On the Measurement of Disease Prevalence

Sotiris Georganas, Alina Velias and Sotiris Vandoros*

June 23, 2021

Abstract

Accurate prevalence figures are necessary for informed policy against epidemics. "Random" testing is commonly used to estimate prevalence. However, since testing is necessarily voluntary, field estimates suffer from *selection bias*. We document experimentally a strong testing bias: people feeling symptoms are up to 42 times likelier to seek testing. This leads to *prevalence bias*: test positivity can inflate true prevalence fivefold, and this bias changes necessarily over time. We validate using external data and confirm the bias varies intertemporally, making comparisons across time and/or countries misleading. We suggest sampling the population to bypass the bias, yielding more accurate estimates, real-time. Our results are relevant to any epidemic, besides Covid-19, when carrier status informs beliefs.

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

JEL Codes: D90, C81, I18

*We thank Andrew Atkeson, Rebecca Thornton, William Hanage, Ichiro Kawachi, Charles Manski, PJ Healy, and Roberto Weber for useful comments, Fay Angelidou for outstanding research support, and the Department of Economics, City University, for funding the experiments. Georganas: City, University of London, sotiris.georganas.1@city.ac.uk Vandoros: King's College London and Harvard University, vandoros@kcl.ac.uk Velias: City, University of London, Alina.Velias@city.ac.uk

1 Introduction

How to measure prevalence for infectious diseases? In the Covid-19 pandemic, health agencies (e.g. ECDC, 2021) 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 (Acemoglu and Johnson, 2007).

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 and indeed find that positivity is inflated by 3.8 to 23.6 times, depending on the time period. This means that we cannot use a constant adjustment to debias positivity figures. Unlike other biases induced by imperfect medical measurement, the testing bias induces a different inflation factor in different stages of the disease (as illustrated in Section 4.3), making it impossible even to track the speed of the disease spreading, besides measuring the absolute prevalence. 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 economic policy makers and health agencies, using a real example from a European country. During Christmas 2020 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 January 29, 941 cases were recorded, almost double the cases a week before (506). Ignoring statistical issues of significance, two questions arise: is that rise in cases a clear sign of a worsening disease, and what is to blame? Due to the selection bias, even the first question is hard to answer. At the same time as cases rose, testing rose too. On 29 January there was twice the testing than on the 22nd. Actually, *test positivity* is similar between these dates. But the 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 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 test positivity estimates in the field. As long as the characteristics of the testing procedures are known by the health agencies (rare to date), 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]

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.

To understand the relevance of these results, note that policy responses (e.g. social distancing) will inevitably be inefficient if we are not aware of real prevalence, by location and age. Mortality and hospitalisation rates 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, added to the time necessary for 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) similarly takes months (see Brauner et al., 2020 on the effectiveness of pharmaceutical interventions (NPIs), using death counts). 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.

The possibility that infection rates in the untested population can be different than in the tested subsample, has been raised in one of the few papers throughly treating selection issues

[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.

in pandemic measurement, econometrically (Manski and Molinari, 2021). The econometric inference problem in that paper though is more related to epidemic phases during which testing is not widely available, and selection is done from the supply side, while in our paper selection is on the demand side. In (Greene et al., 2021) the measurement issue is also treated econometrically, but instead of considering selection, they use statistical nowcasting. While certainly useful, the accuracy of this method is probably not as high as polling, and detection of trend reversals is not possible in real time.

Experimental, incentivized methods have been used to address important policy issues before (see for example Gneezy et al., 2019). Especially regarding virus testing, there is seminal work measuring 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. Should people be averse to learning they are infected, as information avoidance models suggest (Golman, Hagmann, and Loewenstein, 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 known to happen (Oster, Shoulson, and Dorsey, 2013), prevalence figures would be inflated 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; Hsiang et al., 2020), to compare the effect of government policies *between* countries (Haug et al., 2020; Brauner et al., 2020; Hsiang 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 policies aimed to contain the pandemic (Dergiades, Milas, and Panagiotidis, 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 Theoretical Considerations

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: $v_s = a v_h, t_s = b t_h$, $v_s = a v_h$, $t_s = b t_h$. 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

$$\beta = \frac{abp + 1 - p}{(ap + 1 - p)(bp + 1 - p)} \tag{4}$$

For example, suppose the true symptoms prevalence is 10%, $p = 0.10$. Then $\beta = (0.1ab + 0.9)/(0.1a + 0.9)/(0.1b + 0.9)$. 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, one 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

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

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 at-home test for Covid-19[2].

3.1 Design

To estimate the bias in willingness to test and validate the measure – we design an experiment combining hypothetical and incentive-compatible measures. We first elicit the hypothetical willingness-to-test (WTT) and willingness-to-wait (WTW) to take a test, conditional on (i) symptoms, and (ii) non-monetary cost of testing (in our case, proxied by waiting time). We then validate the measure eliciting the real WTW for a Covid-19 test voucher, where we exogenously vary (i) the waiting time, and (ii) size of the cash alternative – to estimate the participant's valuation of testing. To enable comparisons of the bias across different socio-demographic groups, we asked the subjects relevant control questions, as well as questions on exposure to Covid-19-risky environments (e.g. taking public transport). The full flow of the experiment is presented in Figure 1.

