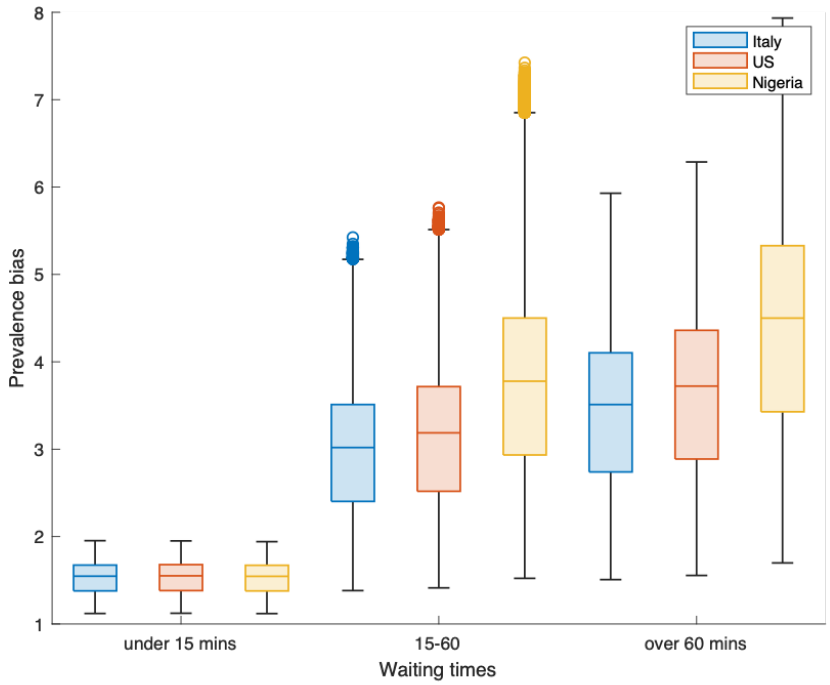

Figure 4: Simulation of the effect of different demographic structures on the prevalence bias.

yielding positive results; (b) the percentage of the general population that reports symptoms; (c) how more likely are people with symptoms to be carrying the virus than those without symptoms; and (d) how more likely are people with symptoms to self-select into testing than those without symptoms. Obtaining estimates for the above parameters is of varying difficulty: (a) is obtained in any country doing "random" street testing (it is important to keep track of important factors, such as waiting times though), (b) can be estimated with standard polling and (d) can be estimated with our experimental methodology. Estimating (c) would require asking subjects at testing stations to self-report their symptoms before testing.

We suggest however a novel, more economical and accurate alternative for prevalence estimation. The important parameter to estimate is the probability of having covid-19 conditional on having symptoms, and on not having symptoms, similar to parameter (c) above. This can be done by asking a simple question at existing testing sites (indeed we

have ongoing parallel work underway to obtain these estimates in cooperation with testing centres in the field). These parameters could be country-specific and time-variant, but we do not expect changes to be too fast. Obtaining a few estimates in each virus season could suffice, and this estimate could be used for many similar countries. The next step is unusual in the context of the pandemic: poll a representative sample regularly, *to obtain symptoms prevalence*. A common misunderstanding involves the argument that laymen cannot measure their symptoms properly. This is not a bug, but a feature of our procedure. Since the testing bias depends on self-reported symptoms, we need to condition on subjects *believing* they have symptoms, not on actually having them. Using both steps above can yield accurate prevalence estimates in real time at very low, comparatively, cost.

In the following we try simulate the novel polling method and compare to data that are as accurate as possible. That is, we need a benchmark figure to approximate true prevalence, derived by a study that does not suffer as much from the self selection bias. We use the REACT study in England (REACT, 2020) and the ONS Infection Survey, which are to our knowledge the two studies, that likely do not suffer from an inordinately high testing bias.[4] REACT is done on a large sample of all ages and locations, and importantly non response is relatively low[5]

REACT has been conducted in eight waves, to date. Two of the waves have been published in two sub-waves, yielding 10 different observations (we match the ONS data to these dates). The data consists of non-overlapping random samples of the population of England at lower-tier local authority level (LTLA, n=315) that were invited to take part in each round of the study based on the National Health Service list of patients.

