What can we learn about SARS-CoV-2 prevalence from testing and hospital data?[*]

Daniel W. Sacks[†]

Nir Menachemi[‡]

Peter Embí[§]

Coady Wing[¶]

August 4, 2020

Abstract

Measuring the prevalence of active SARS-CoV-2 infections is difficult because tests are conducted on a small and non-random segment of the population. But people admitted to the hospital for non-COVID reasons are tested at very high rates, even though they do not appear to be at elevated risk of infection. This sub-population may provide valuable evidence on prevalence in the general population. We estimate upper and lower bounds on the prevalence of the virus in the general population and the population of non-COVID hospital patients under weak assumptions on who gets tested, using Indiana data on hospital inpatient records linked to SARS-CoV-2 virological tests. The non-COVID hospital population is tested fifty times as often as the general population. By mid-June, we estimate that prevalence was between 0.01 and 4.1 percent in the general population and between 0.6 to 2.6 percent in the non-COVID hospital population. We provide and test conditions under which this non-COVID hospitalization bound is valid for the general population. The combination of clinical testing data and hospital records may contain much more information about the state of the epidemic than has been previously appreciated. The bounds we calculate for Indiana could be constructed at relatively low cost in many other states.

Key words: COVID-19, SARS-CoV-2, prevalence, partial identification

---

[†]Business Economics & Public Policy, Kelley School of Business, Indiana University (dansacks@indiana.edu).

[‡]Fairbanks School of Public Health and Kelley School of Business, Indiana University, and Regenstrief Institute, Inc. (nirmena@iu.edu).

[§]School of Medicine, Indiana University, Indiana University Health System, and Regenstrief Institute, Inc. (pembi@regenstrief.org).

[¶]O'Neill School of Public and Environmental Affairs, Indiana University (cwing@indiana.edu).

1 Introduction

Constructing credible estimates of the current prevalence of SARS-CoV-2 in the United States is challenging. Despite growing since the start of the epidemic, testing rates remain low in most of the country: only a small fraction of the population is tested for SARS-CoV-2 on any given day. Moreover, tests are often allocated to people exhibiting COVID-19 symptoms or who are thought to have come into contact with the virus (Abbott and Lovett, 2020). For example, New York State and Texas both use a self-diagnostic tool to screen people for testing.[1] The low rate of testing in the general population means that the number of confirmed SARS-CoV-2 cases almost certainly understates the true number of infections in the population. At the same time, statistics like the fraction of tests that are positive likely overstate population prevalence because the tested population is more likely to be infected than the population as a whole.

In this paper, we propose a new approach to measuring the point-in-time prevalence of active SARS-CoV-2 infections in the overall population using data on patients who are hospitalized for non-COVID reasons. There is less uncertainty about SARS-CoV-2 prevalence for the non-COVID hospital patient population because people in the hospital are tested at much higher rates than the general population, even if they are hospitalized for reasons unrelated to COVID-19 (Sutton et al., 2020). We combine detailed testing data and hospital data from Indiana with a family of weak monotonicity assumptions that seem to have high credibility. The combination of these assumptions with linked testing-hospital data leads to relatively tight upper and lower bounds on the prevalence of active SARS-CoV-2 infections in the overall population in Indiana in each week from mid-March to mid-June. The detailed Indiana data allows us to conduct robustness checks that partially validate some of our assumptions. Our basic method (without the validation checks) could be implemented using data that many states are already collecting and partially reporting. Thus, our approach could help states extract timely prevalence information using existing surveillance data. Importantly, our paper is focused on estimates of the fraction of the population that would test positive in each week. These estimates are distinct from recent efforts to estimate the share of the population *ever infected* with SARS-CoV-2 (Manski and Molinari, 2020). The distinction between active prevalence and cumulative prevalence is important because the prevalence of active infections is a key determinant of the spread of the epidemic, given that the level of immunity in the population is thought to be quite low.

---

[1] See `https://covid19screening.health.ny.gov/covid-19-screening/` and `https://txctt.force.com/ct/s/assessment?language=en_US`.

Estimates and forecasts of the prevalence of active SARS-CoV-2 infections are crucial for public and private responses to the disease. They have shaped decisions about disruptive non-pharmaceutical interventions such as school closures, non-essential business closures, gathering restrictions, stay-at-home mandates, and temporary increases in the generosity of the unemployment insurance system (Gupta et al., 2020). Reported estimates of prevalence likely also motivate individual precautionary behaviors ranging from wearing a mask to reducing demand for goods and services that require physical interaction (Gupta et al., 2020; Allcott et al., 2020; Philipson, 2000, 1996; Kremer, 1996). Finally, estimates of prevalence are a necessary input into efforts to measure other quantities of interest, like the infection fatality rate and the infection hospitalization rate.

Given its importance, researchers have developed several approaches to measuring SARS-CoV-2 prevalence in light of the challenge of non-representative testing. One of the most credible is to conduct a biometric survey in which tests are offered to a representative sample of the population (Menachemi et al., 2020; Richard M. Fairbanks School of Public Health, 2020; Gudbjartsson et al., 2020). Other studies have tested a census of smaller populations such as cruise ships (e.g. Russell et al. (2020)). However, it is difficult and costly to regularly implement a survey with accurate coverage and high response rates, especially a new survey that has not been in the field for long. A second approach involves backcalculation methods, which use data on observed hospitalizations or deaths to infer disease prevalence at earlier dates using assumptions about the unobserved parameters that determine the progression of the disease, hospitalization rates, and case fatality rates (Brookmeyer and Gail, 1988; Egan and Hall, 2015; Flaxman et al., 2020; Salje et al., 2020). Backcalculation may work well if hospitalizations or deaths are well measured, and if previous research has already reached consensus about key parameters related to the disease. However, backcalculation may be less credible for a novel virus because the scientific knowledge base is smaller. And even when backcalculation is based on credible assumptions, it may be of limited value for public health decision making because both hospitalizations and deaths lag current infections (Verity et al., 2020).

An alternative to biometric surveys and backcalculation is to combine non-random clinical testing data with weak distributional assumptions to construct bounds on population prevalence (e.g. Manski (1999); Wing (2010)). In the COVID-19 epidemic, Stock et al. (2020) provide bounds on population prevalence using a testing encouragement design, which is not always available. Manski and Molinari (2020) bound the share of the population that has ever been infected under a "test monotonicity assumption" that the infection rate is weakly higher among the tested than among the untested. This assumption is appealingly credible, and the bounds can be calculated from widely available test

data. However, even when the focus is cumulative prevalence under test monotonicity, the prevalence bounds are often wide because testing is so rare.

We pursue a related strategy in this paper. Our analysis is based on the insight that test rates are especially high among hospitalized patients, even patients hospitalized for reasons that are apparently unrelated to COVID-19, such as labor and delivery or vehicle accidents. The upper and lower bounds on prevalence in these populations will tend to be much tighter than similar bounds in the general population. In addition, it is plausible that some types of hospitalizations occur for reasons that are independent of infection risk. For example, people who are hospitalized for injuries sustained in a traffic accident might be expected to have about the same risk of SARS-CoV-2 infection as the general population. We build on this insight to estimate more informative upper and lower bounds on weekly SARS-CoV-2 prevalence in the population.

Our paper makes some methodological contributions that may be relevant to the development of more informative public health surveillance systems. We describe the conditions under which upper and lower bounds on active prevalence among non-COVID hospitalizations are valid estimates of the upper and lower bounds on prevalence in the general population. We maintain the test monotonicity assumption throughout, and we derive upper and lower bounds on prevalence in the population under two alternative assumptions about the representativeness of non-COVID hospitalizations for the broader population. The first assumption is a relatively weak "hospital monotonicity" assumption that prevalence is at least as high among non-COVID hospital patients as it is in the general population. This assumption would be satisfied even if hospitalized patients are at greater risk of COVID because, for example, people who get into car accidents have more social interactions. The second assumption is a stronger "hospital independence" assumption that prevalence is the same among non-COVID hospital patients as it is in the general population. The resulting bounds are informative for prevalence at a point in time, not just cumulative prevalence. This is important because the information required to estimate the bounds is closely related to the simple statistics that many states already report. It appears possible for many states to use our method to report upper and lower bounds on prevalence in near real time using data that they already collect and report. In particular, states already report the COVID-hospitalization rate, as well as overall test rates and test positivity rates. To report a version of the upper and lower bounds we describe in this paper, states would simply have to compute testing and positivity rates among non-COVID hospitalizations.

Our first empirical contribution uses this framework to estimate upper and lower bounds on weekly prevalence of SARS-CoV-2 in Indiana. To operationalize the basic

idea, we work with two definitions of non-COVID hospitalizations. The first, which we call non-Influenza- or COVID-like-illness (non-ICLI) hospitalizations, simply excludes all hospitalizations with diagnosis for ICLI (Center for Disease Control and Prevention, 2020; Armed Forces Health Surveillance Center, 2015). This definition would be easy to implement in many different hospital data sets and yields a large population of hospital patients. However, we also work with a definition using a narrower set of patients who are hospitalized for six groups of clear non-COVID causes: (i) cancer; (ii) appendicitis and vehicle accidents; (iii) labor and delivery; (iv) AMI and stroke; (v) fractures, crushes, and open wounds; and (vi) other accidents. The clear-cause analysis is more intuitive and transparent, but it is based on smaller samples and might be harder to implement as part of a public health surveillance system.

We show that, in Indiana, test rates are much higher among non-COVID hospital patients than in the general population. For example, in June, about 0.4 percent of the general population was tested in a given week, compared with 24 percent of non-COVID hospital patients; test positivity rates are lower among non-COVID hospital patients than in the general population. In the general population, 4.1 percent of tests were positive. In contrast, among non-COVID hospital patients who were tested, only 2.6 percent of tests were positive. These testing and positivity rates can be combined to estimate upper and lower bounds on prevalence. Under the test monotonicity assumption, between 0.01 and 4.1 percent of the general population was infected with SARS-CoV-2 on June 15. Under the same test monotonicity assumption, the bounds are half as wide for non-COVID hospital patients: prevalence was between 0.6 and 2.6 percent. Under the hospital monotonicity assumption, the upper bound on prevalence in the non-COVID hospital population is a valid upper bound on population prevalence. And under the stronger hospital independence assumption, the upper and lower bounds on prevalence in the non-COVID hospital population are valid upper and lower bounds on population prevalence. The bounds on prevalence based on the non-ICLI definition are typically very similar to the bounds based on the clear-cause definition. We present bounds under alternative hospital representativeness assumptions (none, monotonicity, independence) to allow readers to make their own judgements about which assumptions are credible, and to better understand how much information about prevalence is derived from the data versus assumptions.

We also calculate bounds among ICLI hospitalizations. We find, of course, much higher prevalence among this group, but our bounds still rule out very high prevalence. Specifically, we find that SARS-CoV-2 prevalence among ICLI hospitalizations is no higher than 50 percent at its peak, and no higher than about 30 percent by mid-June. This shows

4

that there is value to testing even highly symptomatic patients, as their SARS-CoV-2 rate is far from 100 percent, and testing outcomes would be informative for treatment and quarantine decisions.

Our second empirical contribution is to assess the credibility of key assumptions that would be difficult to study in other data sets. Manski and Molinari (2020) point out that the accuracy of SARS-CoV-2 virological tests is not well understood. Incorporating information about testing errors alters the bounds on prevalence. We use data on people who were tested twice in a two day period to shed some light on the fraction of people who test negative but are actually infected. We tentatively conclude that test errors are have a negligible effect on the upper and lower bounds on prevalence reported in our paper.

We also assess the credibility of the hospital representativeness assumptions at the core of the paper. The most restrictive hospital independence condition assumes that SARS-CoV-2 prevalence is the same in the non-COVID and general populations, and the weaker hospital monotonicity condition assumes that prevalence is at least as high among the hospitalized. Although we are not able to directly validate these assumptions, we probe their credibility in two ways. First, we compare the hospitalization bounds to estimates of population prevalence obtained from tests of a random sample of Indiana residents in April and June (Menachemi et al., 2020; Richard M. Fairbanks School of Public Health, 2020). Second, we examine the pre-hospitalization test rates of non-COVID hospitalized patients. These validity checks are roughly consistent with the the hospital independence assumption, and highly consistent with hospital monotonicity. This suggests that these assumptions might be considered reasonable in other states where it would be easy to estimate the upper and lower bounds but harder to perform elaborate validation exercises.