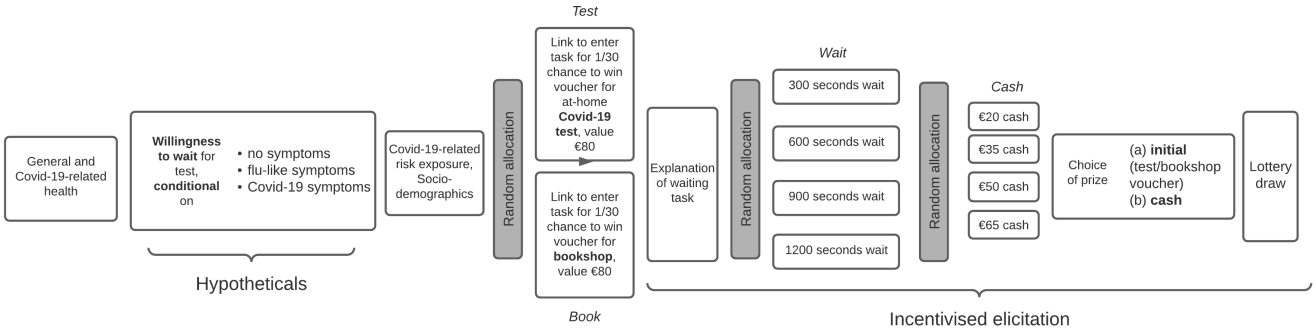

Figure 1: Experiment flow. Italics indicate randomised treatment conditions.

Hypothetical WTW

[2]The service was widely provided by private medical clinics and familiar to the subjects at the time of the study.

To elicit the *hypothetical* willingness to test (WTT) and willingness to wait (WTW) to take a rapid test for Covid-19, *conditional on symptoms*, we presented the participants with the two scenarios of a test being offered by the national health authority (EODY[3]) while the participant was walking down the street (which closely matches the actual procedure implemented in Greece at the time of the study). Scenario 1 asked "How likely would you be to accept the free rapid test if...", Scenario 2 asked "How much time would you be prepared to spend waiting for your turn to take the free rapid test test if..." for the three different self-assessed conditions:

1. ...you did not feel any problem with your health?

2. ...you had some flu symptoms but did not think that they were related to Covid-19?

3. ...you had symptoms that you thought that were related to Covid?

This allowed us to compare the maximum non-monetary cost (as measured in waiting time) the person was willing to incur, depending on their symptoms. The options offered to the participants ranged from "Would not wait at all" to "Over 2 hours", in eight gradual increments.

Incentive-compatible WTW

After completing the questionnaire part of the study, the participants were randomly allocated to one of two lottery prize conditions.[4] In treatment *Test*, the participant was offered a 1/30 chance lottery[5] for a voucher for a home-administered Covid-19 test, worth €80 at the time of the study. In the control treatment *Book*, the participant was offered the same 1/30 chance lottery for a voucher from a large bookshop chain, which we also set to €80 value. The control treatment was designed to measure the baseline WTW for any (non-fungible) prize. Crucially, the participant had to complete a waiting task (methodologically, a real-effort task) to enter the lottery.

[3]The experiment was conducted in Greece.

[4]Delivery of both was guaranteed within 36 hours.

[5]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)

Waiting Task involved waiting in front of the screen for the target amount of time. A button would appear at random times and the participant had to press it within 4 seconds to verify that they were in front of the screen. Upon successful completion of the target wait time, the lottery draw took place. The participants first read the rules of the task, and the next screen revealed their target wait time. In both conditions, we randomly allocated the participant to one of the four *Wait* target conditions $\{300, 600, 900, 1200\}$ seconds. The variation in the waiting times allowed us to measure the willingness to incur the real time cost to receive the prize. To account for the participants who were willing to incur zero waiting cost, we made it clear that the lottery part was optional and they could withdraw at any moment by closing the browser window.

Finally, before the lottery draw we also elicit the cash equivalent of the prize for each participant. Each participant completing the waiting task was randomly allocated to one of the four *Cash* conditions, {€20,€35,€50,€65}. The participant then chose between two lottery options: (a) the original prize $(Book, Test)$, or (b) the displayed *Cash* amount.

3.2 Experiment data

Subjects were recruited from the database paignia.net and invited to participate in a study on general and Covid-related behavior. Data collection took place in Greece, from 11 till 18 December 2020, online and over phone (for participants who were not proficient users of computer, due to senior age). A total of 575 people completed the study (attrition 4.4% [6]). Median age was 39 years (median for Greece 45.6), and the age distribution is shown in the appendix.

