WHAT CAN WE LEARN ABOUT SARS-COV-2 PREVALENCE FROM TESTING AND HOSPITAL DATA?

Daniel W. Sacks, Nir Menachemi, Peter Embí, and Coady Wing*

Abstract—Measuring the prevalence of active SARS-CoV-2 infections in the general population is difficult because tests are conducted on a small and nonrandom segment of the population. However, hospitalized patients are tested at very high rates, even those admitted for non-COVID reasons. We show how to use information on testing of non-COVID hospitalized patients to obtain tight bounds on population prevalence, under conditions weaker than those usually used. We apply our approach to the population of test and hospitalization data for Indiana, and we validate our approach. Our bounds could be constructed at relatively low cost, and for other heavily tested populations.

I. Introduction

PUBLIC and private responses to the spread of an infectious disease rely on accurate and timely estimates of disease prevalence. For example, individual precautionary behaviors respond to the prevalence of COVID-19, HIV/AIDS, influenza, and Ebola (Allcott et al., 2020; Philipson, 2000, 1996). And government decisions about nonpharmaceutical interventions, such as stay-at-home orders, school and business closures, and mask mandates, were also a reaction to the prevalence of SARS-CoV-2 (Gupta et al., 2020). Despite its importance, prevalence is often measured poorly, especially during large outbreaks, when test capacity is scarce and tests are often allocated to symptomatic patients. Simple estimators of prevalence can be misleading in these situations. For example, the fraction of the overall population that tests positive for the disease severely understates prevalence because most infected people are not tested. On the other hand, the fraction of the tested population that tests positive likely overstates prevalence because tested people are more likely to be infected than untested people.

We propose a new approach to measuring prevalence under incomplete and nonrandom testing. Previous work by Manski and Molinari (2020) suggested estimating cumulative prevalence—the fraction of people who have ever been infected—from available data under a “test monotonicity” assumption. We adapt this assumption to the problem of estimating current or point-in-time prevalence. Applying test monotonicity to the general population produces bounds that are typically wide during the COVID-19 pandemic, because testing rates are low. Our primary contribution is to introduce a new approach resulting in tighter bounds. The overarching idea is to exploit data on testing and test results from institutional environments where people are tested at unusually high rates for reasons that are unrelated to disease prevalence. We apply and validate the idea among patients hospitalized for reasons unrelated to COVID-19, such as labor and delivery (Sutton et al., 2020). But the basic strategy is more general. For example, there may also be useful variation in disease testing rates associated with airports, workplaces, conferences, and other settings. The methods we present in this paper could be used in these other situations to estimate COVID-19 prevalence or perhaps the prevalence of other diseases in a future outbreak.

The empirical results we present show that, in highly tested subpopulations such as hospitalized patients, test monotonicity implies relatively tight bounds on point-in-time prevalence. Of course, these tight bounds are only useful if the heavily tested population is in some sense representative of the general population. We therefore introduce two representativeness assumptions. The first assumption is “hospitalization monotonicity,” which requires that patients hospitalized for non-COVID reasons are (if anything) no less likely to be infected with SARS-CoV-2 than the general population. This assumption implies that the upper bound on prevalence in the hospitalized subpopulation is a valid upper bound for the whole population. Because testing rates are higher in the non-COVID hospitalized population, the assumption typically implies a tighter population upper bound, albeit with less statistical precision due to lost sample size. We also consider a second assumption—“hospitalization independence”—which is more restrictive. It says that patients hospitalized for non-COVID reasons have the same risk as the general population. Using the independence assumption means that the hospitalized upper bound remains an upper bound on population prevalence. But it also turns out that—under independence—the lower bound for the non-COVID hospitalized subsample is a lower bound for the general population.

These assumptions, though much weaker than the assumptions that are most often used to estimate prevalence—for example, that all positive cases are tested, or the population is tested at random—are nonetheless restrictive and their validity will likely vary across contexts and subpopulations. Either of these two assumptions permits some extrapolation from the non-COVID hospitalized population to the general population, but neither assumption is sufficient to point-identify prevalence. We present results from both sets of assumptions and allow readers to reach their own conclusions.

Received for publication August 4, 2020. Revision accepted for publication January 20, 2022. Editor: Xiaoxia Shi.

* Sacks: Indiana University; Menachemi: Indiana University and Regenstrief Institute, Inc.; Embí: Vanderbilt University Medical Center and Vanderbilt University; Wing: Indiana University.

We are grateful to Justin Blackburn, Seth Freedman, Aaron Kolb, Charles Manski, and Kosali Simon for helpful comments on earlier drafts, to Evengen Teal for assistance with the data, and to Agrayan Gupta for helpful research assistance. We acknowledge the Indiana University Pervasive Technology Institute for providing HPC resources that have contributed to the research results reported in this paper (https://pti.iu.edu). All errors are our own.