[4]Both studies are aiming to test large, representative samples at home. An important difference is that REACT sends testing kits to homes, and participants can choose to self-test and send back the results, while ONS sends health workers to test citizens at home. It is not clear without further research which method leads to a lower bias.

[5]The overall conversion rate from invitation letter to registration for a swab kit was 23.8% (1,474,824 registrations from 1,474,824 invitations sent), but this is not the main problem, because symptoms at this stage would not last for the duration of the study. Self-selection leading to bias can happen when people receive swabs and decide to test or not. The average proportion of swabs returned was 74.6% (1,100,270 swabs returned from 1,474,824 kits sent). There was some variation in response rates between rounds.

For those registering to take part, a swab kit was sent to a named individual with a request to provide a self-administered throat and nose swab (or for a parent/guardian to obtain the swab for children aged 12 years or younger). The participant was requested to refrigerate the sample and order a courier for same or next day pick-up and transporting to the laboratory for RT-PCR. There is no obligation to take the test at any stage of the process. Participants then completed an online questionnaire (or telephone interview) giving information on history of symptoms, health and lifestyle. The publicly available data includes the raw figures on tests and outcomes, as well as unweighted prevalence estimates and estimates weighted to be representative of the population of England as a whole.

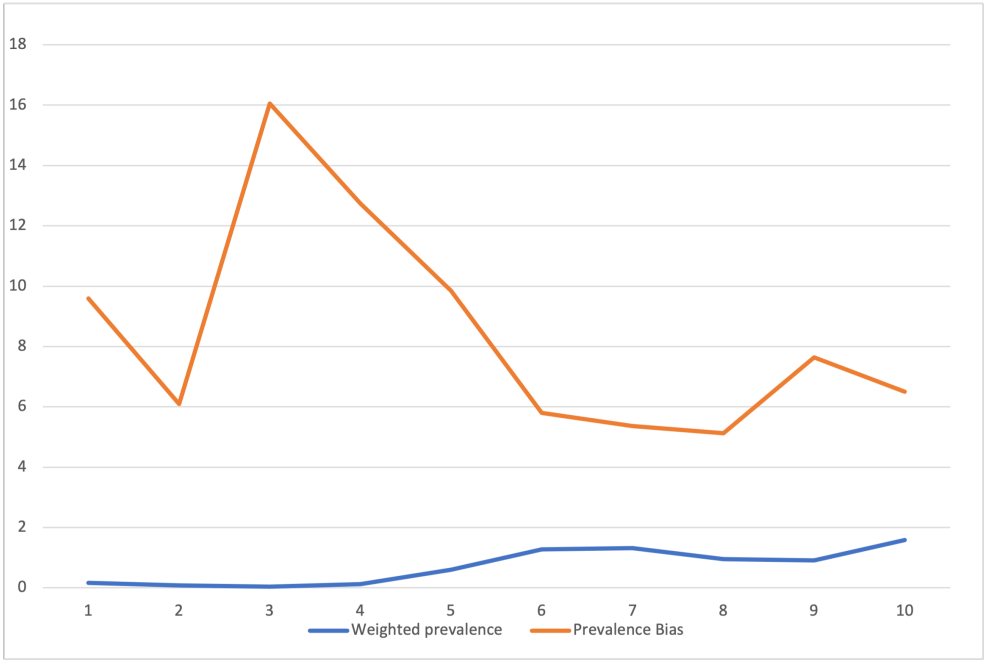

Figure 5: Estimate of the prevalence bias in field testing. Test positivity divided by the best prevalence estimate using REACT and ONS data.

We focus on the weighted prevalence figures, as the most accurate and take the simple average of the two surveys to get our best prevalence estimates. The number of daily tests is publicly available, along with the number of tests being positive, yielding test positivity. We divide test positivity by the best prevalence estimate to obtain an estimate of the prevalence bias in field testing.

From our calculations and the experiment, three main hypotheses follow regarding the prevalence bias:

1. Test-positivity is always inflated due to self-selection, meaning the prevalence bias is large.

2. The prevalence bias is time-varying

3. As virus prevalence in the population increases, so does the bias in its measurement (for reasonable prevalence ranges in the covid-19 pandemic)

As presented in figure 5, in the 10 different subwaves of the study, the estimated prevalence bias indeed is positive, substantial, but also highly variable, ranging from 3.8 to 23.6, thus confirming our two main predictions. Apart from the first waves, during which the testing strategy was changing, complicating comparisons, it seems there is a weak effect for the bias to be rising in prevalence. A proper test of this hypothesis would require more waves and a constant testing strategy.