Overall, our results indicate that combining testing data and information on non-COVID hospitalizations may be a feasible and informative way of measuring SARS-CoV-2 prevalence. In the most recent weeks of our data (and under test monotonicity but without hospitalization data), we can only conclude that at most 4 percent of the overall Indiana population was actively infected. With hospitalization data and the hospital monotonicity assumption, we can conclude that at most 2.6 percent of the Indiana population was infected. Despite this substantial improvement, the bounds remain wide enough that we cannot conclude whether prevalence has fallen since early April. Similar bounds could be constructed in other states using aggregate data on non-COVID hospitalizations and their testing and test positivity rates, potentially improving COVID surveillance systems across the country.

2 Inferring COVID prevalence from incomplete testing

The empirical goal of our study is to bound the fraction of the Indiana population that is infected with SARS-CoV-2 in each week. To fix ideas, we use $i = 1...N$ to index the population of Indiana. Let $C_{it} = 1$ indicate that person i is currently infected with SARS-CoV-2 on date t. The population prevalence of active SARS-CoV-2 infections in Indiana at date t is $Pr(C_{it} = 1) = \frac{1}{N}\sum_{i=1}^{N} C_{it}$, where we leave conditioning on date t implicit to reduce clutter. We are also interested in prevalence among hospital inpatients with various COVID- and non-COVID-related diagnoses. This is simply the probability that a person is infected, conditional on being hospitalized for a specified condition or set of conditions. Let H_{it} be a binary indicator set to 1 if the person was hospitalized with a non-COVID-related diagnosis. Then $Pr(C_{it} = 1|H_{it} = 1)$ is the prevalence of active SARS-CoV-2 infections in the sub-population of people who were admitted with a non-COVID-related diagnosis on date t.

A key inferential challenge in estimating prevalence is that values of C_{it} are unknown for most people on most days, because testing is rare. Let $D_{it} = 1$ if person i was tested on t and $D_{it} = 0$ if the person was not tested. Let $Pr(D_{it} = 1)$ represent the proportion of the population tested on date t, where conditioning on t is implicit. Continuing with the notation laid out above, $Pr(C_{it}|D_{it} = 1)$ and $Pr(C_{it}|D_{it} = 0)$ represent prevalence among people who are tested and not tested, respectively. The value of C_{it} is observed for people where $D_{it} = 1$, but unknown for people where $D_{it} = 0$, which means that $Pr(C_{it}|D_{it} = 0)$ is not identified by the data on testing and test outcomes.[2] We define prevalence in the tested and untested hospital populations similarly, but with conditioning on both testing status and hospitalization status.

2.1 Worst Case Bounds

In the absence of any distributional assumptions, the observed clinical tests partially identify prevalence overall, and in any sub-populations that can be defined by observable covariates. Use the law of total probability to decompose population prevalence:

$$Pr(C_{it}) = Pr(C_{it}|D_{it} = 1)Pr(D_{it} = 1) + Pr(C_{it}|D_{it} = 0)Pr(D_{it} = 0)$$

---

[2] Our analysis here assumes that there is no measurement error involved in COVID-19 testing procedures. In practice, false negatives and false positives could create additional uncertainty. We abstract from such concerns throughout the paper.

The only unknown quantity on the right-hand side of the expression is $Pr(C_{it}|D_{it} = 0)$, which is prevalence among people who were not tested. Without any additional assumptions or data, all that is known is that this value lies between 0 and 1. Substituting 0 and 1 for the unknown prevalence yields worst-case lower and upper bounds L_w and U_w on population prevalence:

$$L_w = Pr(C_{it}|D_{it} = 1)Pr(D_{it} = 1),$$
$$U_w = Pr(C_{it}|D_{it} = 1)Pr(D_{it} = 1) + Pr(D_{it} = 0).$$

These bounds define the set of values for unknown population prevalence that are compatible with the observed data and the logical definition of prevalence. The lower bound is the confirmed positive rate, and the upper bound is that rate plus the untested rate. The width of the worst-case bounds on a given day is decreasing in that day's testing rate. Testing more people can only increase the confirmed positive rate and decrease the untested rate. However if few people are tested, the bounds can be very wide.

2.2 Test monotonicity

To narrow the bounds, Manski and Molinari (2020) propose the "test monotonicity" condition. This condition requires that the prevalence of SARS-CoV-2 is at least as high in the tested population as it is in the untested population. Formally, test monotonicity implies that $Pr(C_{it}|D_{it} = 1) \geq Pr(C_{it}|D_{it} = 0)$. This is an appealing condition because virological tests are typically allocated to symptomatic individuals, who have a higher than average likelihood of infection.

Under test monotonicity, prevalence in the tested sub-population represents an upper bound on the unknown prevalence among the untested sub-population. The lower bound remains the worst-case lower bound. The lower and upper bounds under monotonicity, L_m and U_m, are:

$$L_m = Pr(C_{it}|D_{it} = 1)Pr(D_{it} = 1),$$
$$U_m = Pr(C_{it}|D_{it} = 1).$$

The new upper bound will be lower than the worst-case upper bound as long as prevalence in the tested sub-population is less than 1. In our data, test rates are less than 1 percent and positivity rates in the population are roughly 10 percent, so this assumption

brings the upper bound down from 99 percent to 10 percent or less. A similar assumption could be made for the non-COVID hospitalization population, yielding bounds L_m^H and U_m^H on prevalence among the non-COVID hospitalized population.

2.3   Inferring population prevalence from non-COVID hospital patients

Test monotonicity can be used to narrow the the bounds on prevalence in the population and among non-COVID hospitalized patients. Because testing rates are much higher in hospitals than in the general population, the bounds on prevalence in hospitalized subpopulations are much narrower. Thus, assumptions that link hospital and population prevalence may be a powerful way to reduce uncertainty about population prevalence. We pursue two types of assumptions that enable extrapolation from non-COVID hospital populations to the general population: (i) monotone selection into hospitalization and (ii) risk-independent hospitalization. These are both forms of *hospital instrumental variable* assumptions, and we refer to them collectively as hospital IV assumptions.

Hospital monotonicity lowers the population upper bound

The hospital monotonicity assumption requires that the prevalence of active SARS-CoV-2 infections among non-COVID hospital patients is not lower than the prevalence of active infections in the general (non-hospitalized) population. Formally, $Pr(C_{it}|H_{it} = 1) \geq Pr(C_{it})$. When prevalence is bounded in the hospitalized and general populations, the hospital monotonicity assumption may further reduce the width of both sets of bounds by ruling out values that would violate it.

In particular, under hospital monotonicity, the upper bound on population prevalence cannot be larger than the upper bound on hospital prevalence. When both the hospital monotonicity assumption and the test monotonicity assumption are imposed, the upper bound on SARS-CoV-2 prevalence in the population is

$$
\begin{aligned}
U_{m,h} &= \min\left\{U_m, U_m^H\right\} \\
&= \min\left\{Pr(C_{it} = 1|D_{it} = 1), Pr(C_{it} = 1|D_{it} = 1, H_{it} = 1)\right\},
\end{aligned}
$$

where U_m is the upper bound on prevalence in the population under hospital monotonicity and U_m^H is the upper bound on prevalence among non-COVID hospital patients. In our data, the hospital upper bound is typically lower than the population upper bound, so the hospital monotonicity condition in practice implies that the positivity rate among

non-COVID hospitalizations is an upper bound population prevalence.[3]

Hospital independence tightens both population bounds

A stronger assumption that also facilitates extrapolation from a hospitalized sub-population to the general population is a "hospital independence" assumption, which means that hospitalization for non-COVID-related health conditions is mean independent of infection with SARS-CoV-2. Independence implies that people infected with SARS-CoV-2 have the same probability of being hospitalized for a non-COVID condition as people who are not infected with SARS-CoV-2 so that $Pr(H_{it}|C_{it} = 1) = Pr(H_{it}|C_{it} = 0)$. Equivalently, the independence assumption implies that SARS-CoV-2 prevalence is the same among people who are hospitalized for non-COVID conditions and the general population. That is, under the hospital independence assumption: $Pr(C_{it}|H_{it}^j) = Pr(C_{it})$. The independence assumption implies that non-COVID hospitalizations are an instrumental variable for testing. This assumption would be satisfied if non-COVID hospitalizations arose randomly in the population, for example because of health conditions (such as pregnancy or heart disease) determined prior to the epidemic. The hospital independence assumption would fail, however, if hospitalization risk was systematically related to COVID-19 risk, for example because essential workers have more social interactions and greater likelihood of hospitalizations.

Under the hospital independence assumption, the bounds on the common prevalence parameter are defined by the intersection of the hospital and population bounds. The lower and upper bounds under test monotonicity and hospital independence, $L_{m,ind}$ and $U_{m,ind}$, are:

$$
\begin{aligned}
L_{m,ind} &= \max \left\{ U_m, U_m^H \right\} \\
&= \max \left\{ Pr(C_{it} = 1 | D_{it} = 1), Pr(C_{it} = 1 | D_{it} = 1, H_{it} = 1) \right\}, \\
U_{m,ind} &= \min \left\{ U_m, U_m^H \right\} \\
&= \min \left\{ Pr(C_{it} = 1 | D_{it} = 1), Pr(C_{it} = 1 | D_{it} = 1, H_{it} = 1) \right\}.
\end{aligned}
$$

Under hospital independence (as well as test monotonicity), the upper bound on prevalence is the same as under hospital monotonicity. What the hospital independence assumption buys us is a tighter lower bound, which is now the greater of the lower bounds

---

[3] In principle, the same assumption could be used to derive a potentially higher lower bound on prevalence in the hospital population. But this lower bound condition is essentially non-binding in practice, so we ignore it here.

on population and hospital prevalence under test monotonicity. In practice we find that the lower bound is always higher in the non-COVID hospitalization sub-population than in the general population, so in practice this assumption implies that the lower bound on population prevalence is the confirmed positive rate among non-COVID hospitalizations.

2.4 Measurement Error in Testing

Virological tests for the presence of SARS-CoV-2 may not be perfectly accurate, and so far there are no detailed studies of the performance of the PCR tests that Indiana is using to test people for SARS-CoV-2. To clarify how error-ridden tests complicate our prevalence estimates, we augment our notation to distinguish between test results and virological status. We continue to use C_{it} to represent a person's true infection status, and we still use D_{it} to indicate whether a person was tested at date t or not. But now we introduce R_{it}, which is a binary measure set to 1 if the person tests positive and 0 if the person tests negative. Using this notation, $Pr(C_{it} = 1|D_{it} = 1, R_{it} = 1)$ is called the Positive Predictive Value (PPV) of the test among people who are tested and who test positive. $Pr(C_{it} = 0|D_{it} = 1, R_{it} = 0)$ is called the Negative Predictive Value (NPV) among people who are tested and who test negative. $1 - NPV = Pr(C_{it} = 1|D_{it} = 1, R_{it} = 0)$ is the fraction of people who test negative who are actually infected with SARS-CoV-2.

Our initial worst case bounds assumed no test errors. Relaxing that assumption yields a different set of upper and lower bounds on prevalence. Following Manski and Molinari (2020), we assume that (i) $PPV = 1$ so that none of the positive tests are false, but (ii) $Pr(C_{it} = 1|D_{it} = 1, R_{it} = 0) \in [\lambda_l, \lambda_u]$. The second condition imposes a bound on $1 - NPV$, which is the fraction of people who test negative who are actually infected. Under these two restrictions, the new worst case bounds work out to:

$$L_{w,\lambda} = L_w + \lambda_l Pr(R_{it} = 0|D_{it} = 1) Pr(D_{it} = 1)$$
$$U_{w,\lambda} = U_w + \lambda_u Pr(R_{it} = 0|D_{it} = 1) Pr(D_{it} = 1)$$

Allowing for test errors in this way increases the worst case lower bound by the best-case fraction of missing positives, and increases the worst case upper bound by the worst-case fraction of missing positives. Similar expressions hold for prevalence bounds under test monotonicity and other independence assumptions.