A supplemental appendix is available online at https://doi.org/10.1162/rest_a_01179.
We analyze data on the near-universe of COVID-19 tests and all-cause hospitalizations in Indiana in 2020. We start by using the test monotonicity condition developed by Manski and Molinari (2020) to estimate the weekly prevalence of active COVID-19 infections. These test monotonicity bounds are narrowest in the week of June 12, when they imply that between 0.05% and 4.5% of the Indiana population had an active infection. Using the hospital data and imposing our new assumptions narrows the bounds substantially. Under the hospital monotonicity assumption, the upper bound falls and we find that COVID-19 prevalence in the week of June 12 is between 0.5% and 2.2%. Under hospital independence, the bound narrows further to 0.7%–2.2%. In the average week, imposing hospitalization monotonicity on top of test monotonicity reduces the width of the prevalence bounds by about half. All three sets of bounds are transparent and simple to calculate. In the Results section, we present a range of bounds to enable readers to choose the set of results that correspond to the assumptions they find most credible.

To assess the credibility of our assumptions in the Indiana context, we present three pieces of evidence. First, we compute prevalence bounds for all COVID-unrelated hospitalizations as well as for specific categories of hospitalizations, such as vehicle accidents, appendicitis, labor and delivery, and heart attacks. The results are similar across these subpopulations, which suggests little selection across hospital conditions on the basis of COVID risk. Second, we show that hospitalized patients are drawn from areas with test rates similar to or lower than the general population, again suggesting similar COVID risk. Third, we show that our bounds contain the prevalence estimates from a random sample survey of the Indiana population (Menachemi et al., 2020; Richard M. Fairbanks School of Public Health, 2020).

Our approach complements existing methods of estimating prevalence. Two common methods are biometric survey samples and backcalculation. In a biometric survey, a representative sample from the population is tested for the disease (Menachemi et al., 2020; Richard M. Fairbanks School of Public Health, 2020; Gudbjartsson et al., 2020). The method is slow and expensive, but likely produces accurate estimates. In backcalculation studies, data on observed hospitalizations or deaths are used to infer disease prevalence at earlier dates using assumptions about the progression of the disease, hospitalization rates, and case fatality rates (Brookmeyer & Gail, 1988; Egan & Hall, 2015; Flaxman et al., 2020; Salje et al., 2020). The approach we pursue combines passively collected administrative data with distributional assumptions that are less restrictive than those used in backcalculation (Manski, 1999; Wing, 2010; Stock et al., 2020; Manski & Molinari, 2020). Because states already report test rates, positivity rates, and COVID-related hospitalization rates, an advantage of our approach is that the bounds could be calculated at little additional cost in terms of data infrastructure, and they could also be calculated for other heavily tested subpopulations. More generally, our results show that there is considerable value in reporting testing and positivity rates separately based on the reason for the test; tests that are not driven by symptoms or exposure may be particularly informative about population prevalence.

II. Inferring COVID Prevalence from Incomplete Testing

Our goal is to estimate the weekly prevalence of SARS-CoV-2 infections using the kind of administrative data that are available in most states, accounting for the fact that testing is likely to be unrepresentative. We use partial identification methods to construct upper and lower bounds on prevalence under alternative assumptions. Figure 1 gives a schematic representation of the data, assumptions, and results. Typically, more restrictive assumptions yield tighter bounds. We discuss arguments for and against key assumptions, which should help readers assess the credibility of the restrictions. However, we present estimated bounds under a variety of assumptions, so readers may focus on the approach they find most convincing.

A. Notation and Worst Case Bounds

Use $i = 1, \ldots, N$ to index the population of Indiana. $C_{it}$ is a binary indicator that person $i$ is infected with SARS-CoV-2 on date $t$. Leaving conditioning on the date implicit, the prevalence of SARS-CoV-2 in Indiana at date $t$ is $\Pr(C_{it} = 1) = \frac{1}{N} \sum_{i=1}^{N} C_{it}$. Next, let $H_{it}$ be a binary indicator that the person was hospitalized with a specified diagnosis type. Then $\Pr(C_{it} = 1|H_{it} = 1)$ is SARS-CoV-2 prevalence in the hospitalized subpopulation in Indiana on date $t$.

Let $D_{it}$ indicate whether the person was tested. The share of the population tested on date $t$ is $\Pr(D_{it} = 1)$. Prevalence among the tested is $\Pr(C_{it}|D_{it} = 1)$. $\Pr(C_{it}|D_{it} = 0)$ represents prevalence among people who were not tested. By construction, the value of $C_{it}$ is unknown for people with $D_{it} = 0$. This means that $\Pr(C_{it} = 1|D_{it} = 0)$ is not identified by the data on testing and test outcomes.