In the next graph we compare the best estimates of positivity with the two methods used currently to proxy prevalence, field positivity and case counts (as a percentage of the country's population), along with the our two new methods, the debiasing estimate and a simulation of the polling method.

We simulate the polling method by taking symptom conversion parameters, as published in REACT, but from the immediately preceding wave. The symptoms prevalence numbers we use are then from the current wave. As long as agencies can get a polling estimate that is similarly accurate to REACT, this simulation places a lower bound on the accuracy of the polling method.

We find that the polling method is consistently closest to "true prevalence", while the debiasing estimate is further away and still inflates actual prevalence to some extent. As shown before, field positivity is an order of magnitude higher in most waves, while recorded cases are underestimating prevalence by at least an order of magnitude. Even assuming that cases sum up over several days to 10 times the daily rate, this estimate is still many times

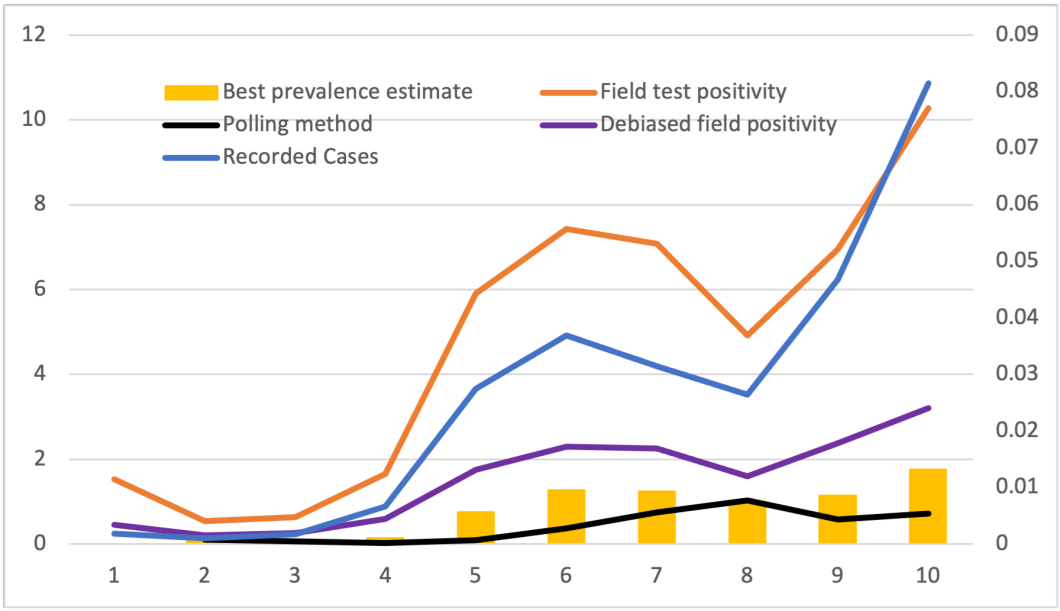

Figure 6: Comparison of the various prevalence estimates, in percentage of the UK population.

lower than estimated prevalence. Also note that all traditional methods are very variable, for example recorded cases increase almost fivefold when true prevalence doubles. Again, this is in line with our bias calculations.

A final note on the usefulness of the REACT and ONS methods: the marked difference between their prevalence estimates and common field test positivity, is driven by the fact that the monetary and non-monetary cost of testing happen are much lower in REACT and the ONS Infection Survey. Crucially, participants were able to administer the test and report symptoms without leaving the house. While this is a step in right direction, other significant non-monetary costs need to be mitigated in order to address self-selection bias. For example, for both studies, the physical unpleasantness of conducting the test may still make those not experiencing symptoms more likely to test. While it is possible to reduce other non-monetary costs of *testing*, we believe that making large-scale regular *self-reporting* of symptoms easy would be a more effective step towards achieving accurate prevalence estimates.

6 Application: Do Open Schools Lead to Transmission?

Closed schools cause problems to working parents, besides hindering the education of young pupils who reportedly find it hard to follow remote teaching. Studies have not yet yielded a clear, conclusive answer regarding the epidemic cost of school opening though and the debate remains heated.