[6]Out of 608 participants starting the study, 24 dropped out after the first 5% of the questions, three participants failed to report year of birth

4 Experiment Results and Prevalence Bias

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

4.1.1 Hypothetical willingness to wait for test

Firstly, we find bias in willingness-to-test conditional on symptoms. In our sample, the proportion of people who state that if they do not experience any symptoms they would not take a test even if there is zero waiting cost involved is 1.5 times higher compared to that if they have symptoms (proportions test $p < 0.01$). Secondly, we find heterogeneity in willingness-to-wait for a (free) Covid-19 test, driven by the age group and self-assessed symptoms. Overall, younger people behave similarly to elder people, while people between the age of 30 and 50 are less willing to wait for a test (see left-hand side of Table 1). This is reconcilable with the fact that this group has the highest employment rate and possibly family obligations, leading to less free time.

Conditioning on symptoms shows further divide. Right-hand side of Table 1 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, up to 42 with a 1-2 hour wait. The ratio for 30-50 year-olds ranges between 1.50 and 17.33. 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 Table 3 in the appendix for details.

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 can change drastically in a short period of time.

	Proportion WTW 30+ min, by symptoms			Bias by age and waiting time						
Age	**None**	**Flu-like**	**Covid-19**	**0'**	**5-15'**	**15-30'**	**30-60'**	**1-2h**	**2h+**	**N**
Under 30	0.156	0.391	0.641	1.50	2.74	4.10	11.67	42.00	42+	192
30-50	0.094	0.279	0.558	1.50	3.29	5.91	9.57	17.33	17.333+	222
50+	0.161	0.373	0.596	1.66	2.67	3.69	7.62	9.4	9.4+	161
Average				1.54	2.91	4.46	9.43	15.67	15.67+	575

Table 1: Bias by symptoms and age groups. LHS: Raw proportions of respondents reporting willingness to wait (WTW) for a Covid-19 PCR test for over half an hour. RHS: Bias (ratio of people with Covid-19 symptoms to people with no symptoms), by hypothetical waiting time for the test.

4.1.2 Validation with incentivised waiting task

Apart from waiting times, self-selecting into testing also depends on the cost associated with it (if applicable – costs can vary from time cost to monetary cost of the test, cost of travel etc). We test whether the hypothetical willingness to wait to take a Covid-19 test correlates with the incentivised real waiting time for the 1/30 lottery for the test– and find a significant positive relationship between the two. Specifically, for each extra unit of WTW for the test in the hypothetical condition, the participants waited, on average, 15-24 seconds longer in the waiting task for the real test ($p < 0.05$, Table 2). We also find an overall lower valuation of the Covid-19 test, compared to the Book voucher of the same monetary value. Participants waited on average 69 seconds less for the test, compared to the book voucher ($p < 0.05$).

As for the cash equivalent valuation of the test, among those who had a test voucher as a prize, 83.8% preferred to swap it for (the lower amount of) cash in case of winning the lottery. This proportion was significantly higher to that of participants willing to swap the bookshop voucher for a lower cash amount in case of winning (48.9%, $p < 0.05$ [7]). This evidence is indicative of subjects valuing the test lower than its market cost.

Note that there were too few people reporting any 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.

[7] 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).

Table 2: Hypothetical vs incentivised waiting time for Covid-19 test

	Dependent variable:	
	Real wait time (seconds)	
	(1)	(2)
Prize:Test (Ref:Book)	-69.470^{**}	-69.244^{*}
	(31.321)	(36.983)
Age	-7.643^{***}	-7.908^{***}
	(1.014)	(1.156)
Hypothetical wait time (No symptoms)	15.175^{**}	24.741^{***}
	(7.703)	(8.878)
Income (proxy)		-0.013
		(0.084)
Constant	576.002^{***}	585.612^{***}
	(50.785)	(62.742)
Observations	537	382
R^2	0.112	0.140
Adjusted R^2	0.107	0.131
Residual Std. Error	362.864 (df = 533)	361.139 (df = 377)
F Statistic	22.422^{***} (df = 3; 533)	15.372^{***} (df = 4; 377)
Note:		*p<0.1; **p<0.05; ***p<0.01

4.2 From Testing Bias to Prevalence Bias

To estimate the prevalence bias one needs estimates of symptoms prevalence and parameters a and b. We recommend polling to estimate symptoms prevalence, experiments for b, while a can be obtained by asking subjects at testing stations to self-report their symptoms before testing.