As a practical matter, testing is very rare in most of the country in any given week, so $C_{it}$ is unknown for most of the population. In the absence of any assumptions about selection into testing, the worst case bounds on prevalence are

$$L_w = \Pr(C_{it} = 1|D_{it} = 1)\Pr(D_{it} = 1),$$

confirmed positive rate

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

untested rate

These bounds define the range of prevalence that is compatible with the data. The lower and upper bounds for a given subpopulation and time period can be formed using relevant proportions from test and hospital data. For inference, we use the bootstrap (Manski & Pepper, 2000; Kreider & Pepper, 2007; Manski & Pepper, 2009); see appendix F for details.
B. Test Monotonicity

To narrow the worst case bounds, we impose a version of the “test monotonicity” assumption that Manski and Molinari (2020) proposed to compute cumulative prevalence:

**Assumption 1** (Test monotonicity). \( Pr(C_{it} = 1|D_{it} = 1) \geq Pr(C_{it} = 1|D_{it} = 0) \).

When we work with specific subpopulations, such as hospitalized patients, we use a conditional version of assumption 1, which requires that \( Pr(C_{it} = 1|D_{it} = 1, H_{it} = 1) \geq Pr(C_{it} = 1|D_{it} = 0, H_{it} = 1) \) for subpopulations defined by values of the covariate \( H_{it} \).

Assumption 1 (Test Monotonicity) requires that prevalence is at least as high among the tested as among the untested, conditional on any other covariates. Test monotonicity allows for a randomly chosen tested person to be arbitrarily more likely to test positive than a randomly chosen untested person. It also allows the possibility that the tested population may be neither positively nor negatively selected. Importantly, the assumption does not need to hold at the individual level, only in aggregate. For example, it is
possible that individual demand for testing responds heterogeneously to situational risk or institutional norms. For instance, some people might avoid testing when they experience additional risk factors, perhaps motivated by dread. Likewise some institutions in some cases might avoid ordering tests in cases in which the test results would be redundant given observed symptoms. These individual level choices would tend to reduce positive selection into testing, and such behaviors are not ruled out by the test monotonicity assumption. The test monotonicity assumption implies only that, on average, these negative selection types do not outnumber the positive selection types.

Imposing the test monotonicity condition constrains the range of prevalence in the untested population. Specifically, assumption 1 implies that \( 0 \leq \Pr(C_h = 1|D_g = 0) \leq \Pr(C_h = 1|D_g = 1) \). Under test monotonicity, the bounds on active prevalence are

\[
\begin{align*}
L_m &= \Pr(C_h = 1|D_g = 1)\Pr(D_g = 1), \\
U_m &= \Pr(C_h = 1|D_g = 1).
\end{align*}
\]

The new upper bound is the prevalence in the tested population, which is often called the test positivity rate and is widely reported. In our data, test rates are often less than 1% and positivity rates in the population are often 10% or less, so this assumption brings the upper bound down from 99% to 100% or less. These bounds address positive selection into testing because they are valid under test monotonicity, and hence they allow for even potentially extreme selection.

C. Inferring Population Prevalence from Non-COVID Hospital Patients

It is straightforward to construct test monotonicity bounds on prevalence overall and in subpopulations, such as hospitalized patients. We emphasize that the test monotonicity assumption must apply in each subpopulation.\(^1\) Testing rates are higher in hospital settings, which means the bounds are much tighter for hospital subpopulations than for the general population. Thus, assumptions that link hospital and population prevalence could substantially reduce uncertainty about population prevalence. We pursue two types of assumptions that enable extrapolation from non-COVID hospital populations to the general population: (i) hospitalization monotonicity and (ii) hospitalization independence.

Hospitalization monotonicity. For some situations or patient types, it is reasonable to assume that hospitalized patients are somewhat adversely selected on health. Applied to COVID-19, adverse selection suggests a hospitalization monotonicity assumption that SARS-CoV-2 prevalence is weakly higher in the hospitalized subpopulation than the general population. Stated formally:

**Assumption 2** (Hospitalization monotonicity). \( \Pr(C_{ht} = 1|H_{ht} = 1) \geq \Pr(C_{ht} = 1) \).

The hospitalization monotonicity assumption may be more credible for some types of hospital patients than others, and such adjustments can be handled with additional conditioning. The important point is that layering the hospital monotonicity assumption on top of the test monotonicity assumption can help reduce the width of the bounds on prevalence in both the general population and the hospital population.