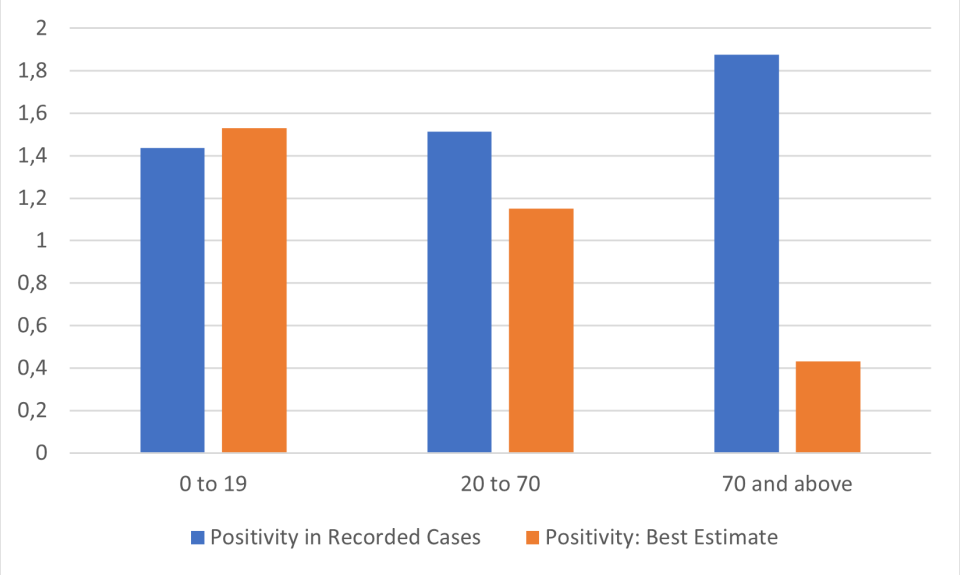

Figure 7: Recorded case positivity by age, vs best estimates.

Understanding the testing bias and how it varies by age group, allows us to reconcile the various pieces of evidence and solve existing puzzles. Looking at case counts, children and youngsters up to 19 years of age, seem not to be major carriers of the disease. Indeed, in a sample of 16 European countries for which data were available, children and teenagers up to 19 are always underrepresented among confirmed cases.[6] Authorities around the world have used this as an argument that school opening is relatively harmless.

However, we know from our experiment that young people are much less likely to test.

[6]The sample includes Belgium, Czechia, Denmark, Estonia, Finland, France, Greece, Germany, Italy, Latvia, Netherlands, Norway, Portugal, Spain, Sweden, Ukraine. The largest percentage was found in Finland and Norway, above 15%, while the lowest were in France, Greece and Spain, at below 7%. For comparison, the population share of 0-19 year olds in a fairly typical country like Germany is 18.7%.

While absolute testing propensities are similar, they are very different between those with symptoms and those without. Combined with the fact that the young have a much lower symptoms prevalence (Qiu et al., 2020; Kelvin and Halperin, 2020), they test much less frequently.[7]

As a consequence of the testing bias, the young are underrepresented in testing, meaning their cases are underreported. Indeed, looking at the data from the ONS Infection Survey in the UK, high-school children seem to have the highest prevalence of all (see figure). This example illustrates the importance of the selection-bias: how it complicates comparisons of prevalence in different age groups and can lead to wrong, in this case missing, pandemic prevention interventions.

7 Discussion

Using an incentivised online experiment, we found that the probability of taking a Covid-19 test for those who have symptoms (or believe they are more likely to have caught the virus) is many times higher than those who do not. In our sample, this testing propensity bias ranged from 1.5 times (for people under 30 years with no waiting time) to 42 times (for people under 30 and a 2-hour waiting time). The bias becomes larger with longer waiting times, and any cost associated with taking the test. Testing stations cannot readily correct this by oversampling (i.e. selecting people without symptoms to test).

A person's age also influences the testing propensity bias, which means that different areas (or countries) will have different biases depending on the age composition. Furthermore, there have been reports of very long waiting times in some cases of community testing, which greatly exacerbates the bias and makes comparisons even within a country hard. Lastly, even keeping everything else constant the bias depends strongly on the actual virus prevalence. All these effects combined mean the bias is very likely to be varying across space

[7] Additonally, there seem to be reasons strictly related to the test itself that contribute to bias, due to the the under-detection of Covid-19 positivity in children, compared to that of adults (Dattner et al., 2020).

and time.