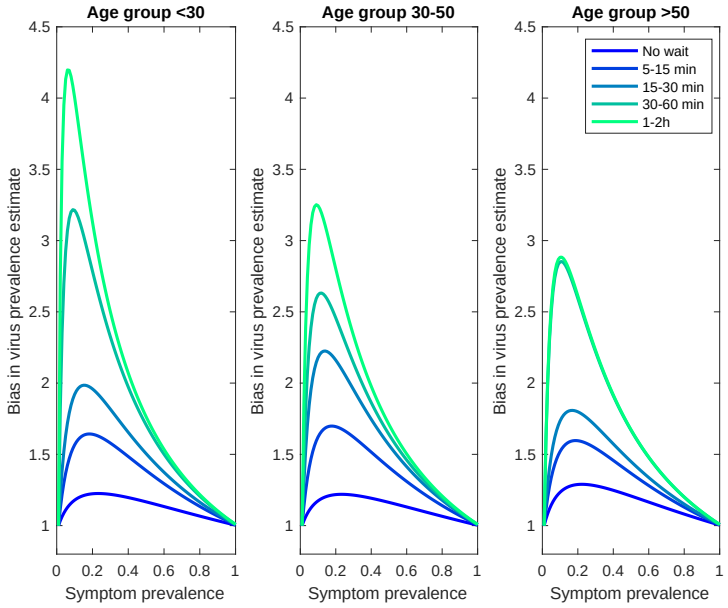

Figure 2: Best estimate of the virus prevalence bias: The ratio between reported prevalence and actual, depending on symptoms prevalence and waiting time, by age group.

Suppose, for example, that community testing reveals 10% positivity, and 5% of the population reported symptoms. If those with symptoms are 5 times more likely to be positive than those without, and waiting time was 0, then the results of community testing exaggerate by 27.71%, and the true prevalence in the population is 7.83%. At a 30-60 minute waiting time, the bias increases to 106.95%, meaning that the true prevalence is 4.83%.

To illustrate our results, Figure 3 depicts our best estimate of the virus prevalence bias depending on symptoms prevalence and waiting time, by age.

Based on these estimates, we can simulate how demography affects the prevalence bias. We use 3 million draws from the plausible parameter space (assuming symptoms prevalence of 5%, and allowing the testing bias parameter to vary uniformly within the 95% confidence interval gained from the experiments) applied to three countries, with different demographic

structures: Nigeria (one of the youngest populations globally), Italy (heavily ageing population) and the USA (between the two extremes). The simulation shows that demography matters: Nigeria could have a substantially higher prevalence bias than Italy. However, waiting times are clearly more important than demographics. Lowering waiting times would result in a low bias for all countries.

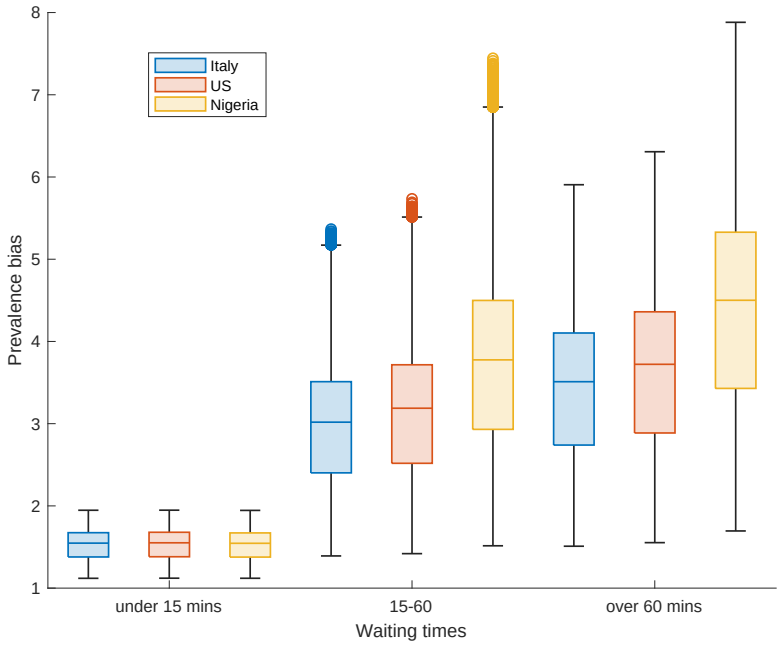

Figure 3: Simulation of the prevalence bias under different demographics.

4.3 The Prevalence Bias over Time

As we have mentioned before, the prevalence bias is necessarily time variant in real world conditions, when epidemics ebb and flow. Even if nothing changes regarding a region's testing strategy and people's individual preferences, the prevalence bias will change substantially over the course of the disease, because the frequency of symptomatic people directly affects the bias (see equation 4 above).

In the following graph, we present a stylised example of a disease that rises exponentially and then subsides (due to non pharmaceutical interventions, vaccinations, or naturally reached immunity), with a subsequent second wave. At first we assume $a = b = 20$, which is likely in initial stages when tests are expensive and/or waiting times high. We then assume

a change of the testing strategy from week 12 to 13, that lowers the testing bias, such that $b = 5$. This change could be induced for example by the availability of free self-tests.