For example, suppose \( U_{m,H}^H \) and \( L_{m,H}^H \) are the upper and lower bounds on prevalence in the hospitalized subpopulation under assumption 1 (test monotonicity). And let \( U_m \) and \( L_m \) represent test monotonicity bounds in the general population. Adding assumption 2 (hospitalization monotonicity) creates a cross-population restriction, which implies that the upper bound on population prevalence (\( U_m^H \)) cannot be larger than the upper bound on hospital prevalence (\( U_{m,H}^H \)). The bounds on population prevalence under both test monotonicity and hospitalization monotonicity are

\[
U_{m,h} = \min \{ U_m, U_{m,H}^H \} = \min \{ \Pr(C_{ht} = 1|D_{ht} = 1), \Pr(C_{ht} = 1|D_{ht} = 1, H_{ht} = 1) \} = \min \{ \text{Population test positivity}, \text{Hospital test positivity} \}.
\]

In practice, the upper bound on prevalence among non-COVID hospital patients is typically lower than the population upper bound. This means that the test positivity rate among non-COVID hospitalizations is an upper bound on population prevalence. Thus assumption 2 typically tightens the bounds on prevalence. However, there are far fewer hospitalized patients than people in the population, so there is a tradeoff between tighter identification and the statistical precision of the estimates of the bounds.

Hospitalization independence. For some types of patients, it may be credible to assume that the risk of hospitalization is actually unrelated to the risk of SARS-CoV-2 infection. Formally, this type of hospitalization independence assumption can be written as follows:

**Assumption 3** (Hospitalization independence). \( \Pr(C_{ht} = 1|H_{ht} = 1) = \Pr(C_{ht} = 1) \).

Assumption 3 (hospital independence) implies that people who are hospitalized for a specified non-COVID health condition have the same probability of being infected with the virus as the general population. An equivalent statement is that people with SARS-CoV-2 have the same probability of being hospitalized for a non-COVID condition as people without SARS-CoV-2.
Combining the test monotonicity and hospitalization independence assumptions can narrow the bounds on population prevalence. As before, let $U_m$ and $L_m$ represent test monotonicity bounds in the general population. Let $U^H_m$ and $L^H_m$ be test monotonicity bounds for the hospital population. Under both test monotonicity and hospital independence, the bounds on population prevalence are

$$L_{m, ind} = \max\{L_m, L^H_m\} = \max\{\Pr(C_{it} = 1|D_{it} = 1)\Pr(D_{it})\},$$

$$U_{m, ind} = \min\{U_m, U^H_m\} = \min\{\Pr(C_{it} = 1|D_{it} = 1, H_{it} = 1)\Pr(D_{it})\}.$$

It turns out that $U_{m, ind} = U_{m, h}$, so the upper bound is the same under hospitalization independence and hospitalization monotonicity. The hospital independence assumptions affect the population lower bound. In practice, the lower bound is higher among hospitalized subpopulations, which means that—under the independence assumption—the lower bound on population prevalence is the confirmed positive rate among non-COVID hospitalizations.

**D. Summary and Data Requirements**

Figure 1 summarizes our methodological results and serves as a guide for interpreting our empirical findings. The overall approach requires data on the tested population that can be linked to hospital inpatient records, which contain diagnosis information. We work with three main assumptions: two weak monotonicity assumptions, and one conditional independence assumption. The flow chart shows which assumptions yield which bounds on population prevalence. Using only data and no assumptions, we have worst-case bounds for prevalence in the general population and for hospitalized subpopulations. Under assumption 1 (test monotonicity) the bounds tighten. The lower bound becomes the confirmed positive rate and the upper bound the test positivity rate.

Assumptions 2 and 3 let us extrapolate from the hospitalized subpopulation to the general population. But these bounds also turn out to be fairly simple objects. Under assumption 2 (hospitalization monotonicity) the upper bound on population prevalence tightens to the test positivity rate in the hospitalized population. Under assumption 3 (hospitalization independence), the lower bound on population prevalence tightens to the confirmed positive rate in the hospitalized population and the upper bound is the same as it is under hospital monotonicity.

An appealing feature of these bounds is that they can be calculated without new data collection efforts. Every state already reports the number of tests and the number of positive tests, and many states report the number of COVID-related hospitalizations. To release all of the bounds we report in the paper, 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 (Arizona Department of Health Services, 2020; Illinois Department of Public Health, 2020a,b).

**III. Indiana Hospital and Testing Data**

**A. Test and Hospitalization Data Sets**

Our test data consist of all polymerase chain reaction (PCR) tests for SARS-CoV-2 conducted in Indiana between January 1, 2020 and December 18, 2020, reported to the Regenstrief Institute and available for research. This is the near-universe of PCR tests in the state. Appendix figure A.1 shows that the number of cases in our tests data matches the state's reported number almost perfectly, until November, when they diverge somewhat, as our data appear to be missing some tests. The consequence of these missing tests, in our framework, is a reduced lower bound.