The upshot of this analysis is that knowledge of test accuracy is important for efforts to learn about prevalence. In their study of the cumulative prevalence of SARS-CoV-2 infections, Manski and Molinari (2020) computed upper and lower bounds on prevalence

under the assumption that $\lambda_l = .1$ and $\lambda_u = .4$, citing Peci et al. (2014). Manski and Molinari (2020) view this choice of $.1 \leq 1 - NPV \leq .4]$ as an expression of scientific uncertainty about test errors, and they refer to the resulting prevalence bounds as "illustrative". However, the structure of the test error bounds makes it clear that assumptions about the numerical magnitude of test errors have inferential consequences. For example, setting $\lambda_u = .4$ implies that, regardless of the outcome of the test, at least 40 percent of the people who are tested for SARS-CoV-2 are infected.

Although there is little published evidence on the properties of the SARS-CoV-2 PCR test, previous research suggests that PCR test errors are uncommon in other settings. For example, Peci et al. (2014) study the performance of rapid influenza tests using PCR-based tests as a *gold standard*. PCR tests are used as a gold standard because they are expected to have very high PPV and NPV.

To shed more light on test errors, we constructed a sample of people who were (i) tested on day t, (ii) not tested on day $t-1$, and (iii) were tested again on day $t+1$. A total of 16,401 test pairs met this criteria. Using $R1_i$ and $R2_i$ to represent the results of a person's first and second test, we found that $Pr(R1_i = 1, R2_i = 1) = .11$ and $Pr(R1_i = 0, R2_i = 0) = .88$ among the people in the twice-tested sample. The two tests were discordant for less than 1 percent of the twice-tested sample. In Appendix C, we estimate prevalence and NPV in the twice-tested sample under the strong assumptions that the tests have specificity equal to 1, sensitivity does not depend on initial test result, and retesting is random. Using this method, we find that prevalence was 12 percent and $NPV = 0.995$ in the twice-tested sample. With $1 - NPV = 0.005$, the estimates imply only about 0.5 percent of people who test negative are actually infected.

The twice-tested sample is likely not representative of the population, of course. People who are re-tested after a negative test are probably highly symptomatic.[4] This suggests that $1 - NPV = Pr(C_{it}|D_{it}, R_{it} = 0)$ is probably higher in the twice-tested sample than in the population. Accordingly, we think that a plausible value for λ_l is nearly zero, and a plausible value for λ_u is 0.005. Accounting for test errors in this range would have almost no effect on the upper and lower bounds reported in the paper.

2.5 Summary and data requirements

Our bounds turn out to be fairly simple objects. Under test monotonicity, the lower bound on prevalence is the confirmed positive rate, the share of the population that tests

---

[4] Furthermore, it is possible that the first sample was drawn with less care than the second sample, suggesting that specificity may have differed across the repeated tests.

positive. The corresponding upper bound under monotonicity is the test positivity rate, the share of tests that are positive. Under hospital monotonicity, the upper bound becomes the test positivity rate among non-COVID hospitalizations. And under hospital representativeness, the lower bound becomes the confirmed positive rate among non-COVID hospitalizations.

An appealing feature of these bounds is that they can be calculated with little additional data beyond what public health organizations already report. Every state already reports the number of tests and the number of positive tests, and many states report the number of COVID-related hospitalizations.[5] States would only have to report test and positivity rates for non-COVID-related hospitalizations. This appears possible because many states already report "suspected" or "under investigation" COVID hospitalizations, defined as hospitalized patients exhibiting COVID-like illness. Some states actually report both the number of hospitalizations of patients with COVID- or influenza-like illness and, separately, the number of hospitalizations of patients with a positive SARS-CoV-2 test (e.g. Arizona and Illinois (Arizona Department of Health Services, 2020; Illinois Department of Public Health, 2020a,b)).[6] Thus states have the capacity to identify ICLI-related hospitalizations and link hospitalization and testing data.

3 Indiana Hospital and Testing Data

3.1 Data sets

Our analysis is based on two main data sources managed by the Regenstrief Institute. First we obtain data on the near universe of clinical virological tests for SARS-CoV-2 conducted in Indiana between January 1, 2020 and June 18, 2020. Second we obtained data on all inpatient hospital admissions from hospitals that belong to the Indiana Network for Patient Care (INPC), which is a health information exchange that centralizes and stores data from health providers across the state of Indiana, including all hospitals with emergency departments.[7] The hospital data are derived from the same database that the state uses for reporting hospitalizations on its dashboard (Indiana State Department of Health, 2020). We link the testing and hospital data using an encrypted common identifier. Of

---

[5] See, e.g., The COVID Tracking Project (2020).

[6] States reporting both confirmed SARS-CoV-2 hospitalizations and hospitalizations of suspected cases or cases under investigation include California (California Department of Health, 2020), Colorado (Colorado Department of Public Health and Environment, 2020), Mississippi (Mississippi State Department of Health, 2020), Tennessee (Tennessee Department of Health, 2020), and Vermont (Vermont Department of Health, 2020).

[7] See Grannis et al. (2005) for more details.

course, only a subset of hospital patients are tested and only a subset of tested people appear in the inpatient hospital data. For both data sets, we are also able to link the data with basic demographic data collected by the INPC; this information is available only for a subset of patients.

The test data contain individual records for nearly all of the SARS-CoV-2 tests conducted in Indiana during 2020. A small number of tests are excluded from our data because some institutions that conduct tests provide data to INPC but do not allow the data to be used for research purposes. The consequence of these exclusions is that we are missing some tests, which will result in a reduced lower bound in our framework. Despite these exclusions, our data set tracks the state's official case counts fairly closely; see Appendix Figure A.1. As an aggregate summary, our data contain 39,472 total positive tests, and the state reports 41,541 positive tests as of June 18 (Indiana State Department of Health, 2020).[8] Each test record in our testing data includes information on the date the test was run, the outcome of the test (positive, negative, or inconclusive), and a patient identifier that we use to link the test data to demographic files and inpatient hospital files.

The hospital inpatient data contain separate observations for each admission. We always observe admission time and a patient identifier that we use to link to the test data and inpatient files. We observe discharge time and diagnosis information only for a subset of admissions. (Not all fields are available for all admissions because different institutions contribute different information to the INPC.) Because the INPC data come from health care providers and payers, the same hospitalization can appear in the data set multiple times. To de-duplicate these records, we keep one observation per admission time (defined second-by-second), keeping the observation with the most diagnosis codes.[9]

3.2 Key measures

In-hospital testing, positivity rate, and confirmed positives The fraction of people who are tested in the hospital is an important quantity of interest in our analysis because hospital patients are tested at a higher rate than the general population, and hospital testing may be less correlated with COVID symptoms. Our data do not distinguish whether a person was tested in the hospital or whether the test was initiated independently of the

---

[8] We compare positives rather than tests because some testing institutions appear not to report negative tests to the state. (Such institutions are not included in our data.)

[9] This procedure could in principle result in us dropping relevant diagnostic information, but in practice this would happen only very rarely. There are 311,127 admissions that are unique in terms of patient identifier and time, and 270,576 non-unique admissions. In 579 of these non-unique cases are there multiple records with the same number of diagnostic codes; we keep one of these multiples at random. In 30 of the 270,576 cases there are multiple records with diagnosis codes that provide conflicting first diagnoses.

hospital visit. We say that a hospitalized patient was tested in-hospital if she had at least one SARS-CoV-2 test dated between 2 days prior to admission to 4 days after admission. We chose to focus on this week-long period, rather than strictly between admission and discharge, for three reasons. First, some patients will be tested prior to admission, as part of their preparation for admission. Thus it is valuable to look prior to admission. Second, we observe the date the test is run, not the the date the sample is collected. Backlogs in the testing system may mean that a patient is discharged before the test is run. Third, for some admissions, we lack discharge dates, but we can still define in-hospital tests using this measure.

We say that a patient tests positive in the hospital if she has at least one positive COVID test between 2 days prior to admission and 4 days after admission. We define the positivity rate as the fraction of tests that are positive; the confirmed positive rate is the fraction of the population with a positive test.

Population Testing and Positivity Rates In some analyses we compare hospital testing and positivity to population testing and positivity. Hospital testing and positivity are defined over a week-long span for a given hospitalization. To make the comparison with the general population clean, we examine test rates and positivity in a given week-long period. We say that a person was tested if she was tested at least once in a given week, and we say she was positive if she was positive at least once in that period.

3.3   Sample construction

Throughout, a patient is in the "test sample" if they are tested at least once. We say a patient is in the "inpatient sample" if they are hospitalized at least once. We limit our analysis to admissions with non-missing diagnostic information, and we say a patient is in the "inpatient diagnoses sample" if they meet this restriction. This limitation is important because diagnostic information is necessary for distinguishing COVID-related admissions from non-COVID-related admissions.

ICLI and non-ICLI Hospitalizations We construct three analytic samples from the inpatient data. We start by defining hospitalizations for influenza- and COVID-like illness (ICLI) using ICD-10 codes. We identify admissions with any of a standard set of ICD-10 codes for ICLI following Armed Forces Health Surveillance Center (2015). Then we identify admissions with any of the additional ICD-10 codes that the CDC recommends using for coding COVID hospitalizations (Center for Disease Control and Prevention, 2020). Both the influenza-like and COVID-like diagnoses include general symptoms such as cough or fever, as well as more specific diagnoses like acute pneumonia, viral influenza,

or COVID-19. We classify hospitalizations as ICLI-related if they have any influenza- or COVID-like illness (ICLI) diagnoses, and we classify hospitalizations as non-ICLI if they are not ICLI-related. Appendix B lists the ICD-10 codes used to define the analytic samples.

We view the non-ICLI sample as a useful starting point for our analysis for two reasons. First, our hospital IV assumptions are most plausible for hospitalizations that are not obviously COVID-related, and this sample meets that criteria. Second, as we have noted, many states already classify hospitalizations as ICLI-related; thus non-ICLI hospitalizations are identifiable and measurable in near-real time, so this sample can be studied more broadly.

We acknowledge, however, that the non-ICLI sample may not satisfy the hospital IV assumptions for at least two reasons. First, it may condition on COVID itself, since a patient with a reported COVID diagnosis would be excluded from it. (In practice we observe many patients with positive COVID tests but no COVID diagnosis.) Second, COVID is a new disease with heterogeneous symptoms, so even if a patient is hospitalized because of COVID, she may not have one of our flagged diagnoses, and we may incorrectly call her hospitalization non-ICLI (Yang et al., 2020).

Clear-cause sample To avoid these problems, we study a third sample, which we call the "clear cause" sample. These are hospitalizations with a clear cause that is not obviously COVID-related. We define clear-cause hospitalizations as hospitalizations with a diagnosis code for labor and delivery, AMI, stroke, fractures, crushes, open wounds, appendicitis, vehicle accidents, other accidents, or cancer. For all of these conditions except cancer, we flag hospitalizations with a diagnosis at any priority. For cancer, we flag hospitalizations with a cancer diagnosis code as the admitting diagnosis, the primary final diagnosis, or any chemotherapy diagnosis. Appendix B lists the ICD-10 codes used for these classifications.

We view the clear-cause sample as important for two reasons. First, we believe the hospital IV assumptions are most plausible for this sample, so we believe the bounds on prevalence are most likely to be valid. Second, we view the clear-cause sample as offering a test of the validity of the non-ICLI sample. To the extent that the two samples generate similar bounds, we can be more confident that the non-ICLI sample is informative of broader population COVID prevalence, despite the problems with the non-ICLI classification. This would be valuable because classifying hospitalizations as ICLI-related or not requires less information than ascertaining a clear cause of the hospitalization.

Summary statistics We show summary statistics for all of our samples in Table 1, as well as for the state as a whole (from Census Fact Finder and United States Census Bu-

reau (2019)). The average tested and hospitalized patient is substantially older than the population as a whole, and also more likely to be female. Because the tested and hospitalized samples are not age representative of the general population, we reweight all samples to match the population age distribution. The tested and hospitalized samples are fairly similar to the general population in terms of racial composition. Limiting the inpatient sample to admissions with diagnoses reduces our sample size substantially, but it does not appear to change its demographic profile. Although our main analysis looks at test rates at the admission level (rather than the person level), the summary statistics show that hospitalized patients are vastly more likely to have ever been tested than the population as a whole – about 21 percent of ever-hospitalized patients, compared to about 280,000 out of 6.64 million (4.2 percent) for the general population. ICLI-related hospitalizations are especially likely to have ever been tested.