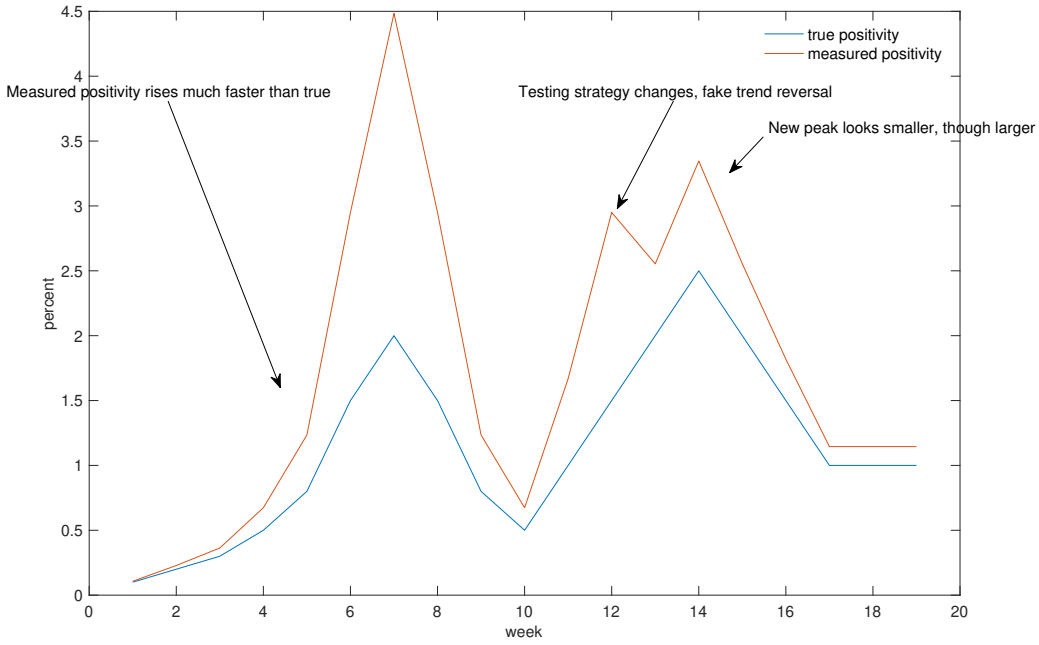

Figure 4: Stylised example of the prevalence bias' time-varying nature.

In the first weeks the prevalence bias is small, because there are too few people feeling symptoms. As the percentage of people with symptoms rises, the prevalence bias rises too, making the prevalence look much higher than it is. From weeks 5 to 7, true prevalence rose slightly more than two times, but measured prevalence rises more than fourfold. This effect, uncorrected for, could lead to excessively strong measures, since policy makers would think the disease is spreading much faster than it actually is.

The change in the testing strategy has two effects: first it leads to an apparent trend reversal, that is not true, and second it makes the second wave of the disease look smaller than the first, while the opposite is true.

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 the relevant parameters, which can be hard.

We suggest 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 (in fact ongoing parallel work is underway in cooperation with testing centres in the field, to obtain such estimates). 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 simulate the novel polling method and compare to data that are as accurate as possible. We use the REACT study in England (REACT, 2020) and the ONS Infection Survey as benchmarks, since these suffer, to our knowledge, from the lowest testing bias.[8]

REACT has been conducted in eight waves, including two sub-waves, yielding 10 different observations (we match the ONS data to these dates). For those registering to take part, a swab kit was sent with a request to provide a self-administered throat and nose swab, and a history of symptoms was also asked. The publicly available data includes the raw figures, as well as estimates weighted to be representative of the population of England as a whole.

We focus on weighted prevalence, 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

[8]Both studies aim to test large, representative samples at home. Importantly, REACT sends testing kits to homes and participants can choose to send back results, while ONS sends health workers to test citizens. REACT non response, after kits are sent, is 74.6%, but unknown for ONS. Further research is needed.

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

In the 10 different sub-waves 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 (see the appendix for details). 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. We use the symptoms prevalence numbers 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.

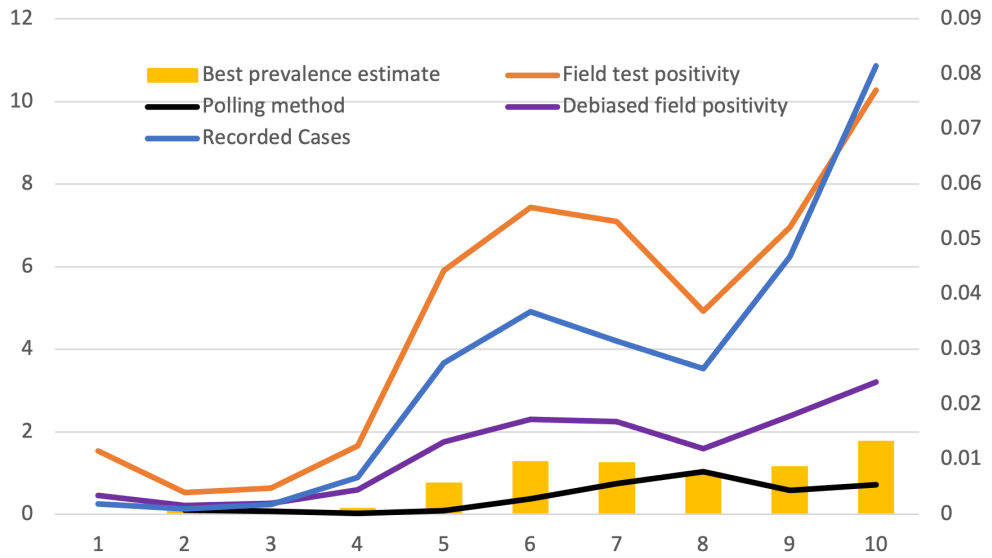

Figure 5: Comparison of the various prevalence estimates, in percentage of the UK population. The recorded cases curve refers to the right hand axis.