Our findings imply that virus positivity results from community testing sites are heavily biased. Contrary to conventional wisdom in the health policy community that suggested the bias would be, if anything, downward, our results suggest that prevalence is inflated by up to 5 times, even if tests are not costly.

We recognise the importance of giving people the opportunity to test, as this identifies positive cases, thus allowing them to self-isolate and stop spreading the disease. If the goal of street testing is just to allow random people to have a quick and free test, then this possibly meets its goal. Note, however, that random testing is not efficient, economically, or epidemiologically: subsidising tests specifically for populations with a high risk of getting infected and infecting others would probably save more lives at lower cost (say, tests for young people working in service industries and living with their parents). These questions remain open for future research.

What we have shown is that "random" voluntary testing is not really random. As such, it does not provide accurate information on disease prevalence, which is important to design and implement urgent policy responses to the pandemic, in terms of type, intensity and geographic area. Since voluntary testing is always biased, aggregate results on prevalence should be corrected. We have explained a method to do such debiasing. Note that debiasing can be useful to get better estimates of prevalence in real time, but also to correct the past time series that are used to estimate and calibrate many models related to the pandemic. The object of such studies ranges from the effectiveness of measures against the pandemic (Brauner et al., 2020; Hsiang et al., 2020), to health outcomes and economic effects. Furthermore, the probability to test is recognised as an important parameter in macroeconomic models evaluating economically optimal lockdown strategies (Alvarez et al., 2020),

Our methodology is not limited to correcting the results of community testing. We showed that the number of confirmed cases reported daily is also biased, strongly downward in this case. People might not test because of costs, or the inconvenience of going to a testing

site, or even due to being afraid of losing income. According to our results, more than 85% of the people who are not feeling any symptoms, would not wait more than 30 minutes (a likely time in many street testing procedures) to have a test, even if it is provided free of charge. For people feeling symptoms the estimated percentage of non-testers is still about 40%. These percentages rise even further when tests have a non-negligible cost to the citizen.

Using polling results from a representative sample can correct the error both in recorded cases and field test positivity. Our proposed method is more accurate than these traditional proxies. Moreover the polling method is not costly, and does not require an extraordinary testing capacity, which means it can be used daily, allowing real-time prevalence estimation in myriads of communities worldwide.

The REACT study in the UK (along with the ONS Survey) is an interesting special case of large-scale community testing on a nationally representative sample. The authors claim that this sample is truly random. While we use REACT data as the best estimate we have, our experimental results suggest the sample might still not be truly random. Even for people taking a free test at home (compare to the no waiting time condition in the experiment), a substantial testing bias exists. Importantly, REACT is also very expensive to run, while simultaneously less timely than our polling proposal. REACT has been done monthly or less often, while our procedure can be run daily.

This paper also contributes to the literature on testing regimens (Mina et al., 2020). Mass testing, extending to a very large part of the population, is useful as it can provide more accurate figures, and also identifies positive cases. It has been used, among others, in Liverpool, Slovakia and South Korea (Pavelka et al., 2020; BBC, 2020; Bloomberg, 2020; Brauner et al., 2020). However, mass testing is extremely expensive, and might be infeasible, especially at frequent intervals, due to capacity and technical constraints.

In the absence of mass testing, obtaining unbiased prevalence estimates is of paramount importance for health and the economy. Underestimating disease prevalence can trigger inadequate measures and further spread of disease, while overestimating can be detrimental

to economic activity. We thus urge policy makers to redesign "random" testing as a matter of priority in the effort to tackle the pandemic.

As a final note, our methodology is applicable to the prevalence measurement of any epidemic, when carriers have informative private information about their health status. Fighting disease is hard, even without the added complication of not knowing the location and magnitude of the fight. Our work offers tools to measure prevalence in real time. Further work is needed, to estimate specific selection-bias parameters for every disease, as they are necessarily related to the health burden and life expectancy reduction caused by the specific pathogen.