4   Bounds on COVID-19 prevalence

4.1   Bounds by broad sample

Figure 1 shows the SARS-CoV-2 testing rate for each of the sub-populations in our analysis. We report the exact values of each of the test rates and the weekly number of admissions in Appendix Table A.1. Because the tested and hospitalized populations have very different age distributions compared to the rest of the population, we reweight the hospital sub-populations to match the coarsened age distribution in the population. Specifically, we calculate test rates in week-by-age-group cells, for age groups 0-17, 18-30, 30-50, 50-64, 65-74, and 75 and older. Then we average these age-specific testing rates across the age groups, weighting each group by its population share. We report unweighted test rates in Appendix Table A.2.

Figure 1 shows vastly higher test rates in the hospitalized samples than in the general population. The testing rate in the general population grew from 0.2 percent in April to between 0.4 and 0.6 percent in May and June. So despite tripling, the weekly test in the Indiana population rate remained below 1 percent. In contrast, people hospitalized for ICLI were tested at a very high rate, between 65 and 75 percent in most weeks. Testing rates among non-ICLI hospital patients and among the clear-cause non-COVID hospital patients were lower than the ICLI sample but much higher than the population overall, about 20 percent in April and 25-30 percent in May and June. In other words, testing rates among non-COVID hospital inpatients are about 50 times higher than testing rates in the general population, but they are typically less than half as high as testing rates in the ICLI

population.

These high test rates imply tighter bounds on population prevalence under test monotonicity, as we show in Figure 2 and report in Appendix Table A.3. We age-weight the bounds, analogously to our weighting of test rates, and we report unweighted bounds in Appendix Table A.4. Several patterns are clear in the figure. First, the ICLI hospitalized population has higher upper and lower bounds on prevalence than the other groups. For the ICLI patients, the prevalence bounds are 5-22 percent in the first week of our sample, then increase to 30-40 percent in the last week of March, before declining steadily to 11-18 percent in the final week of the sample. These bounds rule out the possibility that all or nearly all symptomatic patients are infected with SARS-CoV-2. In most weeks the ICLI bounds lie outside the other groups' bounds, implying (unsurprisingly) unambiguously higher COVID prevalence. This separation shows that the data are sensible and that the bounds are informative enough to tell apart these highly distinct populations.

The second clear pattern in the figure is that the prevalence bounds are tighter for the non-ICLI and clear-cause hospitalization samples than for the all-test sample. In fact the bounds for both of these hospitalizations samples are always contained within the all-test bounds. Even at their tightest, the bounds for the all-test sample are as wide as 0.02 to 3.6 percent in the last week of our data. In that week the bounds for non-ICLI hospitalizations are [0.6%, 2.4%] and for clear-cause hospitalizations they are [0.5%, 1.9%]. This tight bound implies that the hospitalization data could be informative for population prevalence (under hospital monotonicity or mean independence).

The final pattern evident in Figure 2 is that the bounds for the non-ICLI hospitalization sample and for the clear-cause hospitalization sample are nearly indistinguishable. The only noticeable difference is that the upper bound for non-ICLI hospitalization is perhaps slightly higher. This fact is important because non-ICLI hospitalizations are potentially easier to measure, but they may be negatively selected in the sense that by construction they may exclude COVID-likely cases. The similarity of the non-ICLI bounds with the bounds for the clear-cause sample (which is not selected based on COVID-likelihood) provides some evidence in support of using non-ICLI hospitalizations to measure general prevalence.

4.2 Bounds by cause of admission

Our clear-cause hospitalization sample pools many distinct causes, including among others labor and delivery, vehicle accidents, and other accidents, including falls. In principle these hospitalizations may differ in their COVID likelihood. One might worry, for

17

example, that pregnant women are especially cautious and careful not to become infected, whereas people getting into vehicle accidents may be less cautious (either because they are not careful drivers, or because they are out of the house at all).

We therefore report test rates and bounds for disaggregated causes as well as the overall clear-cause sample. We focus on six sets of causes: cancers, appendicitis and vehicle accidents, injuries (fractures/crush/wounds), non-vehicle accidents, and AMI/stroke. These six groups have reasonable sample sizes throughout (they each have 1,000-1,500 admissions per month), and the demographic profiles within each group are roughly similar; see Appendix Table A.5 for age profiles of admitted patients by cause of admission. All ages are represented in the cancer sample: appendicitis and vehicle accidents both afflict young people; AMI, stroke, and other accidents—primarily falls—afflict older people; and labor and delivery is limited, of course, to women of childbearing age.[10] Because not all age groups are represented in every category, we do not age-weight these results.

We report the monthly test rates for each of these groups in Figure 3. We also report sample sizes, exact tests rates, and bounds in Appendix Tables A.6 and A.7, by month and cause. All causes had relatively low test rates in March before rapid increases in April and May. By June there are some differences in the test rates, with higher rates for the injury, accident, and AMI/stroke admissions, and less testing for cancer and labor and delivery.

We report the bounds on prevalence by cause of admission in Figure 4. In March there is little testing; the bounds are wide and uninformative. The bounds tighten in April, May, and especially June. Importantly, we do not see obvious, systematic differences across the groups. Typically the bounds overlap, and there is no strong evidence that the upper or lower bounds differ by cause of admission. It is true that in June, cancer patients had zero positive tests, and the upper bound for appendicitis and vehicle accidents was below the lower bounds for injuries, other accidents, and AMI/stroke. However in April and May the cancer and appendicitis/vehicle patients had similar or even higher upper bounds than did the injury and other accident patients. This evidence shows that patients admitted to the hospital for different reasons and with different demographic profiles are all nonetheless tested at a high rate and with similar bounds on prevalence. This is perhaps reassuring for the view that pooling many distinct causes of admissions can nonetheless generate meaningful bounds on prevalence.

---

[10] These groups are not necessarily mutually exclusive, and in particular there is overlap between injury and accidents.

5 Assessing the hospital representativeness assumptions

The results so far show that the bounds on prevalence are much tighter for the non-ICLI hospitalized population than for the population as a whole. This tighter bound is informative for general population prevalence only under assumptions about hospital representativeness, either a monotonicity assumption or an equal prevalence assumption. How valid are these assumptions? Assessing them directly is of course impossible because we lack data on prevalence in the population as a whole or in the hospital sample.

We have already provided one piece of indirect evidence in support of our hospital representativeness assumptions. The non-ICLI and clear-cause samples generate similar bounds, and, within the clear-cause sample, there are not large differences in bounds across different causes of admission. This suggests that prevalence does not vary with the exact set of hospitalizations studied, although of course this does not prove monotonicity or representativeness.

In this section, we provide two additional pieces of evidence on the hospital IV assumptions. First we show that the hospital bounds are consistent with the estimates of population prevalence from the Indiana COVID-19 Random Sample Study (Menachemi et al., 2020; Richard M. Fairbanks School of Public Health, 2020).[11] Second, we compare the hospital sample to the general population in terms of their likelihood of prior testing (prior to the hospital data) and the test rate of their home counties. We take these to be proxies for their concern about COVID, although other interpretations are possible.

5.1 Comparison to random sample testing

A valuable benchmark for the hospital-based prevalence bounds comes from a large-scale study of SARS-CoV-2 prevalence in Indiana. The study invited a representative sample of Indiana residents (aged 12 and older) to obtain a SARS-CoV-2 test. The first wave of the study took place April 25-29, and the second wave took place June 3-7. The preliminary results are reported in Menachemi et al. (2020) and Richard M. Fairbanks School of Public Health (2020). The response rate was roughly 25 percent, and no attempt was made to correct for non-random response. Nonetheless this survey appears to be the best benchmark available. We report the point estimates for prevalence (assuming random nonresponse) and their confidence intervals in the top panel of Table 2. The first

---

[11] Our data do not contain the test results from the Random Sample Study, so we compare our bounds to the published results.

wave estimates 1.7 percent prevalence and the second 0.5 percent.[12]

We compare our prevalence bound during the same time periods in the bottom panel of the table. We limit our sample to tests of people aged 12 and older, for comparison with the population study. Using population testing we obtain very wide bounds that contain the random sample study estimates. This fact provides some support for the test monotonicity assumption. For both the non-COVID hospitalization and clear-cause hospitalization samples, the bounds are much tighter. Both sets of upper and lower bounds contain the April prevalence estimate, and both sets exclude the June estimate point estimates. However the confidence interval for the June 3-7 point estimates overlaps substantially with the non-ICLI and clear-cause bounds. Thus for both dates the prevalence point estimates are consistent with the bounds obtained from the non-COVID hospitalizations. As a comparison we also report the bounds from the ICLI-related hospitalizations, which always exclude the random sample estimates.

5.2   Comparison of prior testing and community testing

A standard way of measuring representativeness is to compare the distribution of covariates in a study population to their distribution in the target population. In our case, this approach is most convincing if we have well-measured covariates that proxy for having COVID-19. Two candidate covariates are the community SARS-CoV-2 testing rate and the prior testing rate. The idea behind these proxies is that people who come from areas with high test rates, or who have been tested in the past, may themselves have a higher current likelihood of having COVID-19.

To operationalize these measures, we define the community testing rate for person i as the fraction of people in i's county who have ever been tested, as of the end of our sample period. We define the prior test rate of person i as of date t as the probability that i was tested at least once during the week-long period $[t - 15, t - 9]$. We focus on this window because it is the second week prior to our hospital testing window (which runs from $t - 2$ to $t + 4$ for a patient admitted at t). We allow for a week of time to elapse between the hospitalization and the "prior" testing because it is possible that some pre-hospital testing would occur in the window $[t - 8, t - 3]$. When studying prior tests, we limit the sample to each person's first hospitalization after March 1, 2020, to avoid picking up the higher testing that mechanically results from the fact that people hospitalized once are more likely than the general population to have been previously hospitalized. As with

---

[12] The estimates in Table 2 are slightly different from those reported by Richard M. Fairbanks School of Public Health (2020). We report updated calculations, based on correspondence with the authors.

our bounds, here we weight the data to match the population age distribution.

Table 3 shows the community testing rate. The average county has a testing rate of 3.5%, with an interquartile range of 2.5% to 4.3%. The average person lives in a county with a test rate of 3.6%. The average non-ICLI hospitalized patient comes from a county with a test rate of 4.1%: for clear-cause hospitalizations it is 4.0%, and for ICLI hospitalizations it is 4.2%. These rates are all significantly different from the population average.

Figure 5 shows the prior testing rate as a function of admission date for the non-ICLI hospitalization sample, the clear-cause hospitalization sample, and the general population (for which the prior test rate on day t is defined as the fraction tested between $t - 15$ and $t - 9$). The rates in the hospitalization samples are initially close to the population rate (when testing is low in general), but the lines diverge. By the last week of the sample, the prior testing rate is about 1.5 percentage points in the hospitalization samples, relative to approximately .75 percentage points in the population. Although the weekly differences are not individually statistically significant, overall the greater rate of prior testing is consistent with positive selection into hospitalization. However it is also consistent with the possibility that the hospitalization sample simply has more contact with the medical sector, resulting in greater testing at a given SARS-CoV-2 prevalence.

6   Conclusion

We have calculated weekly bounds on the prevalence of SARS-CoV-2 for the Indiana population as a whole and for three hospitalized populations: people hospitalized for influenza- and COVID-like illness, people hospitalized for other reasons, and people with clear (and clearly not COVID) causes of hospitalization. The bounds are valid under weak monotonicity assumptions. The bounds for the general population are wide but narrow over time. The bounds for the hospitalized population are much tighter, because the hospitalized populations are tested at a much higher rate than the general population.

The hospitalized populations are informative for the general population only under additional, stronger assumptions. In particular, if the hospitalized population is representative of the general population in terms of SARS-CoV-2 prevalence, then both the upper and lower bounds are valid. If the hospitalized population has a higher prevalence than the general population, then only the upper bound is valid. We assess these assumptions in multiple ways. We find that the non-COVID hospitalized population bounds contained the point estimate of prevalence from a random sample in late April. By early June the point estimate from the random sample was below our lower bound, although we cannot reject that it was inside the bound. Hospitalized patients also appear to be tested

for SARS-CoV-2 at somewhat higher rates than the general population, even outside the hospital. This last fact suggests that the representativeness assumption may be violated (although the magnitude may not be too severe), but both are consistent with a weaker monotonicity condition. Even under this monotonicity condition, the hospital data are still useful, bringing down the upper bound on population prevalence by a third or more.