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 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.

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.[9] Authorities around the world have used this as an argument that school opening is relatively harmless.

However, our experimental results imply that young people are much less likely to test. While absolute testing propensities are similar, they are very different between those with symptoms and those without. The young have lower symptoms prevalence: conditional on

[9]Finland and Norway had percentages above 15%, while the lowest were in France, Greece and Spain, below 7%. For comparison, the population share of 0-19 year olds in, e.g., Germany is 18.7%.

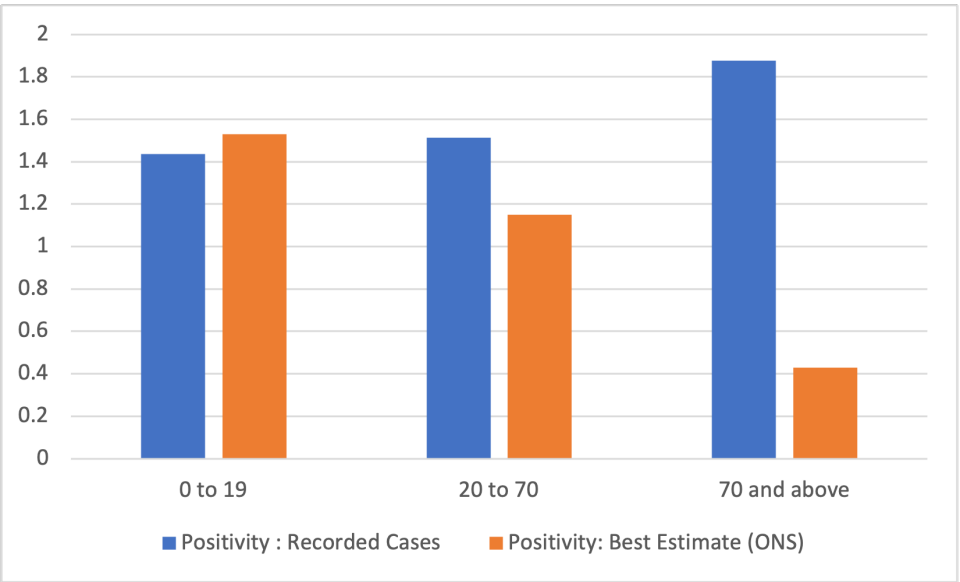

Figure 6: Recorded case positivity by age, vs best estimates.

testing positive, the frequency of people aged 5-17 reporting at least one of the four classic symptoms is 10.5% in REACT wave 8, vs 31.4% for those 18-54 and 22.1% for those 55 and above (also see Qiu et al., 2020; Kelvin and Halperin, 2020),[10] Combined with the large testing bias found experimentally, the different symptoms prevalence means the young test substantially less frequently.

As a consequence of the testing bias, the young are underrepresented in testing, meaning they are underreported in recorded cases. Indeed, although children aged 0-19 have the smallest presence in UK recorded cases (weighted by cohort size), high-school children are estimated to have the highest prevalence of all in the ONS survey (see figure 6).

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

[10] 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).

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 locations will have different biases depending on demography. In line with this, Serra-Garcia and Szech, 2020, find variation in testing demand by race and political views – but the effect of symptoms is left to future research. 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 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 under free testing.

We recognise the importance testing epidemiologically, to identify positive cases, allowing self-isolation to break disease transmission. If the goal of street testing is just to allow quick and free testing, 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), proposals for new remedies such as test and tracing, to general 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, Argente, and Lippi, 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 and ONS studies are an interesting special case of large-scale community testing on a nationally representative sample. It is claimed that the sample is truly random. While we use such data as the best available estimate, our experimental results suggest randomness might be wanting. Even for free tests at home (see 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 is done monthly or less often, while our procedure can be run daily.

This paper also contributes to the literature on testing regimens (Mina, Parker, and Larremore, 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 insights into how the bias changes in different phases of the disease, and 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

[1] Daron Acemoglu and Simon Johnson. "Disease and Development: The Effect of Life Expectancy on Economic Growth". In: *Journal of Political Economy* 115.6 (2007), pp. 925–985. DOI: 10.1086/529000. eprint: https://doi.org/10.1086/529000. URL: https://doi.org/10.1086/529000.

[2] Jérôme Adda. "Economic activity and the spread of viral diseases: Evidence from high frequency data". In: *The Quarterly Journal of Economics* 131.2 (2016), pp. 891–941.