References

Flavio A Cadegiani, John A McCoy, Carlos Gustavo Wambier, and Andy Goren. Clinical diagnosis of COVID-19: A prompt, feasible, and sensitive diagnostic tool for COVID-19 based on a 1,757-patient cohort (The AndroCoV Clinical Scoring for COVID-19 diagnosis). *medRxiv*, pages 2020–12, 2021.

Steven Riley, Kylie EC Ainslie, Oliver Eales, Caroline E Walters, Haowei Wang, Christina Atchinson, Claudio Fronterre, Peter J Diggle, Deborah Ashby, Christl A Donnelly, et al. React-1 round 8 interim report: SARS-CoV-2 prevalence during the initial stages of the third national lockdown in England. *medRxiv*, 2021.

Andrew Atkeson. How deadly is COVID-19? Understanding the difficulties with estimation of its fatality rate. Technical report, National Bureau of Economic Research, 2020.

Jan M. Brauner, Sören Mindermann, Mrinank Sharma, David Johnston, John Salvatier, Tomáš Gavenčiak, Anna B. Stephenson, Gavin Leech, George Altman, Vladimir Mikulik, Alexander John Norman, Joshua Teperowski Monrad, Tamay Besiroglu, Hong Ge, Meghan A. Hartwick, Yee Whye Teh, Leonid Chindelevitch, Yarin Gal, and Jan Kulveit. Inferring the effectiveness of government interventions against COVID-19. *Science*, 2020. ISSN 0036-8075. doi: 10.1126/science.abd9338. URL `https://science.sciencemag.org/content/early/2020/12/15/science.abd9338`.

ECDC. Covid-19 testing strategies and objectives. Technical report, Available at: https://www.ecdc.europa.eu/sites/default/files/documents/TestingStrategy-Objective-Sept-2020.pdf, 2020.

ECDC. Data on testing for Covid-19 by week and country. Technical report, Available at: https://www.ecdc.europa.eu/en/publications-data/covid-19-testing, 2021.

Neil Pearce, Jan P Vandenbroucke, Tyler J VanderWeele, and Sander Greenland. Accurate statistics on COVID-19 are essential for policy guidance and decisions, 2020.

Charles F Manski and Francesca Molinari. Estimating the COVID-19 infection rate: Anatomy of an inference problem. *Journal of Econometrics*, 220(1):181–192, 2021.

Sharon K Greene, Sarah F McGough, Gretchen M Culp, Laura E Graf, Marc Lipsitch, Nicolas A Menzies, and Rebecca Kahn. Nowcasting for real-time COVID-19 tracking in New York City: An evaluation using reportable disease data from early in the pandemic. *JMIR Public Health and Surveillance*, 7(1):e25538, 2021.

Rebecca L. Thornton. The demand for, and impact of, learning HIV status. *American Economic Review*, 98(5):1829–63, December 2008. doi: 10.1257/aer.98.5.1829. URL `https://www.aeaweb.org/articles?id=10.1257/aer.98.5.1829`.

Russell Golman, David Hagmann, and George Loewenstein. Information avoidance. *Journal of Economic Literature*, 55(1):96–135, 2017.

Emily Oster, Ira Shoulson, and E Dorsey. Optimal expectations and limited medical testing: Evidence from Huntington disease. *American Economic Review*, 103(2):804–30, 2013.

Elisabeth Mahase. Covid-19: Mass testing in Slovakia may have helped cut infections. 2020.

Nils Haug, Lukas Geyrhofer, Alessandro Londei, Elma Dervic, Amélie Desvars-Larrive, Vittorio Loreto, Beate Pinior, Stefan Thurner, and Peter Klimek. Ranking the effectiveness of worldwide COVID-19 government interventions. *Nature Human behaviour*, 4(12):1303–1312, 2020.

Kelly Osezele Elimian, Chinwe Lucia Ochu, Blessing Ebhodaghe, Puja Myles, Emily E Crawford, Ehimario Igumbor, Winifred Ukponu, Adobola Olayinka, Olusola Aruna, Chioma Dan-Nwafor, et al. Patient characteristics associated with COVID-19 positivity and fatality in Nigeria: Retrospective cohort study. *BMJ open*, 10(12):e044079, 2020.

Steven Riley, Kylie EC Ainslie, Oliver Eales, Benjamin Jeffrey, Caroline E Walters, Christina J Atchison, Peter J Diggle, Deborah Ashby, Christl A Donnelly, Graham Cooke, et al. Community prevalence of SARS-CoV-2 virus in England during may 2020: React study. *medRxiv*, 2020.