We believe that the main promise of the non-COVID hospitalization population is that it can provide near-real time information about population prevalence. Only three numbers are necessary to calculate bounds from the non-COVID hospitalizations: the count of non-COVID admissions, the number of tests among this group, and the number of positive results. Although these numbers are not currently reported, many states already report related numbers, including both the number of COVID tests and the count of ICLI-related hospitalizations. Thus the infrastructure largely exists already to calculate these bounds. The results here help validate this approach for real-time surveillance.

References

Abbott, B. and I. Lovett (2020, July). Covid-19 test shortages prompt health authorities to narrow access. *Wall Street Journal*. `https://www.wsj.com/articles/covid-19-test-shortages-prompt-health-authorities-to-narrow-access-115955`

Allcott, H., L. Boxell, J. Conway, M. Gentzkow, M. Thaler, and D. Y. Yang (2020). Polarization and public health: Partisan differences in social distancing during the coronavirus pandemic. *NBER Working Paper* (w26946).

Arizona Department of Health Services (2020). Data dashbaord. `https://www.azdhs.gov/preparedness/epidemiology-disease-control/infectious-disease-epidemiology/covid-19/dashboards/index.php`. Last accessed July 15, 2020.

Armed Forces Health Surveillance Center (2015, October). Influenza-like illness. AFHSC Standard Case Definitions, available at `https://health.mil/Reference-Center/Publications/2015/10/01/Influenza-Like-Illness`, last accessed June 28, 2020.

Brookmeyer, R. and M. H. Gail (1988). A method for obtaining short-term projections and lower bounds on the size of the aids epidemic. *Journal of the American Statistical Association 83*(402), 301–308.

California Department of Health (2020). Covid-19. `https://www.cdph.ca.gov/Programs/CID/DCDC/Pages/Immunization/ncov2019.aspx`. Last accessed July 15, 2020.

Center for Disease Control and Prevention (2020, April). Icd-10-cm official coding and reporting guidelines april 1, 2020 through september 30, 2020. Available at `https://www.cdc.gov/nchs/data/icd/COVID-19-guidelines-final.pdf`, last accessed June 28, 2020.

Colorado Department of Public Health and Environment (2020). Case data. `https://covid19.colorado.gov/data/case-data`. Last accessed July 15, 2020.

Egan, J. R. and I. M. Hall (2015). A review of back-calculation techniques and their potential to inform mitigation strategies with application to non-transmissible acute infectious diseases. *Journal of The Royal Society Interface 12*(106), 20150096.

Flaxman, S., S. Mishra, A. Gandy, H. J. T. Unwin, T. A. Mellan, H. Coupland, C. Whittaker, H. Zhu, T. Berah, J. W. Eaton, et al. (2020). Estimating the effects of non-pharmaceutical interventions on covid-19 in europe. *Nature*, 1–8.

Grannis, S. J., P. G. Biondich, B. W. Mamlin, M. G. Wilson, L. Jones, and J. M. Overhage (2005). How disease surveillance systems can serve as practical building blocks for a health information infrastructure: the indiana experience. In *AMIA Annual Symposium Proceedings*, Volume 2005, pp. 286. American Medical Informatics Association.

Gudbjartsson, D. F., A. Helgason, H. Jonsson, O. T. Magnusson, P. Melsted, G. L. Norddahl, J. Saemundsdottir, A. Sigurdsson, P. Sulem, A. B. Agustsdottir, et al. (2020). Spread of sars-cov-2 in the icelandic population. *New England Journal of Medicine*.

Gupta, S., T. D. Nguyen, F. Lozano Rojas, S. Raman, B. Lee, A. Bento, K. I. Simon, and C. Wing (2020). Tracking public and private response to the covid-19 epidemic: Evidence from state and local government actions. Technical report, National Bureau of Economic Research.

Gupta, S., K. I. Simon, and C. Wing (2020). Mandated and voluntary social distancing during the covid-19 epidemic. *Brookings Papers on Economic Activity (Conference Draft)*.

Illinois Department of Public Health (2020a). Covid-19 hospital resource utilization. `https://www.dph.illinois.gov/covid19/hospitalization-utilization`. Last accessed July 15, 2020.

Illinois Department of Public Health (2020b). Covid-19 syndromic surveillance. `https://www.dph.illinois.gov/covid19/syndromic-surveillance`. Last accessed July 15, 2020.

Indiana State Department of Health (2020). Novel coronavirus (covid-19). `coronavirus.in.gov`, last accessed July 24, 2020.

Kremer, M. (1996). Integrating behavioral choice into epidemiological models of aids. *The Quarterly Journal of Economics 111*(2), 549–573.

Manski, C. F. (1999). *Identification problems in the social sciences*. Harvard University Press.

Manski, C. F. and F. Molinari (2020). Estimating the covid-19 infection rate: Anatomy of an inference problem. *Journal of Econometrics*.

Menachemi, N., C. Yiannoutsos, B. Dixon, and et al. (2020). Population point prevalence of sars-cov-2 infection based on a statewide random sample indiana, april 2529, 2020. *MMWR Morbidity and Mortality Weekly Report*, 960–964. doi: `http://dx.doi.org/10.15585/mmwr.mm6929e1external`.

Mississippi State Department of Health (2020). Interactive chart: Mississippi covid-19 hospitalizations. `https://msdh.ms.gov/msdhsite/_static/14,21994,420,873.html`. Last accessed July 15, 2020.

Peci, A., A.-L. Winter, E.-C. King, J. Blair, and J. B. Gubbay (2014). Performance of rapid influenza diagnostic testing in outbreak settings. *Journal of clinical microbiology 52*(12), 4309–4317.

Philipson, T. (1996). Private vaccination and public health: an empirical examination for us measles. *Journal of Human Resources*, 611–630.

Philipson, T. (2000). Economic epidemiology and infectious diseases. In *Handbook of Health Economics*, Volume 1, pp. 1761–1799. Elsevier.

Richard M. Fairbanks School of Public Health (2020). Iupui, isdh release findings from phase 2 of covid-19 testing in indiana. Available at `https://fsph.iupui.edu/news-events/news/iupui-isdh-release-findings-from-phase-2-of-covid-19-testing-in-indiana.html`, last accessed July 5, 2020.

Russell, T. W., J. Hellewell, C. I. Jarvis, K. Van Zandvoort, S. Abbott, R. Ratnayake, S. Flasche, R. M. Eggo, W. J. Edmunds, A. J. Kucharski, et al. (2020). Estimating the infection and case fatality ratio for coronavirus disease (covid-19) using age-adjusted data from the outbreak on the diamond princess cruise ship, february 2020. *Eurosurveillance 25*(12), 2000256.

Salje, H., C. T. Kiem, N. Lefrancq, N. Courtejoie, P. Bosetti, J. Paireau, A. Andronico, N. Hozé, J. Richet, C.-L. Dubost, et al. (2020). Estimating the burden of sars-cov-2 in france. *Science*.

Stock, J. H., K. M. Aspelund, M. Droste, and C. D. Walker (2020). Identification and estimation of undetected covid-19 cases using testing data from iceland. *medRxiv*.

Sutton, D., K. Fuchs, M. Dalton, and D. Goffman (2020). Universal screening for sars-cov-2 in women admitted for delivery. *New England Journal of Medicine 382*(22), 2163–2164.

Tennessee Department of Health (2020). Current covid-19 hospitalizations. `https://www.tn.gov/health/cedep/ncov/data/hospitalization-data/current-covid-hospitalizations.html`. Last accessed July 15, 2020.

The COVID Tracking Project (2020). The covid tracking project. `https://covidtracking.com/`. Last accessed July 15,2020.

UCSF Health Hospital Epidemiology and Infection Prevention (2020). Covid-19 diagnostic testing. `https://infectioncontrol.ucsfmedicalcenter.org/sites/g/files/tkssra4681/f/COVID-19-testing-reference-sheet.pdf`, last accessed July 31, 2020.

United States Census Bureau (2019, July). Quick facts: Indiana. Available at `https://www.census.gov/quickfacts/IN`, last accessed July 4, 2020.

Verity, R., L. C. Okell, I. Dorigatti, P. Winskill, C. Whittaker, N. Imai, G. Cuomo-Dannenburg, H. Thompson, P. G. Walker, H. Fu, et al. (2020). Estimates of the severity of coronavirus disease 2019: a model-based analysis. *The Lancet infectious diseases*.

Vermont Department of Health (2020). Current activity in vermont. `https://www.healthvermont.gov/response/coronavirus-covid-19/current-activity-vermont`. Last accessed July 15, 2020.

Wing, C. (2010). *Three essays on voluntary HIV testing and the HIV epidemic*. Syracuse University.

Yang, B. Y., L. M. Barnard, J. M. Emert, C. Drucker, L. Schwarcz, C. R. Counts, D. L. Murphy, S. Guan, K. Kume, K. Rodriquez, et al. (2020). Clinical characteristics of patients with coronavirus disease 2019 (covid-19) receiving emergency medical services in king county, washington. *JAMA Network Open 3*(7), e2014549–e2014549.

Figure 1: Weekly test rates by sample
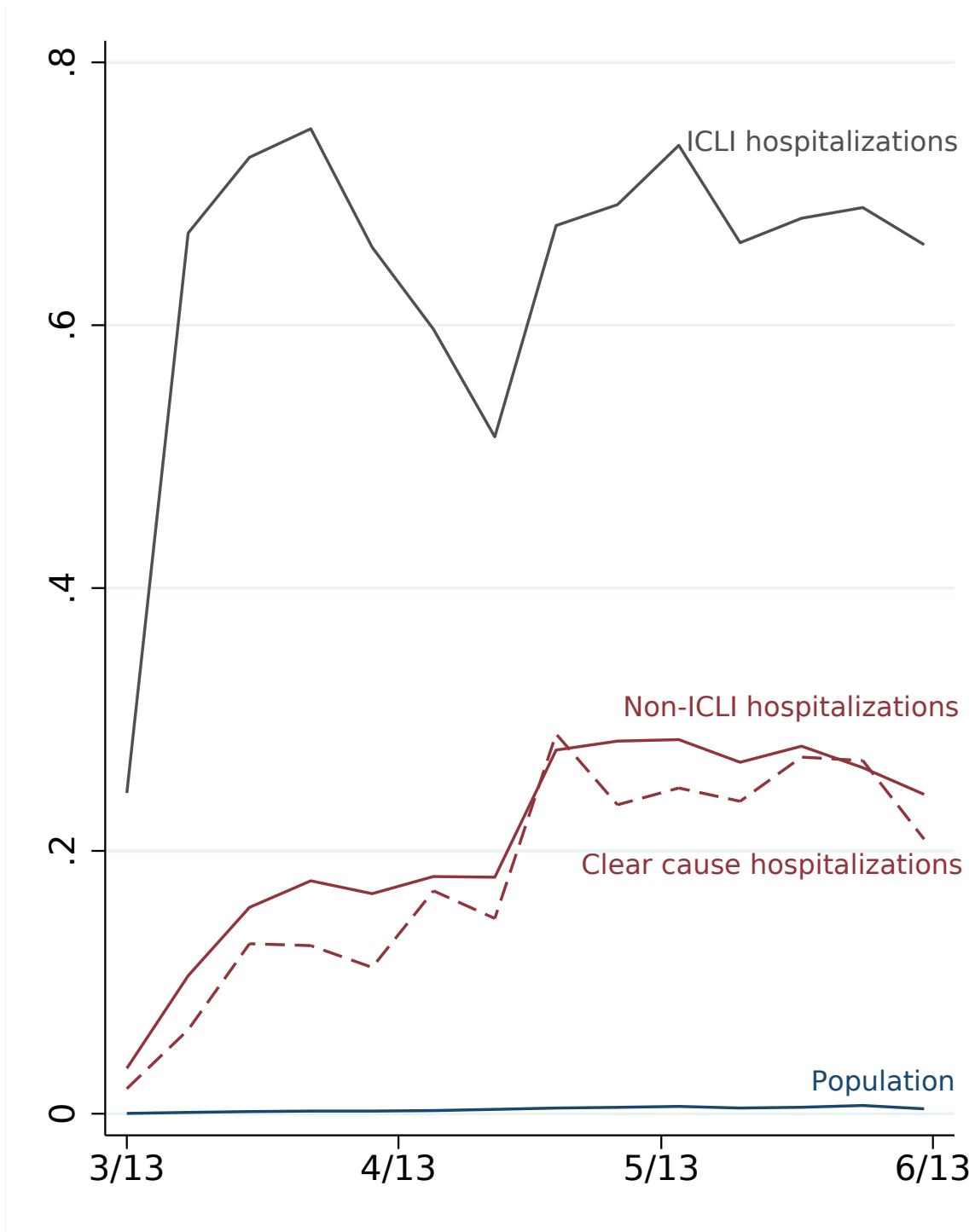
Notes: Figure plots the test rate in each seven-day period of our data, for four samples: the general population, ICLI hospitalizations, non-ICLI hospitalizations, and clear cause hospitalizations. ICLI hospitalizations have at least one diagnosis for influenza-like or COVID-like illness. Clear cause hospitalizations are hospitalizations for cancer, labor and delivery, AMI, stroke, fracture or crush, open wound, appendicitis, or accidents (vehicle or other). See Appendix B for definitions. For the general population, the test rate is the fraction of people tested at least once in that week. For the hospitalizations, the test rate is the fraction of hospitalizations admitted in that week with a test between two days prior to admission and four days after.