[3] Fernando E Alvarez, David Argente, and Francesco Lippi. *A simple planning problem for Covid-19 lockdown*. Tech. rep. National Bureau of Economic Research, 2020.

[4] Andrew Atkeson. *How deadly is COVID-19? Understanding the difficulties with estimation of its fatality rate*. Tech. rep. National Bureau of Economic Research, 2020.

[5] 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.

[6] 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.

[7] Jan M. Brauner et al. "Inferring the effectiveness of government interventions against COVID-19". In: *Science* (2020). ISSN: 0036-8075. DOI: 10.1126/science.abd9338. eprint: https://science.sciencemag.org/content/early/2020/12/15/science.abd9338.full.pdf. URL: https://science.sciencemag.org/content/early/2020/12/15/science.abd9338.

[8] Flavio A Cadegiani et al. "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)." In: *medRxiv* (2021), pp. 2020–12.

[9] Itai Dattner et al. "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". In: *medRxiv* (2020).

[10] Theologos Dergiades, Costas Milas, and Theodore Panagiotidis. "Effectiveness of government policies in response to the COVID-19 outbreak". In: *Available at SSRN 3602004* (2020).

[11] ECDC. *Data on testing for Covid-19 by week and country.* Tech. rep. Available at: https://www.ecdc.europa.eu/en/publications-data/covid-19-testing, 2021.

[12] Kelly Osezele Elimian et al. "Patient characteristics associated with COVID-19 positivity and fatality in Nigeria: Retrospective cohort study". In: *BMJ open* 10.12 (2020), e044079.

[13] Uri Gneezy et al. "Measuring Success in Education: The Role of Effort on the Test Itself". In: *American Economic Review: Insights* 1.3 (2019), pp. 291–308. DOI: 10.1257/aeri.20180633. URL: https://www.aeaweb.org/articles?id=10.1257/aeri.20180633.

[14] Russell Golman, David Hagmann, and George Loewenstein. "Information avoidance". In: *Journal of Economic Literature* 55.1 (2017), pp. 96–135.

[15] Sharon K Greene et al. "Nowcasting for Real-Time COVID-19 Tracking in New York City: An Evaluation Using Reportable Disease Data From Early in the Pandemic". In: *JMIR Public Health and Surveillance* 7.1 (2021), e25538.

[16] Nils Haug et al. "Ranking the effectiveness of worldwide COVID-19 government interventions". In: *Nature Human behaviour* 4.12 (2020), pp. 1303–1312.

[17] Solomon Hsiang et al. "The effect of large-scale anti-contagion policies on the COVID-19 pandemic". In: *Nature* 584.7820 (2020), pp. 262–267.

[18] Steven J Kachelmeier and Mohamed Shehata. "Examining risk preferences under high monetary incentives: Experimental evidence from the People's Republic of China". In: *The American Economic Review* (1992), pp. 1120–1141.

[19] Alyson A Kelvin and Scott Halperin. "COVID-19 in children: the link in the transmission chain". In: *The Lancet Infectious Diseases* 20.6 (2020), pp. 633–634.

[20] Elisabeth Mahase. "Covid-19: Mass testing in Slovakia may have helped cut infections". In: *BMJ* 371:m4761 (2020). DOI: https://doi.org/10.1136/bmj.m4761.

[21] Charles F Manski and Francesca Molinari. "Estimating the COVID-19 infection rate: Anatomy of an inference problem". In: *Journal of Econometrics* 220.1 (2021), pp. 181–192.

[22] Rutger A Middelburg and Frits R Rosendaal. "COVID-19: How to make between-country comparisons". In: *International Journal of Infectious Diseases* 96 (2020), pp. 477–481.

[23] Michael J. Mina, Roy Parker, and Daniel B. Larremore. "Rethinking Covid-19 Test Sensitivity — A Strategy for Containment". In: *New England Journal of Medicine* 383.22 (2020), e120. DOI: 10.1056/NEJMp2025631. eprint: https://doi.org/10.1056/NEJMp2025631. URL: https://doi.org/10.1056/NEJMp2025631.

[24] Emily Oster, Ira Shoulson, and E Dorsey. "Optimal expectations and limited medical testing: Evidence from Huntington disease". In: *American Economic Review* 103.2 (2013), pp. 804–30.

[25] Martin Pavelka et al. "The effectiveness of population-wide, rapid antigen test based screening in reducing SARS-CoV-2 infection prevalence in Slovakia". In: *medRxiv* (2020).

[26] Haiyan Qiu et al. "Clinical and epidemiological features of 36 children with coronavirus disease 2019 (COVID-19) in Zhejiang, China: an observational cohort study". In: *The Lancet Infectious Diseases* 20.6 (2020), pp. 689–696.

[27] REACT. *REACT-1: Summary of sample extraction and fieldwork dates, and response rates, by round*. Tech. rep. Retreived on 2021-01-30. 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.

[28] Steven Riley et al. "Community prevalence of SARS-CoV-2 virus in England during May 2020: REACT study". In: *medRxiv* (2020).