Our hospitalization data consist of all admissions to hospitals belonging to the Indiana Network for Patient Care (INPC), a health information exchange that centralizes and stores data from health providers across the state of Indiana. 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). The hospital inpatient data contain separate observations for each admission. The same hospitalization can appear in the data set multiple times, for example because an insurer and a hospital both report it. To avoid duplicating these records, we keep one observation per admission time (defined second-by-second), keeping the observation with the most diagnosis codes. We link the testing and hospital data sets to each other and, for a subset of patients, to demographic information.

**B. Measuring Tests and Cases**

*In-hospital testing, positivity rate, and confirmed positives.* Because our data do not record in-hospital testing, we match tests to hospitalizations based on date. We say a hospitalized patient is tested in-hospital if she had at least one SARS-CoV-2 test dated between 5 days prior to admission

---

2See, for example, The COVID Tracking Project (2020).
3States 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).
4See Grannis et al. (2005) for more details.
5In 0.02% of cases where there is ambiguity about which record to keep; in those cases, we choose at random.
We examine the tested and hospitalized samples. These diagnoses include general symptoms such as cough or fever, as well as more specific diagnoses like acute pneumonia, viral influenza, or COVID-19. A hospitalization is ICLI-related if it has any influenza- or COVID-like (ICLI) diagnoses. Non-ICLI hospitalizations are ones that have no ICLI-related diagnoses. Non-ICLI hospitalizations are quite rare.

Population testing and positivity rates. We examine population-wide test rates and positivity rates on a week-by-week basis. We define test rates as the share of the population tested at least once in a given week, and the positive rate as the share of people with at least one positive test in a given week, among people tested that week.

C. Sample Construction

Throughout, a patient is in the “test sample” if they are tested at least once, and in the “inpatient sample” if they are hospitalized at least once. We construct three hospitalization samples, as follows.

ICLI and non-ICLI hospitalizations. We start by defining hospitalizations for influenza- and COVID-like illness (ICLI) using ICD-10 codes. We collect diagnoses codes for influenza-like illness from Armed Forces Health Surveillance Center (2015), and codes for COVID-like illness from Center for Disease Control and Prevention (2020). Appendix B lists the ICD-10 codes used to define the analytic samples. These diagnoses include general symptoms such as cough or fever, as well as more specific diagnoses like acute pneumonia, viral influenza, or COVID-19. A hospitalization is ICLI-related if it has any influenza- or COVID-like illness (ICLI) diagnoses. Non-ICLI hospitalizations are ones that have no ICLI-related diagnoses.

We start our analysis with the non-ICLI sample for two reasons. First, our hospital independence assumption is most plausible for hospitalizations that are not obviously COVID-related, and this sample meets that criterion. Second, 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.

However, the non-ICLI sample may not satisfy the hospital representativeness assumptions. First, inclusion in this sample implicitly risks conditioning on COVID itself. 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 hospitalizations. 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. Although the clear causes do not include COVID-like diagnoses (such as respiratory distress), a clear-cause admission can nonetheless also be an ICLI hospitalization. This is because if a patient is admitted for, say, cancer, but has respiratory distress as a comorbidity, we would treat it as both a cancer admission and an ICLI admission.

Summary statistics and test rates. 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 Bureau, 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, in what follows 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. About one in three Hoosiers has ever had a COVID test, whereas about half of hospitalized Hoosiers have had a test.

Although hospitalized patients are about 44% more likely to have ever been tested than the general public, during the period of their actual hospitalization they are at least ten times more likely to be tested, as we show in figure 2, which plots weekly age-adjusted testing rates for each sample. The testing rate in the general population grew from less than 1% in May and June to a peak of 3% in mid-November. Test rates for ICLI hospitalizations varied between 60% and 75% in most weeks. Testing rates among non-ICLI hospital patients who we report in appendix table A.1 and age-adjusted rates in appendix table A.2.
patients and among the clear-cause non-COVID hospital patients were 25%–40% in May and later months. Despite their high rates, hospitalized patients are not always tested. Even ICLI patients are tested only about two-thirds of the time. Several factors explain incomplete testing. Highly symptomatic patients may not be tested because a test would not necessarily influence care, and could generate a false negative, and testing capacity was sometimes limited. They also might not receive a SARS-CoV-2 test if they had a positive influenza test, as that provides an alternative explanation for the symptoms. For asymptomatic patients, hospital policy encourages testing, but does not always require it. The Chief Medical Officer of a large Indiana hospital system indicated that asymptomatic patients would typically be

Figure plots the age-standardized 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 five days prior to admission and one day after. To age-standardize, we reweight the hospitalization samples to match the population age distribution.