Rutger A Middelburg and Frits R Rosendaal. Covid-19: How to make between-country comparisons. *International Journal of Infectious Diseases*, 96:477–481, 2020.

Jérôme Adda. Economic activity and the spread of viral diseases: Evidence from high frequency data. *The Quarterly Journal of Economics*, 131(2):891–941, 2016.

Theologos Dergiades, Costas Milas, and Theodore Panagiotidis. Effectiveness of government policies in response to the COVID-19 outbreak. *Available at SSRN 3602004*, 2020.

Chen Shen, Derrick VanGennep, Alexander F Siegenfeld, and Yaneer Bar-Yam. Unraveling the flaws of estimates of the infection fatality rate for COVID-19. *Journal of Travel Medicine*, 2021.

Steven J Kachelmeier and Mohamed Shehata. Examining risk preferences under high monetary incentives: Experimental evidence from the People's Republic of China. *The American Economic Review*, pages 1120–1141, 1992.

REACT. React-1: Summary of sample extraction and fieldwork dates, and response rates, by round. Technical report, 2020. URL `https://www.imperial.`

ac.uk/media/imperial-college/institute-of-global-health-innovation/REACT1-Fieldwork-info-table-for-Imperial-website_with-Round-7.pdf. Retreived on 2021-01-30.

Haiyan Qiu, Junhua Wu, Liang Hong, Yunling Luo, Qifa Song, and Dong Chen. Clinical and epidemiological features of 36 children with coronavirus disease 2019 (COVID-19) in Zhejiang, China: an observational cohort study. *The Lancet Infectious Diseases*, 20(6): 689–696, 2020.

Alyson A Kelvin and Scott Halperin. Covid-19 in children: the link in the transmission chain. *The Lancet Infectious Diseases*, 20(6):633–634, 2020.

Itai Dattner, Yair Goldberg, Guy Katriel, Rami Yaari, Nurit Gal, Yoav Miron, Arnona Ziv, Yoram Hamo, and Amit Huppert. The role of children in the spread of COVID-19: Using household data from Bnei Brak, Israel, to estimate the relative susceptibility and infectivity of children. *medRxiv*, 2020.

Solomon Hsiang, Daniel Allen, Sébastien Annan-Phan, Kendon Bell, Ian Bolliger, Trinetta Chong, Hannah Druckenmiller, Luna Yue Huang, Andrew Hultgren, Emma Krasovich, et al. The effect of large-scale anti-contagion policies on the COVID-19 pandemic. *Nature*, 584(7820):262–267, 2020.

Fernando E Alvarez, David Argente, and Francesco Lippi. A simple planning problem for Covid-19 lockdown. Technical report, National Bureau of Economic Research, 2020.

Michael J. Mina, Roy Parker, and Daniel B. Larremore. Rethinking Covid-19 test sensitivity — a strategy for containment. *New England Journal of Medicine*, 383(22):e120, 2020. doi: 10.1056/NEJMp2025631. URL https://doi.org/10.1056/NEJMp2025631.

Martin Pavelka, Kevin van Zandvoort, Sam Abbott, Katharine Sherratt, Marek Majdan, Pavel Jarcuska, Marek Krajci, Stefan Flasche, Sebastian Funk, CMMID COVID-19 working group, et al. The effectiveness of population-wide, rapid antigen test based screening in reducing SARS-CoV-2 infection prevalence in Slovakia. *medRxiv*, 2020.

BBC. Covid: Mass testing in Liverpool sees 'remarkable decline' in cases. Available at: https://www.bbc.com/news/uk-england-merseyside-55044488 accessed 19 december 2020., 2020.

Bloomberg. Seoul's full cafes, apple store lines and show mass testing success. Available at: https://www.bloomberg.com/news/articles/2020-04-18/seoul-s-full-cafes-apple-store-lines-show-mass-testing-success accessed 19 december 2020., 2020.