[29] Marta Serra-Garcia and Nora Szech. "Demand for COVID-19 antibody testing, and why it should be free". In: 8340 (2020). URL: https://ssrn.com/abstract=3623675.

[30] Chen Shen et al. "Unraveling the flaws of estimates of the infection fatality rate for COVID-19". In: *Journal of Travel Medicine* (2021).

[31] Rebecca L. Thornton. "The Demand for, and Impact of, Learning HIV Status". In: *American Economic Review* 98.5 (2008), pp. 1829–63. DOI: 10.1257/aer.98.5.1829. URL: https://www.aeaweb.org/articles?id=10.1257/aer.98.5.1829.

Appendix

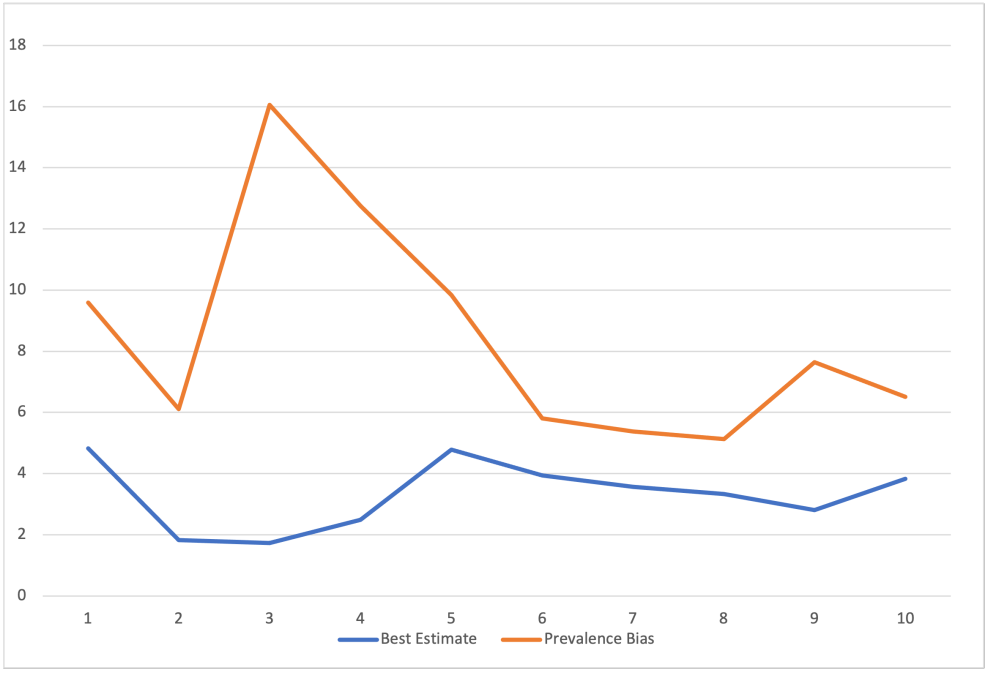

Figure 7: Estimate of the prevalence bias in field testing. Test positivity divided by the best prevalence estimate using REACT and ONS data.

Table 3: Odds of 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			
	Ref:30-50 — No Symptoms		Ref:30-50 — Symptoms	
	(1)	(2)	(3)	(4)
30-50 — Symptoms	-1.279^{***}	-1.186^{***}		
	(0.105)	(0.132)		
30-50 — No Symptoms			1.279^{***}	1.173^{**}
			(0.105)	(0.563)
50+ — No symptoms	-0.107	0.018	1.172^{***}	0.612
	(0.105)	(0.125)	(0.112)	(0.633)
50+ — Symptoms	-1.403^{***}	-1.251^{***}	-0.124	0.028
	(0.119)	(0.140)	(0.121)	(0.372)
Under 30 — No symptoms	-0.114	0.003	1.165^{***}	1.775^{***}
	(0.099)	(0.121)	(0.108)	(0.518)
Under 30 — Symptoms	-1.342^{***}	-1.270^{***}	-0.064	0.693^{**}
	(0.109)	(0.133)	(0.112)	(0.320)
Income (proxy)		0.0004^{**}		
		(0.0002)		
Observations	1,150	808	1,150	219
R^2	0.264	0.259	0.264	0.068
Max. Possible R^2	1.000	1.000	1.000	0.963
Log Likelihood	$-6{,}177.707$	$-4{,}069.409$	$-6{,}177.707$	-352.124
Wald Test	351.360^{***} (df = 5)	240.100^{***} (df = 6)	351.360^{***} (df = 5)	17.750^{***} (df = 5)
LR Test	352.721^{***} (df = 5)	242.062^{***} (df = 6)	352.721^{***} (df = 5)	15.353^{***} (df = 5)
Score (Logrank) Test	388.121^{***} (df = 5)	264.183^{***} (df = 6)	388.121^{***} (df = 5)	20.340^{***} (df = 5)

Note: *p<0.1; **p<0.05; ***p<0.01