Estimating the cumulative rate of SARS-CoV-2 infection

By: Christopher R. Bollinger and Martijn van Hasselt

Bollinger, C. R. & Van Hasselt, M. (2020). Estimating the Cumulative Rate of SARS-CoV-2 Infection. *Economics Letters, 197*, 109652. DOI: 10.1016/j.econlet.2020.109652.

Abstract:

Accurate estimates of the cumulative incidence of SARS-CoV-2 infection remain elusive. Among the reasons for this are that tests for the virus are not randomly administered, and that the most commonly used tests can yield a substantial fraction of false negatives. In this article, we propose a simple and easy-to-use Bayesian model to estimate the infection rate, which is only partially identified. The model is based on the mapping from the fraction of positive test results to the cumulative infection rate, which depends on two unknown quantities: the probability of a false negative test result and a measure of testing bias towards the infected population. Accumulating evidence about SARS-CoV-2 can be incorporated into the model, which will lead to more precise inference about the infection rate.

Keywords: Bayesian inference | partial identification | measurement error | non-random sampling

Article:

1. Introduction

Since its emergence in late 2019, the spread of the new severe acute respiratory coronavirus 2 (SARS-CoV-2) and COVID-19, the illness caused by the virus, has become a global pandemic (Wiersinga et al., 2020). While quantitative data about the spread of SARS-CoV-2 continues to accumulate, concerns remain about data reliability and measurement error. For example, estimating the fatality rate of COVID-19 is complicated by the fact that not all cases of COVID-19 are reported and included in public health statistics, and the decision whether or not to attribute a death to COVID-19 can be difficult (Basu, 2020).

In this article, we consider the problem of estimating the cumulative incidence of infection (for simplicity referred to as the infection rate), which still remains largely unknown (Brown and Walensky, 2020). From a public health perspective, estimating and monitoring infection rates is critical both for predicting the demands on the healthcare system and for understanding the proportion of the population that remains vulnerable to infection. One way to learn about the infection rate is to estimate the presence of antibodies to the virus in large, population-based

samples (e.g., Havers et al., 2020, Stringhini et al., 2020). A difficulty with this approach is that infection rate estimates are difficult to update in real time. Also, antibody levels tend to decrease over time; an undetectable level therefore does not guarantee that an individual was not infected with SARS-CoV-2 at some point in the past (Patel et al., 2020).

An alternative approach, and the focus of this paper, is to estimate infection rates from cumulative testing data. These data are frequently updated, which enables real-time monitoring of the spread of infection. Testing data, however, present two major challenges. First, SARS-CoV-2 tests are not randomly assigned to individuals. In practice, receiving a test is more likely for individuals who show symptoms of COVID-19 and individuals who face a higher risk of exposure to the virus (e.g., healthcare workers). The infection rate among tested individuals is then likely to overestimate the infection rate in the general population. Second, emerging evidence suggests that the sensitivity of the reverse transcription-polymerase chain reaction (RT-PCR) test, the most common test for an active SARS-CoV-2 infection, can be as low as 70% (e.g., West et al., 2020). Put differently, the probability of a false negative—the event where an infected individual has a negative test result—can be as high as 30%. This implies that it is also possible that the actual infection rate is significantly higher than the observed fraction of positive test results.

Manski and Molinari (2020) show that the lack of random testing and the inaccuracy of tests imply that the infection rate is only partially identified. They derive non-trivial parameter bounds, but, as evident from their Table 2, estimates of the bounds can be very far apart. Stoye (2020) notes the width of the bounds, and, similar in spirit to the approach taken here, uses prior bounds on test accuracy and the selection of test subjects to tighten the bounds on the prevalence. Another approach, by Sacks et al. (2020), uses testing from non-Covid hospital patients to provide bounds. Both papers, as well as Manski (2020) and this paper, focus on the *period prevalence* or cumulative incidence: the total fraction of the population who have been infected over a certain time period. In contrast, Peracchi and Terlizzese (2020) focus on bounding the *point prevalence*: the percentage of people who are infected at a given point in time.

In this paper, we make two contributions. First, we present a simple way, similar to Stoye (2020), to parameterize the relation between the (observed) rate of positive test results and the (unobserved) population infection rate. This relation depends on two key parameters: a measure of the randomness in testing and the false negative rate for the average SARS-CoV-2 test. Second, we use a simple Bayesian model to estimate the infection rate. The prior distribution accounts for uncertainty about the two key parameters, which in turn affects posterior uncertainty about the infection rate. In the (non-Bayesian) partial identification approach of Stoye (2020), uncertainty is expressed through deterministic bounds on the key parameters. The use of a prior distribution, on the other hand, provides additional flexibility, since the prior can reflect deterministic bounds as well as probabilistic beliefs. As such, the Bayesian model can be seen as intermediate between a highly parameterized, point-identified model, and a fully nonparametric but much less informative bounding approach.

The Bayesian approach also highlights the practical value of our parameterization: we discuss how accumulating evidence about the sensitivity of SARS-CoV-2 testing, the distribution of symptoms in the infected population, and the eligibility criteria for testing can be used to inform

the prior and can lead to more precise estimates of the infection rate. In Section 2 below, we discuss the Bayesian model. Section 3 presents results and Section 4 concludes.