Figure 2: Weekly bounds on prevalence under test monotonicity, by sample
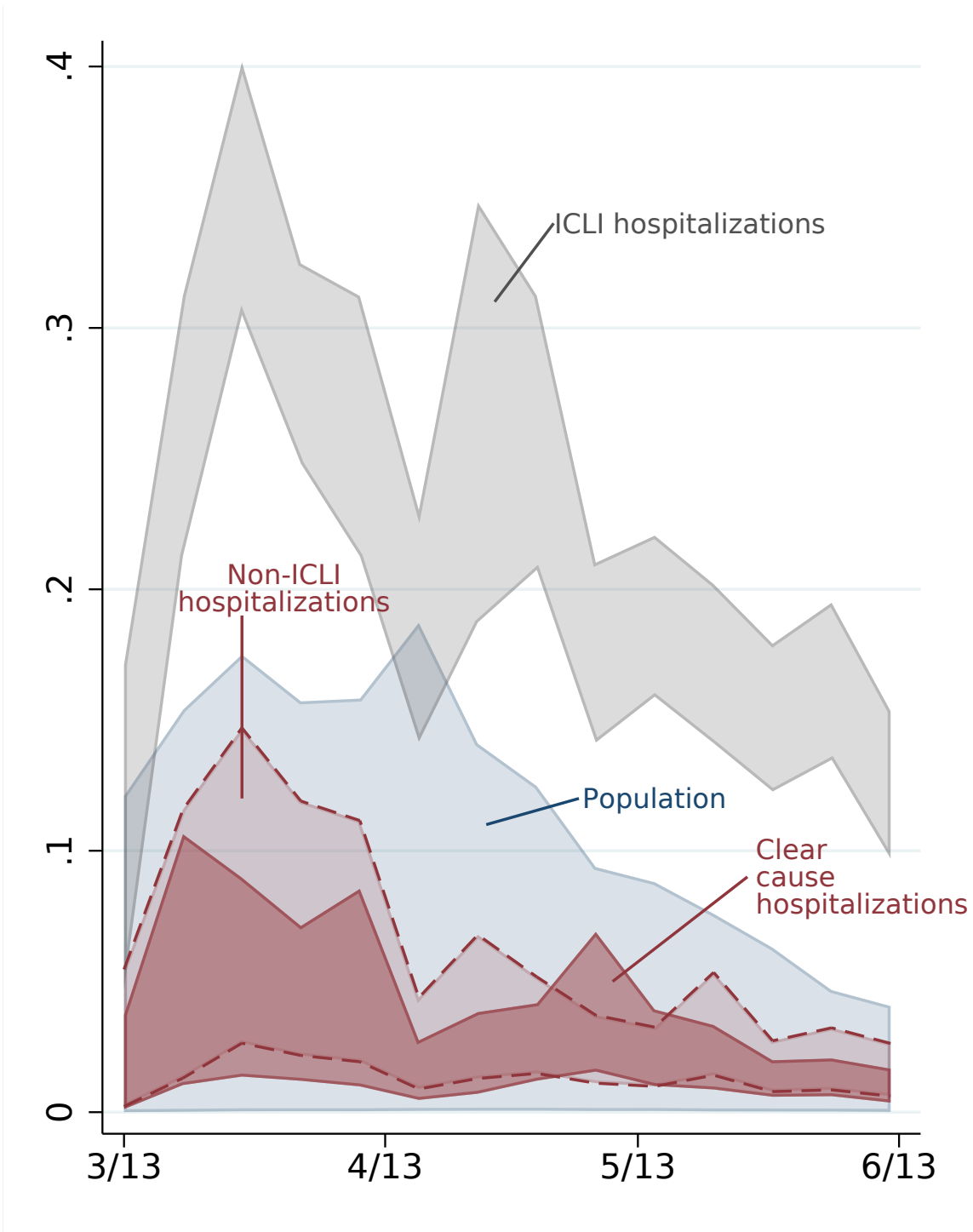
Notes: Figure plots the weekly bounds on prevalence under test monotonicity, for four samples: the general population, ICLI hospitalizations, non-ICLI hospitalizations, and clear cause hospitalizations. The lower bound is the confirmed positive rate and the upper bound is the positivity rate. ICLI hospitalizations have at least one diagnosis for influenza-like or COVID-like illness. Clear cause hospitalizations are hospitalizations for cancer, labor and delivery, AMI, stroke, fracture or crush, open wound, appendicitis, or accidents (vehicle or other). See Appendix B for definitions.

Figure 3: Monthly test rates by cause of admission
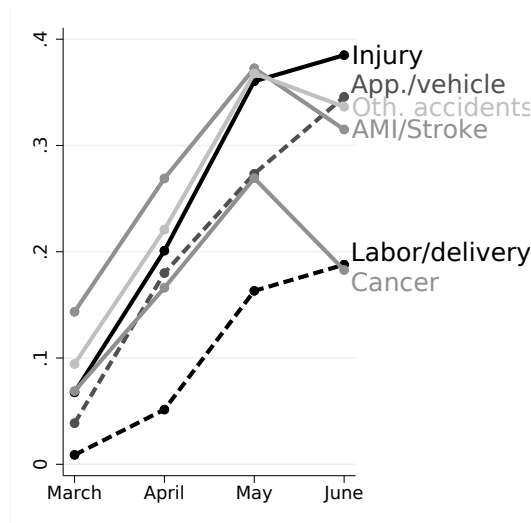
Notes: Figure plots the monthly test rates among people admitted to the hospital for the indicated condition. See Appendix B for definitions.

Figure 4: Monthly bounds on prevalence by cause of admission under test monotonicity
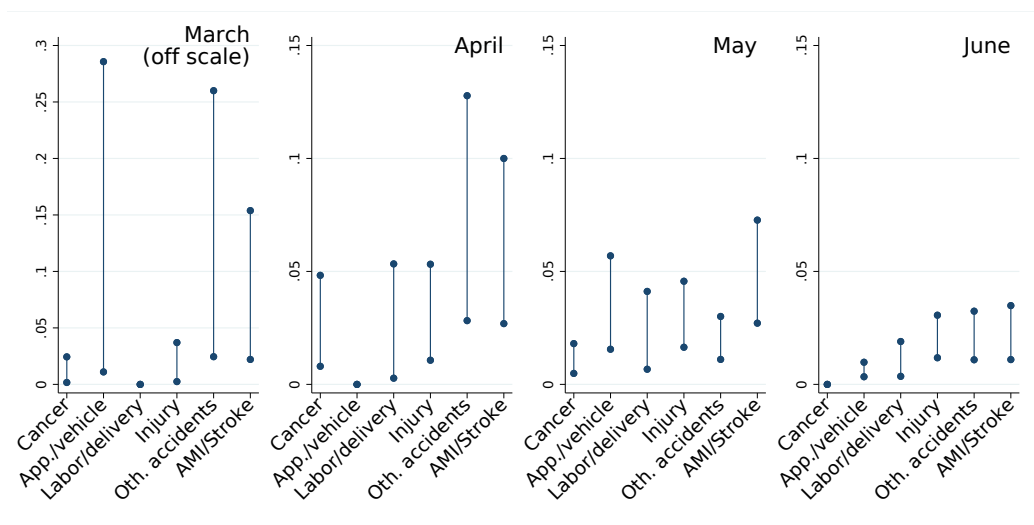
Notes: Figure plots the lower and upper bounds on prevalence, by cause of cause of admission and month, assuming test monotonicity. See Appendix B for definitions.

Figure 5: Prior test rates, population and hospitalization samples
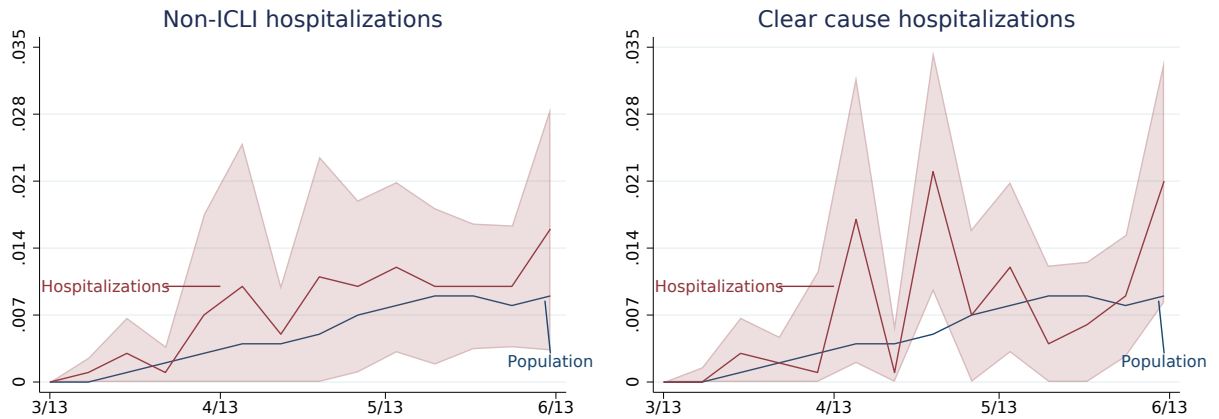
Notes: The prior test rate is the fraction of the group at date t that was tested between $t - 15$ and $t - 9$. Figure plots the average prior test rate for the population, for non-ICLI hospitalizations (in the left panel) and for clear cause hospitalizations (right panel). The shaded area is the 95% confidence interval for each week and hospitalization sample.

Table 1: Person-level summary statistics

			Hospitalized				
Sample	Full State (1)	Ever Tested (2)	Ever (3)	Has Diagnosis (4)	Not ICLI (5)	Clear cause (6)	ICLI (7)
Age as of 1/1/2020							
Born after 1/1		0.002	0.110	0.120	0.134	0.003	0.003
0-17	0.237	0.045	0.038	0.033	0.030	0.026	0.047
18-29	0.166	0.152	0.119	0.120	0.129	0.179	0.039
30-50	0.250	0.296	0.183	0.179	0.183	0.184	0.137
50-64	0.197	0.256	0.204	0.199	0.190	0.193	0.263
65-74	0.087	0.137	0.163	0.160	0.153	0.173	0.222
75+	0.063	0.111	0.183	0.189	0.179	0.243	0.288
Age unknown		0.028	0.001	0.001	0.001	0.000	0.000
Gender							
Male	0.493	0.402	0.417	0.429	0.422	0.391	0.486
Female	0.507	0.569	0.582	0.570	0.577	0.609	0.514
Unknown		0.015	0.000	0.000	0.000	0.000	0.000
Race/ethnicity							
White	0.848	0.853	0.858	0.826	0.829	0.851	0.806
Black	0.099	0.125	0.121	0.149	0.146	0.123	0.175
Race unknown		0.000	0.000	0.000	0.000	0.000	0.000
Test variables							
Ever tested	0.042	1.000	0.214	0.254	0.225	0.256	0.523
Confirmed positive	0.004	0.097	0.025	0.034	0.017	0.017	0.154
People	6,637,426	280,816	256,062	153,716	137,176	30,619	23,014
Counties	92	92	92	92	92	92	90

Notes: Column 1 reports characteristics for the set of people appearing in the test data, and columns 2-6 for people appearing the hospital data, ever (column 2), with at least one diagnosis (column 3), at least one non-ICLI hospitalization for ICLI (column 4), at least one clear cause hospitalization (column 5, see text for details), or at least one ICLI hospitalization with a diagnosis and not for ICLI (column 6).