**Table 1—Person-Level Summary Statistics**

| Sample | Full State | Ever Tested | Ever Has Diagnosis | Not ICLI | Clear cause | ICLI |
|--------|------------|-------------|--------------------|---------|-------------|------|
| Age as of 1/1 | | | | | | |
| Born after 1/1 | 0.004 | 0.118 | 0.129 | 0.143 | 0.003 | 0.005 |
| 0–17 | 0.237 | 0.142 | 0.036 | 0.032 | 0.030 | 0.029 |
| 18–29 | 0.166 | 0.208 | 0.121 | 0.123 | 0.133 | 0.181 |
| 30–50 | 0.250 | 0.279 | 0.184 | 0.180 | 0.183 | 0.182 |
| 50–64 | 0.197 | 0.202 | 0.204 | 0.197 | 0.189 | 0.191 |
| 65–74 | 0.087 | 0.095 | 0.162 | 0.158 | 0.150 | 0.171 |
| 75+ | 0.063 | 0.070 | 0.175 | 0.182 | 0.172 | 0.243 |
| Age unknown | 0.001 | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 |
| Gender | | | | | | |
| Male | 0.493 | 0.442 | 0.417 | 0.428 | 0.420 | 0.393 |
| Female | 0.507 | 0.558 | 0.583 | 0.572 | 0.580 | 0.607 |
| Unknown | 0.012 | 0.007 | 0.001 | 0.000 | 0.000 | 0.000 |
| Race/ethnicity | | | | | | |
| White | 0.848 | 0.877 | 0.862 | 0.834 | 0.835 | 0.853 |
| Black | 0.099 | 0.101 | 0.121 | 0.147 | 0.146 | 0.128 |
| Race unknown | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| Test variables | | | | | | |
| Ever tested | 0.343 | 1.000 | 0.489 | 0.526 | 0.498 | 0.550 |
| Confirmed positive | 0.058 | 0.169 | 0.085 | 0.092 | 0.061 | 0.069 |
| People | 6,637,426 | 2,278,910 | 539,903 | 325,410 | 291,650 | 66,887 |
| Counties | 92 | 92 | 92 | 92 | 92 | 92 |

Column 1 reports characteristics on the state population, column 2 reports characteristics for the set of people appearing in the test data, and columns 3–7 for people appearing the hospital data, ever (column 3), with at least one diagnosis (column 4), at least one non-ICLI hospitalization for ICLI (column 5), at least one clear-cause hospitalization (column 6), see text for details), or at least one ICLI hospitalization with a diagnosis and not for ICLI (column 7).
tested at admission, but this might vary across hospitals depending on their capacity to isolate patients in private or semiprivate rooms (Weaver, 2020). At another large hospital system, the Chief Medical Officer reported that testing was at times based on capacity, but patients coming into particular divisions were more likely to be tested, as were patients coming in for operations (Crabb, 2020). Our personal experience was that hospitals encouraged testing but did not strictly require it. This anecdotal evidence indicates that non-ICLI hospitalizations provide strong encouragement but not a mandate for testing among asymptomatic or mildly symptomatic patients. Greater testing of mildly symptomatic patients would be consistent with our test monotonicity assumption applied to non-ICLI hospitalizations.

D. Justification of Hospital Representativeness Assumptions

Our goal is to use the hospitalization data and hospital representativeness assumptions to tighten bounds on population COVID-19 prevalence. In this section, we briefly provide three justifications for the hospital representativeness assumptions in our data; appendix G provides more detail. First, we benchmark our bounds against prevalence estimates obtained from the Indiana COVID-19 Random Sample Study (Menachemi et al., 2020; Richard M. Fairbanks School of Public Health, 2020). Our bounds under hospital monotonicity contain the random sample estimates, and we cannot reject the hypothesis that our bounds under hospital independence do as well. Second, we show that non-COVID hospitalized patients are not significantly different from the general population in the likelihood of having been tested for SARS-CoV-2 prior to their hospitalization. Interpreting prior testing as a proxy for risk, this suggests that non-COVID hospitalized patients are at a similar risk, consistent with hospital monotonicity and independence. Third, we show below that bounds obtained for a variety of detailed hospitalization subsamples (such as labor and delivery, vehicle accidents, appendicitis, or cancer) are fairly similar. Under the assumption that some of these conditions (e.g., appendicitis) occur independently of COVID risk, this result is consistent with hospitalization monotonicity and independence. It also suggests robustness to the exact conditions used to define non-COVID hospitalization.

IV. Bounds on COVID-19 Prevalence

A. Bounds by Broad Samples

The high test rates among the hospitalized populations shown in figure 2 imply relatively tight bounds on hospital prevalence under test monotonicity, and on population prevalence under our representativeness assumptions. We plot these population bounds in figure 3. The top panel plots population bounds using ICLI-hospitalizations under test monotonicity. The remaining panels plot population bounds for non-ICLI hospitalizations and clear-cause hospitalizations, under each representativeness assumption. We also plot 95% confidence intervals for the bounds in dashed lines.