Appendix A. Supplementary Figures

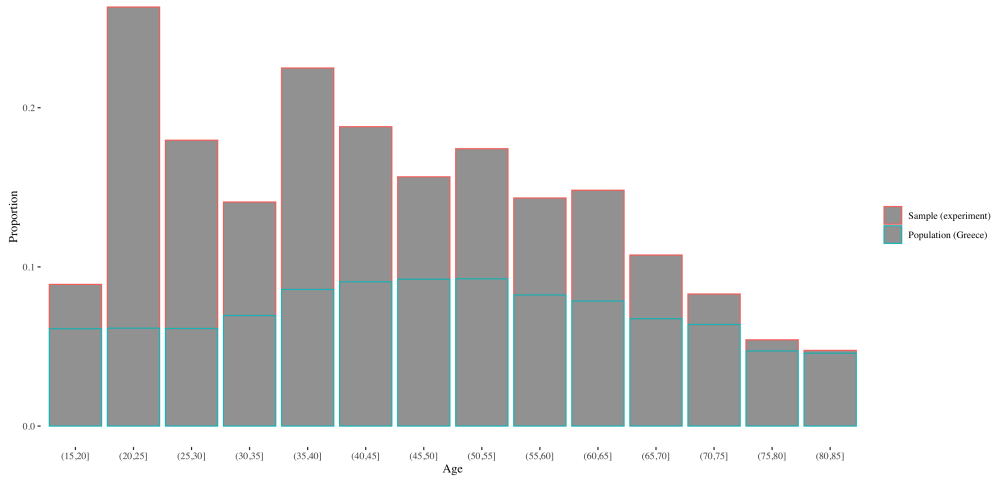

Figure 8: Age distribution in the experiment (n=578) and in population of Greece (source: populationpyramid.net).

Appendix B. Descriptive Statistics and Hazard Models for the Experiment

	Mean	Std dev	min	max
Covid symptoms (0=no; 1=yes)	1.38	0.12	0	1
Age	40.404	15.302	18	84

Table 3: Summary statistics of sample demographics and symptoms.

Hypothetical willingness to test (N=575)		
	By symptoms	**By waiting time**
	No Symptoms	
Mean (SD)	2.96 (1.48)	2.39 (2.04)
Median (Min, Max)	3.00 (1.00, 5.00)	2.00 (0, 8.00)
	Flu Symptoms	
Mean (SD)	2.00 (1.20)	3.81 (2.26)
Median (Min, Max)	2.00 (1.00, 5,00)	4.00 (0, 8.00)
	Covid Symptoms	
Mean (SD)	1.46 (0.951)	5.19 (2.35)
Median (Min, Max)	1.00 (1.00, 5,00)	5.00 (0, 8.00)

By-symptoms key: 1: certainly yes; 2: probably yes; 3: maybe; 4: probably no; 5: certainly no

By-wait-time key: : 0: would not wait at all; 1: would only take it if available immediately; 3: 5 - 15 minutes; 4: 15 - 30 minutes; 5: 30 - 45 minutes; 6: up to an hour; 7: 1 - 2 hours; 8 over 2 hours

Table 4: Summary statistics for hypothetical willingness to wait to take the test, by symptoms and waiting time.

Prize	Not entered	Dropped upon learning waiting time	Dropped after some wait	Swapped prize for cash	Kept prize	N
Book voucher	103	31	38	46	48	266
Test voucher	138	25	31	62	12	268
Bias	1,263	1,267	1.28	4.03	0.25	534

Table 5: Willingness to wait for a 1/30 chance of winning a prize. Number of subjects by level of task completion and incentive (rows 1-2), bias by incentive (row 3).

Table 6: Proportional hazard ratio for dropping out from (hypothetical) wait for a free Covid-19 test, by age group and symptoms.

	Dependent variable:
	Odds of not waiting for Covid-19 test (Reference: Age 30-50 |No symptoms)
Under 30 |No symptoms	−0.114
	(0.099)
Under 30 |Symptoms	−1.342***
	(0.109)
30-50 |Symptoms	−1.279***
	(0.105)
50+ |No symptoms	−0.107
	(0.105)
50+ |Symptoms	−1.403***
	(0.119)
Observations	1,150
R^2	0.264
Max. Possible R^2	1.000
Log Likelihood	−6,177.707
Wald Test	351.360*** (df = 5)
LR Test	352.721*** (df = 5)
Score (Logrank) Test	388.121*** (df = 5)
Note:	*p<0.1; **p<0.05; ***p<0.01