Table 2: Do our bounds contain estimates of prevalence from random-sample testing?

Time period	April 25-29	June 3 -7
Random Sample Study		
Prevalence estimates	0.017	0.005
95% confidence interval	(0.011, 0.025)	(0.002, 0.013)
Bounds from...		
Population testing	[0.000, 0.145]	[0.000, 0.056]
Non-ICLI hospitalizations	[0.015, 0.069]	[0.009, 0.031]
Clear cause hospitalizations	[0.009, 0.042]	[0.006, 0.018]
ICLI-related hospitalizations	[0.242, 0.415]	[0.166, 0.261]

Notes: The first two rows of the table report the estimated population prevalence and 95% confidence interval from the Indiana COVID-19 Random Sample Study, conducted over the indicated dates, which assumes random nonresponse (Menachemi et al., 2020; Richard M. Fairbanks School of Public Health, 2020). The remaining rows report the (age-adjusted) bounds on prevalence from our different samples: population testing, non-ICLI hospitalizations, clear cause hospitalizations, and ICLI-related hospitalizations. We limit our sample to people aged 12 and older, for consistency with the Random Sample Study.

Table 3: County testing rates

	County test rate
County-level	
Average county	.035
25th percentile	.025
75th percentile	.043
Person level	
Average person	.036
Non-ICLI hospitalization	.041
T-statistic	157.4
Clear cause hospitalization	.04
T-statistic	66.4
ICLI hospitalization	.042
T-statistic	78.8

Notes: The county test rate is the share of the county population tested at least once in our test data. Table reports county-level statistics, as well as the average county test rates for the general population, the non-ICLI hospitalizations, clear cause hospitalizations, and ICLI hospitalizations, as well as t-statistic for the null hypothesis that the average person and the average hospitalization have the same county test rate.

A Appendix figures and tables

Figure A.1: Positive cases in our data and on state dashboard
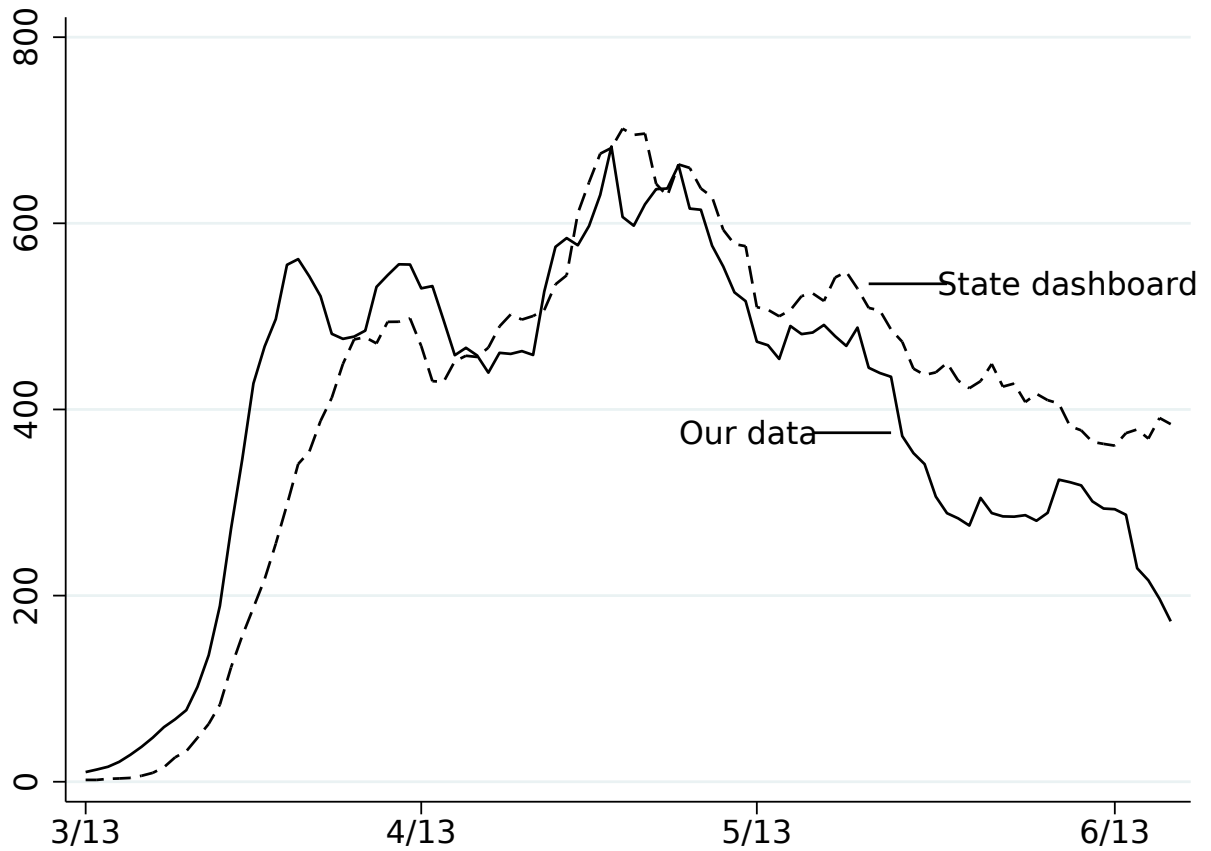
Notes: Figure plots 7-day moving average of the number of positive cases reported on Indiana's COVID-19 dashboard (Indiana State Department of Health, 2020), as well as the number of positive cases observed in our data.

Table A.1: Weekly test rates, by sample

| Sample | Population | Non-ICLI | | Clear Cause | | ICLI | |
| | % Tested | N | % Tested | N | % Tested | N | % Tested |
Week	(1)	(2)	(3)	(4)	(5)	(6)	(7)
3/13/2020	0.000	5,937	0.035	1,342	0.019	1,102	0.244
3/20/2020	0.001	5,054	0.105	1,262	0.064	1,179	0.670
3/27/2020	0.002	4,585	0.157	1,178	0.129	1,292	0.728
4/03/2020	0.002	4,573	0.177	1,246	0.128	1,150	0.749
4/10/2020	0.002	4,702	0.167	1,225	0.111	1,041	0.659
4/17/2020	0.002	4,913	0.180	1,329	0.170	1,146	0.597
4/24/2020	0.003	5,126	0.180	1,377	0.149	1,052	0.515
5/01/2020	0.004	5,618	0.277	1,420	0.289	1,143	0.676
5/08/2020	0.005	6,031	0.284	1,531	0.235	1,050	0.692
5/15/2020	0.006	6,658	0.285	1,549	0.248	1,081	0.737
5/22/2020	0.004	6,663	0.267	1,565	0.238	1,026	0.663
5/29/2020	0.005	6,763	0.280	1,542	0.271	947	0.681
6/05/2020	0.006	7,172	0.263	1,684	0.269	943	0.690
6/12/2020	0.004	6,091	0.243	1,411	0.209	776	0.661

Notes: Table reports the weekly test rate for the population, and the number of hospitalizations and test rate, by type of hospitalizations, weighted to match the population age distribution.. (The population size is 6.64 million in all weeks.) ICLI hospitalizations have at least one diagnosis for influenza-like or COVID-like illness. Clear cause hospitalizations are hospitalizations for cancer, labor and delivery, AMI, stroke, fracture or crush, open wound, appendicitis, or accidents (vehicle or other). See Appendix B for definitions.

Table A.2: Weekly test rates, by sample, not age weighted

Sample	Population	Non-ICLI		Clear Cause		ICLI	
	% Tested	N	% Tested	N	% Tested	N	% Tested
Week	(1)	(2)	(3)	(4)	(5)	(6)	(7)
3/13/2020	0.000	5,937	0.035	1,342	0.019	1,102	0.244
3/20/2020	0.001	5,054	0.105	1,262	0.064	1,179	0.670
3/27/2020	0.002	4,585	0.157	1,178	0.129	1,292	0.728
4/03/2020	0.002	4,573	0.177	1,246	0.128	1,150	0.749
4/10/2020	0.002	4,702	0.167	1,225	0.111	1,041	0.659
4/17/2020	0.002	4,913	0.180	1,329	0.170	1,146	0.597
4/24/2020	0.003	5,126	0.180	1,377	0.149	1,052	0.515
5/01/2020	0.004	5,618	0.277	1,420	0.289	1,143	0.676
5/08/2020	0.005	6,031	0.284	1,531	0.235	1,050	0.692
5/15/2020	0.006	6,658	0.285	1,549	0.248	1,081	0.737
5/22/2020	0.004	6,663	0.267	1,565	0.238	1,026	0.663
5/29/2020	0.005	6,763	0.280	1,542	0.271	947	0.681
6/05/2020	0.006	7,172	0.263	1,684	0.269	943	0.690
6/12/2020	0.004	6,091	0.243	1,411	0.209	776	0.661

Notes: Table reports the weekly test rate for the population, and the number of hospitalizations and test rate, by type of hospitalizations. (The population size is 6.64 million in all weeks.) ICLI hospitalizations have at least one diagnosis for influenza-like or COVID-like illness. Clear cause hospitalizations are hospitalizations for cancer, labor and delivery, AMI, stroke, fracture or crush, open wound, appendicitis, or accidents (vehicle or other). See Appendix B for definitions.

Table A.3: Weekly bounds on prevalence under monotonicity, by sample

Sample	Population		Non-ICLI		Clear Cause		ICLI	
	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
Week	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
3/13/2020	0.0000	0.121	0.0022	0.055	0.0012	0.037	0.0468	0.171
3/20/2020	0.0002	0.154	0.0130	0.116	0.0104	0.106	0.2128	0.312
3/27/2020	0.0004	0.175	0.0264	0.147	0.0136	0.090	0.3050	0.402
4/03/2020	0.0004	0.157	0.0217	0.119	0.0120	0.071	0.2480	0.325
4/10/2020	0.0004	0.158	0.0193	0.112	0.0099	0.085	0.2127	0.312
4/17/2020	0.0005	0.187	0.0089	0.044	0.0047	0.027	0.1417	0.230
4/24/2020	0.0006	0.141	0.0129	0.068	0.0071	0.038	0.1873	0.348
5/01/2020	0.0006	0.125	0.0148	0.052	0.0120	0.042	0.2074	0.312
5/08/2020	0.0005	0.094	0.0111	0.037	0.0155	0.069	0.1414	0.210
5/15/2020	0.0005	0.088	0.0099	0.033	0.0100	0.039	0.1590	0.221
5/22/2020	0.0003	0.076	0.0141	0.053	0.0087	0.033	0.1411	0.202
5/29/2020	0.0003	0.063	0.0079	0.027	0.0059	0.020	0.1227	0.179
6/05/2020	0.0003	0.047	0.0085	0.032	0.0061	0.021	0.1348	0.195
6/12/2020	0.0001	0.041	0.0062	0.026	0.0037	0.017	0.0970	0.153

Notes: Table reports weekly bounds on COVID prevalence in the indicated sample under test monotonicity, weighted to match the population age distribution. The lower bound is the confirmed positive rate and the upper bound is the positivity rate. ICLI hospitalizations have at least one diagnosis for influenza-like or COVID-like illness. Clear cause hospitalizations are hospitalizations for cancer, labor and delivery, AMI, stroke, fracture or crush, open wound, appendicitis, or accidents (vehicle or other). See Appendix B for definitions.