Several patterns are clear in the figure. First, under test monotonicity, the ICLI hospitalized population has higher upper and lower bounds on prevalence than the other groups. For the ICLI patients, the prevalence bounds begin at 6%–18% in the first week of our sample, they increase to 33%–39% in the last week of March, they decline steadily to roughly 12%–18% in the summer and 8%–15% in early fall, and they increase dramatically in November. Although high, these bounds rule out the possibility that even a majority of symptomatic patients are infected with SARS-CoV-2 almost every week. The bounds for ICLI-hospitalized patients separate from the population-wide bounds, implying rejection of representativeness conditions for the ICLI sub-population.

Second, the hospitalized populations have tight prevalence bounds; in fact, the bounds for the COVID-unrelated hospitalizations sample are always contained within the population-based bounds. Under (conditional) test monotonicity and hospitalization monotonicity, the width of the bounds falls (relative to just test monotonicity) by 60% on average. Under hospitalization independence, the width falls by 70%. Accounting for the greater statistical uncertainty of the hospital-based estimates, the width still falls by 45%–55% on average. Indeed, despite some sampling error, in all but three weeks, the 95% confidence interval for bounds under hospital independence are contained in the point estimate for the bounds under test monotonicity alone.

Third, the bounds for the non-ICLI hospitalization sample and for the clear-cause hospitalization sample are nearly indistinguishable, although the confidence intervals are somewhat wider for clear-cause hospitalizations. The main differences are that the upper bound for non-ICLI hospitalization is perhaps slightly higher, and the confidence intervals wider. 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-like cases. (However appendix E shows the bias is likely small in our application.) Empirically, the similarity of the clear-cause and non-ICLI bounds provides some evidence in support of using non-ICLI hospitalizations to measure general prevalence.

Fourth, meaningful trends are evident from the bounds under hospitalization independence. The upper bound for all samples shows a U-shaped pattern, with lower and upper bounds high in the spring, falling in the summer, and falling again in the fall, the pattern more pronounced for the ICLI patients. The bounds for non-ICLI hospitalizations are somewhat wider for clear-cause hospitalizations. The main differences are that the upper bound for non-ICLI hospitalization is perhaps slightly higher, and the confidence intervals wider.

9One of us had a child born in August at a hospital in our sample. The mother was encouraged to obtain a SARS-CoV-2 test prior to admission, but the father (who attended the birth) was not. The mother’s test result was not available until after admission. Happily, it was negative.

10Our data do not contain the test results from the Random Sample Study, so we compare our bounds to the published results.

11See appendix tables A.3–A.6 for exact values and age-unadjusted bounds.
The scale in the top figure differs from the others. Figure plots age-standardized bounds on prevalence under test monotonicity and hospital monotonicity or independence, for the indicated hospitalized population, in dark gray. For comparison, the figure plots (in light gray) the bound using only population-wide data. See notes to figure 2 for sample definitions. The ICLI hospitalization figure plots bounds for hospitalized patients under test monotonicity, but does not plot population bounds, because the representativeness assumptions are rejected. The dashed line depicts 95% confidence intervals for the bound. To age-standardize, we reweight the hospitalization samples to match the population age distribution.

The pattern does not necessarily indicate that prevalence follows a U-shaped trend, because the lower bound for the population as a whole remains fixed at essentially zero. However, the non-ICLI hospitalization bounds are sufficiently tight to confirm that prevalence is lower in midsummer than late fall. For example, the upper bound in the week of September 18 is 1.6%; this is lower than the lower bound in any week after October 30. Our 95% confidence intervals also allow us to reject that the bounds overlap between September 18 and any week in November. Thus under hospital independence and test monotonicity, our hospital-based bounds are tight enough to show that prevalence unambiguously rose from summer to fall.

**B. Bounds by Cause of Admission**

Our overall 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 SARS-CoV-2 infection risk. One might worry, for example, that pregnant women are especially cautious and careful not to become infected. In contrast, people who get into vehicle accidents during the epidemic might be a less cautious group either because they are not careful drivers, or because they are out of the house at all.

Since the credibility of key assumptions may vary across different clear causes, we estimated test rates and bounds separately for each of our clear causes of hospitalizations. Because each individual cause has relatively few hospitalizations, we aggregate across all time periods to form these estimates. We focus on nine sets of causes: AMI (i.e., heart attack), appendicitis, cancer, fractures, labor/delivery, non-vehicle accidents, stroke, vehicle accidents, and wounds. These nine groups have between 1,900 and 14,000 hospitalizations each. The age profile varies considerably across groups by
cause of admission (appendix table A.7). All ages are represented in the cancer sample. Appendicitis and vehicle accidents both afflict more young people. AMI, stroke, and other causes of admission are tested often and for reasons unrelated to COVID-19 risk, such as students, international travelers, and some workers. Of course, validating the representativeness assumptions (monotonicity or independence) would be important. Overall, our results show the value of reporting test rates and test results separately based on reason for testing.