Table A.4: Weekly bounds on prevalence under monotonicity, by sample, not age-weighted

Sample	Population		Non-ICLI		Clear Cause		ICLI	
	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
Week	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
3/13/2020	0.0000	0.150	0.0036	0.112	0.0022	0.079	0.0534	0.228
3/20/2020	0.0002	0.189	0.0153	0.158	0.0142	0.176	0.2713	0.416
3/27/2020	0.0004	0.217	0.0303	0.196	0.0195	0.128	0.3927	0.510
4/03/2020	0.0004	0.191	0.0252	0.141	0.0223	0.133	0.3131	0.410
4/10/2020	0.0004	0.189	0.0232	0.128	0.0155	0.093	0.2620	0.374
4/17/2020	0.0005	0.214	0.0124	0.068	0.0090	0.046	0.1922	0.315
4/24/2020	0.0006	0.167	0.0149	0.081	0.0109	0.062	0.1896	0.369
5/01/2020	0.0006	0.131	0.0157	0.057	0.0169	0.052	0.2047	0.308
5/08/2020	0.0005	0.097	0.0126	0.046	0.0137	0.047	0.1698	0.248
5/15/2020	0.0005	0.087	0.0122	0.043	0.0148	0.047	0.1616	0.224
5/22/2020	0.0003	0.074	0.0134	0.050	0.0159	0.055	0.1654	0.236
5/29/2020	0.0003	0.060	0.0088	0.032	0.0084	0.027	0.1309	0.188
6/05/2020	0.0003	0.045	0.0089	0.033	0.0089	0.029	0.1391	0.207
6/12/2020	0.0002	0.036	0.0058	0.024	0.0049	0.019	0.1121	0.183

Notes: Table reports weekly bounds on COVID prevalence in the indicated sample under test monotonicity. The lower bound is the confirmed positive rate and the upper bound is the positivity rate. ICLI hospitalizations have at least one diagnosis for influenza-like or COVID-like illness. Clear cause hospitalizations are hospitalizations for cancer, labor and delivery, AMI, stroke, fracture or crush, open wound, appendicitis, or accidents (vehicle or other). See Appendix B for definitions.

Table A.5: Demographics and test rates among hospitalized patients, by group

| | Number of | Age | | | | | | |
Group	Admissions	Newborn	0-17	18-29	30-49	50-64	65-74	>74
All	236,267	0.085	0.029	0.108	0.185	0.226	0.176	0.191
Has diagnosis	106,354	0.108	0.025	0.117	0.182	0.209	0.167	0.192
ICLI	14,992	0.004	0.024	0.039	0.153	0.274	0.231	0.276
Non-ICLI	91,362	0.125	0.026	0.130	0.187	0.198	0.157	0.178
Clear cause	19,721	0.003	0.032	0.173	0.183	0.194	0.178	0.237
Cancer	3,159	0.000	0.052	0.027	0.119	0.330	0.288	0.184
Labor/delivery	4,690	0.010	0.015	0.595	0.380	0.000	0.000	0.000
AMI	2,567	0.000	0.000	0.005	0.110	0.324	0.273	0.288
Stroke	2,652	0.000	0.005	0.011	0.088	0.286	0.263	0.346
Fracture	4,147	0.002	0.038	0.061	0.122	0.175	0.193	0.409
Open wound	1,125	0.000	0.050	0.101	0.186	0.208	0.172	0.283
Appendicitis	633	0.000	0.212	0.196	0.275	0.185	0.085	0.047
Vehicle accident	582	0.002	0.089	0.194	0.304	0.215	0.134	0.062
Other accident	3,072	0.002	0.035	0.034	0.079	0.154	0.213	0.483

Notes: Table reports the number and age distribution of admissions, for different categories of admissions, over the time period March 13, 2020 through June 18, 2020. See Appendix B for definitions of the different causes of admissions (Cancer-other accident).

Table A.6: Test rates and prevalence bounds under monotonicity, by month and cause of admission

Month	# Admits	% Tested	Lower bound	Upper bound
A. Cancer				
March	595	0.069	0.002	0.024
April	873	0.166	0.008	0.048
May	1,029	0.269	0.005	0.018
June	662	0.183	0.000	0.000
B. Labor and delivery				
March	904	0.009	0.000	0.000
April	1,456	0.052	0.003	0.053
May	1,489	0.163	0.007	0.041
June	841	0.188	0.004	0.019
C. AMI				
March	437	0.188	0.023	0.122
April	706	0.344	0.041	0.119
May	839	0.451	0.041	0.090
June	585	0.373	0.012	0.032
D. Stroke				
March	469	0.102	0.021	0.208
April	744	0.198	0.013	0.068
May	932	0.303	0.015	0.050
June	507	0.249	0.010	0.040
D. Fractures and crushes				
March	616	0.055	0.002	0.029
April	1,109	0.210	0.011	0.052
May	1,483	0.373	0.014	0.038
June	939	0.408	0.011	0.026

Notes: Table reports monthly number of admissions for cancer, along with the percent of admissions tested in hospital, and the bound on prevalence under test monotonicity. The lower bound is the confirmed positive rate and the upper bound is the positivity rate. See Appendix B for definitions.

Table A.7: Test rates and prevalence bounds under monotonicity, by month and cause of admission (continued)

Month	# Admits	% Tested	Lower bound	Upper bound
E. Open wounds				
March	182	0.110	0.005	0.050
April	294	0.167	0.010	0.061
May	401	0.314	0.025	0.079
June	248	0.298	0.016	0.054
F. Vehicle accidents				
March	61	0.049	0.000	0.000
April	136	0.154	0.000	0.000
May	222	0.212	0.009	0.043
June	163	0.288	0.000	0.000
G. Other accidents				
March	530	0.094	0.025	0.260
April	815	0.221	0.028	0.128
May	1,085	0.368	0.011	0.030
June	642	0.336	0.011	0.032
H. Appendicitis				
March	120	0.033	0.017	0.500
April	153	0.203	0.000	0.000
May	228	0.333	0.022	0.066
June	132	0.417	0.008	0.018

Notes: Table reports monthly number of admissions for cancer, along with the percent of admissions tested in hospital, and the bound on prevalence under test monotonicity. The lower bound is the confirmed positive rate and the upper bound is the positivity rate. See Appendix B for definitions.

B Defining causes of admissions

This section provides more details on our definition of ICLI, non-ICIL, and "clear cause" hospitalization, listing the ICD-10 codes used to define each.

Following Armed Forces Health Surveillance Center (2015), the codes for influenza-like illness are B97.89, H66.9, H66.90, H66.91 H66.92, H66.93, J00, J01.9, J01.90, J06.9, J09, J09.X, J09.X1, J09.2, J09.X3, J09.X9, J10, J10.0, J10.00, J10.01, J10.08, J10.1, J10.2, J10.8, J10.81, J10.82, J10.83, J10.89, J11, J11.0, J11.00, J11.08, J11.1, J11.2, J11.8, J11.81, J11.82, J11.83, J11.89, J12.89, J12.9, J18, J18.1, J18.8, J18.9, J20.9, J40, R05, and R50.9. We say a hospitalizaiton is for an influenza-like illness if it has any of these diagnosis codes in any position. We say a hospitalization is for a COVID-like illness if it has any ICD-10 code among those that is among the CDC's lists of diagnosis codes for COVID-19 Center for Disease Control and Prevention (2020). These codes are J12.89, J20.8, J22, J40, J80, J98.8, O95.5, R05, R06.02, R50.9, U07.1, Z03.818, Z11.58, and Z20.828.

We define ICLI-related hospitalizations as ones with at least one ILI or CLI diagnosis code. We define non-ICLI related hospitalizations as hospitalized with diagnosis codes, but no ILI or CLI code.

We also define "clear cause" hospitalizations. These are hospitalizations for labor and delivery, AMI, stroke, fractures and crushes, wounds, vehicle accidents, other accidents, appendicitis, or cancer. With the exception of cancer, we define a hospitalization as belonging to one of these groups if it has any diagnosis codes for that group, listed below. Cancer is treated differently because it can be a comorbidity. We say a hospitalization is for cancer if a cancer diagnosis (listed below) is an admitting diagnosis, the primary final diagnosis, or if chemotherapy diagnosis is present. We use the following ICD-10 codes.

- **AMI** I21, I22.

- **Appendicitis** K35-K38.

- **Cancer** C00-C97 (in primary or admitting diagnosis), or Z51.0-Z51.2 (in any position).

- **Fracture/Crush** S02, S12, S22, S32, S42, S52, S62, S72, S82, S92, T02, S07, S17, S37, S47, S57, S67, S77, S87, S97, T07.

- **Labor and delivery** O60-O75, O80-O84.

- **Other accidents** W00-W99, X00-X59.

- **Stroke** I61-I64.

- **Vehicle accident** V01-V99.

- **Wound** S01, S11, S21, S31, S41, S51, S61, S71, S81, S91, T01.

C Calculating negative predictive values with test-retest data

Setup and identification Here we show how to use data on multiple tests to simultaneously identify prevalence, test error rates, and how to use this information to obtain the negative predictive value, NPV. Assume in particular that people have been tested exactly twice, with $R1_i$ the outcome of the first test and $R2_i$ the outcome of the second test for person i, and C_i person i's true infection status, which we assume is fixed between the tests. Let $p = Pr(C_i = 1)$ be the prevalence of active SARS-CoV-2 infections in this twice tested population.

Test outcomes may differ from true infection status because of test errors. In general, therefore, there are four possible sequences of test outcomes: $(0, 0), (0, 1), (1, 0), (1, 1)$. We let $P_{ab} = Pr(R1_i = a, R2_i = b)$ for $(a, b) \in \{0, 1\}^2$.

We make three strong assumptions to simplify the analysis.

1. The specificity of the test is 1. That is, $\beta = Pr(Rj_i = 0 | C_i = 0) = 1$.

2. The sensitivity of the test, $\alpha = Pr(Rj_i = 1 | C_i = 1)$, does not depend on the initial test result.

3. Retesting is random, i.e. independent of $R1_i$ and C_i.

Assumption 1 is the weakest of these assumptions. It implies that there are no false positives, which is consistent with typical practice (UCSF Health Hospital Epidemiology and Infection Prevention, 2020). The remaining assumptions are stronger. Assumption 2 says that the test errors are independent of the initial test result. It would be violated, for example, if false negatives are more common for patients with high levels of mucus, and mucus levels are correlated across test results. Assumption 3 says that retesting rates do not depend on possible testing errors. We would expect this condition to fail if highly symptomatic people with negative tests are especially likely to test negative. We view this assumption as the most suspect.

Under these assumptions, the probabilities P_{ab} simplify considerable. As the probabilities sum to one, and the assumptions imply that $P_{10} = P_{01}$, the only non-redundant probabilities are

$$P_{00} = (1 - p) + p(1 - \alpha)^2$$
$$P_{11} = p\alpha^2.$$

We can observe P_{00} and P_{11}. Solving for the unknowns p and α, we have

$$p = \frac{(P_{00} - P_{11} - 1)^2}{4P_{11}}$$

$$\alpha = \frac{2P_{11}}{1 - P_{00} + P_{11}}$$

This shows how to get p and α from two tests and the assumption that specificity equals 1. Our goal is to find the negative predictive value, but we can calculate it given knowledge of α, β and p. In general the NPV of a single test is $Pr(C_i = 0 | R_i = 0)$. Applying Bayes rule,

$$NPV = \frac{1 - p}{p(1 - \alpha) + (1 - p)}$$

Results To implement this approach, we construct a sample of all people who are tested on a given day, not tested the previous day, and then tested again in the next day. There are 16,401 such test pairs. We find $P_{00} = 0.88$ and $P_{11} = 0.11$. Nearly all the mass is on the diagonals; test results switch less than 1% of the time. This fact, together with the assumption that specificity is equal to 1, implies very low false negative rates. Plugging these values into our formula, we have $p = 0.12$ and $\alpha = 0.96$, which imply $NPV = 0.995$. Using instead, all people who are retested once within a three day period, we find similar results: $p = 0.13, \alpha = 0.91, NPV = 0.986$.

We emphasize that these estimates are valid for the twice-tested population and under assumptions 1-3, in particular, random retesting. The prevalence estimate is the prevalence among people tested twice, not the population prevalence. And it is only a valid estimate under assumptions 1-3. In reality, it is likely that retests are most common among suspected false negatives (i.e. when a highly symptomatic patient tests negative). We see some evidence for this: $P_{01} = 0.006$ and $P_{10} = 0.003$, inconsistent with the random retesting assumption. We therefore do not view our estimates of prevalence and sensitivity as definitive; rather we think of the sensitivity estimate as a lower bound on sensitivity, because we have selected a retest sample which has a disproportionate number of false negatives. As NPV is increasing in sensitivity, α, our implied estimate of $1 - NPV$ is likely an upper bound on $1 - NPV$.