Our bounds could be calculated by states or other health agencies with little additional data infrastructure. Similar bounds could be constructed for other groups that are tested often and for reasons unrelated to COVID-19 risk, such as students, international travelers, and some workers.

V. Conclusion

In fast-moving pandemics, testing is often limited and rationed to the most symptomatic, making it difficult to infer population prevalence. We propose examining populations that are heavily tested for reasons plausibly unrelated to their underlying disease prevalence, focusing on patients hospitalized during the COVID-19 pandemic for reasons unrelated to COVID-19. A test monotonicity assumption yields tight bounds on prevalence for this population. To extrapolate from this population to the general population, we introduce and validate a relatively weak hospitalization monotonicity assumption, and a stronger hospitalization independence assumption. Under either assumption, the hospitalized population yields useful bounds on population prevalence, tighter than those obtained with population-wide testing data.

Table 2.—Bounds on Prevalence by Cause of Admission, Pooling All Time Periods

| Cause of admission | # of Admissions | Test rate | Bound (H-M) | Bound (H-I) |
|--------------------|-----------------|-----------|-------------|-------------|
| AMI                | 8,624           | 0.382     | [0.002, 0.085] | [0.033, 0.085] |
| Appendicitis       | 1,961           | 0.384     | [0.002, 0.045] | [0.017, 0.045] |
| Cancer             | 9,585           | 0.337     | [0.002, 0.059] | [0.012, 0.059] |
| Fracture           | 13,718          | 0.363     | [0.002, 0.041] | [0.015, 0.041] |
| Labor/Delivery     | 13,304          | 0.197     | [0.002, 0.048] | [0.007, 0.048] |
| Other Accident     | 9,782           | 0.313     | [0.002, 0.088] | [0.027, 0.088] |
| Stroke             | 8,297           | 0.250     | [0.002, 0.066] | [0.017, 0.066] |
| Vehicle Accident   | 1,944           | 0.271     | [0.002, 0.077] | [0.014, 0.077] |
| Wound              | 3,642           | 0.310     | [0.002, 0.077] | [0.024, 0.077] |

Table reports the number of admissions, in-hospital test rate, and bounds on COVID prevalence by cause of admission. H-M bounds are valid under hospital monotonicity, and H-I are valid under hospital independence (both require test monotonicity). The H-I and H-M upper bound is identical, and the lower bound for H-M is the same across causes because it is determined by the population confirmed positive rate. The sample consists of all admissions with the indicated cause between March 13 and December 18, 2020. See appendix B for a precise definition of each cause. We report the point estimate for the bounds in brackets, and 95% confidence interval in parentheses.

REFERENCES

Allcott, H., L. Boxell, J. Conway, M. Gentzkow, M. Thaler, and D. Y. Yang, “Polarization and Public Health: Partisan Differences in Social Distancing During the Coronavirus Pandemic,” NBER working paper 26946 (2020).

Arizona Department of Health Services, “Data Dashboard” (2020), 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, “Influenza-Like Illness” (2015), aFHSC Standard Case Definitions, https://health.mil/Reference-Center/Publications/2015/10/01/Influenza-Like-Illness, last accessed June 28, 2020.

Brookmeyer, R., and M. H. Gail, “A Method for Obtaining Short-Term Projections and Lower Bounds on the Size of the Aids Epidemic,” Journal of the American Statistical Association 83 (1988), 301–308. 10.1080/01621459.1988.10479599

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

Center for Disease Control and Prevention, “ICD-10-CM Official Coding and Reporting Guidelines April 1, 2020 Through September 30, 2020” (2020), https://www.cdc.gov/nchs/data/icd/COVID-19-guidelines-final.pdf, last accessed June 28, 2020.

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

Crabb, D., Email Correspondence with Nir Menachemi (December 22, 2020).

Egan, J. R., and I. M. Hall, “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 (2015), 20150096. 10.1098/rsif.2015.0096

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., “Estimating the Effects of Non-Pharmaceutical Interventions on Covid-19 in Europe,” Nature 584 (2020), 257–261. 10.1038/s41586-020-2405-7.

Garniss, S. J., P. G. Biondich, B. W. Mamil, M. G. Wilson, L. Jones, and J. M. Overhage, “How Disease Surveillance Systems can Serve as Practical Building Blocks for a Health Information Infrastructure: The Indiana Experience” (pp. 286–290), in AMIA Annual Symposium Proceedings, Vol. 2005 (Bethesda, MD: American Medical Informatics Association, 2005).

12These groups are not necessarily mutually exclusive, and in particular there is overlap between injury and accidents.