2. The model

Following Manski and Molinari (2020) and Stoye (2020), let C_t, T_t and R_t be three binary indicators for being infected with SARS-CoV-2 by time period t, for being tested by time t, and for receiving a positive test result by time t, respectively. Case count data (for example, from www.covidtracking.com) informs us about the fraction of positive test results $\mu_t = P(R_t = 1|T_t = 1)$ We are interested, however, in the cumulative incidence rate (or simply "infection rate") $\pi_t = P(C_t = 1)$, which is not identified (Manski and Molinari, 2020, Manski, 2020).

Before we consider the mapping between μ_t and π_t, we make two, arguably plausible assumptions. The first is that

$$0 < P(T_t = 1|C_t = 0) \leq P(T_t = 1|C_t = 1) \tag{1}$$

This assumption states that there is a non-zero probability that a non-infected individual will get tested, and that being tested is more likely when someone is infected with SARS-CoV-2. The latter is plausible, because infection increases the likelihood of having symptoms of COVID-19, which in turn makes it more likely that you become eligible or are encouraged to get tested. The second assumption is about the properties of the test:

$$P(R_t = 0|C_t = 1, T_t = 1) > 0, \qquad P(R_t = 1|C_t = 0, T_t = 1) \tag{2}$$

Thus, the probability of a false positive is zero but the probability of a false negative is strictly positive (and, for simplicity, fixed over time). Given what is currently known about the RT-PCR test—see Section 1—this assumption is also plausible.

Using assumptions (1), (2) and Bayes' rule, it can be shown that

$$\pi_t = \frac{\mu_t \gamma_t}{\mu_t \gamma_t + (1 - \mu_t - q)}. \tag{3}$$

where $\gamma_t = P(T_t = 1|C_t = 0)/P(T_t = 1|C_t = 1)$ is the relative likelihood of getting tested without and with the infection, respectively, and $q = P(R_t = 0|C_t = 1, T_t = 1)$ is the probability of a false negative test result. Note that (1) implies that $0 < \gamma_t \leq 1$. The parameter γ_t has a useful interpretation as a measure of the randomness in testing: if testing is done randomly, then $\gamma_t = 1$ and (3) shows that the prevalence of SARS-CoV-2 infection πt is higher than the test positive rate μ_t (due to the presence of false negatives). On the other hand, $\gamma_t < 1$ indicates that testing is not random. The closer γ_t is to zero, the more testing is geared towards the infected population only. Depending on the value of q, the value of π_t could be higher or lower than μ_t. We note that there are other ways to parameterize testing selectivity. Peracchi and Terlizzese (2020) use the relative likelihood of being infected, whereas (Stoye, 2020) uses the odds ratio of getting tested. Our use of γ_t, however, is convenient for incorporating prior information, as we will discuss in Section 3.

We start by assigning (μ_t, γ_t, q) a prior distribution. Since μ_t is identified but (γ_t, q) is not, it follows from the results in Poirier (1998) and Moon and Schorfheide (2012) that the joint posterior is the product of the posterior of μ_t and the conditional prior of (γ_t, q), given μ_t. Thus, even as the sample size grows, the prior will still exert significant influence over the posterior. The prior itself is additional information brought to the problem. The approach is appealing because it properly accounts for uncertainty about the unknowns and, as more evidence about testing and test reliability becomes available, allows us to adjust the prior distribution to reflect the current state of knowledge. While in some cases the exact mathematical form of the posterior of πt may be derived from the posterior of (μ_t, γ_t, q), in practice it is easier to generate a set of random draws from it. Our data consists of (n_{t1}, n_{t0}), the cumulative numbers of positive and negative test results at time t, respectively. Throughout, we use a uniform prior for μ_t, which is the natural conjugate prior for the binomial likelihood. We now repeat the following three simple steps: (1) generate a random draw μ_t from its beta posterior distribution with parameters $(n_{t1} + 1, n_{t0} + 1)$; (2) generate a random draw (γ_t, q) from the prior; and (3) use the values (μ_t, γ_t, q) and Eq. (3) to calculate πt.

3. Results

To illustrate our approach, we use testing data for North Carolina from www.covidtracking.com As of September 14, 2020, a total of 185,781 individuals tested positive and 2,449,038 tested negative for SARS-CoV-2 infection. As a benchmark, we first consider the case $q = 0$ and $\gamma_t = 1$, so that $\pi_t = \mu_t$. Summary statistics for 5,000 draws from the posterior are shown in the first row of Table 1. Clearly, the posterior is highly concentrated around the mean of 7.05%.

Table 1. Posterior summary statistics for πt for different priors of (γ_t, q).

q	γ_t	Mean	Median	Standard deviation	95% HPD interval
0	1.00	0.0705	0.0705	0.0002	[0.0702,0.0708]
TN(0.20,0.07)	1.00	0.0890	0.0882	0.0079	[0.0744,0.1043]
TN(0.20,0.07)	0.75	0.0683	0.0677	0.0062	[0.0568,0.0803]
TN(0.20,0.07)	0.50	0.0466	0.0461	0.0043	[0.0386,0.0550]
TN(0.20,0.07)	0.25	0.0238	0.0236	0.0023	[0.0197,0.0283]
TN(0.20,0.07)	unif. mixture	0.0413	0.0408	0.0243	[0.0000,0.0816]
TN(0.20,0.07)	beta (4,3)	0.0524	0.0528	0.0162	[0.0209,0.0827]

Next, we allow for false negatives. Recent estimates of the false negative rate range from around 10% to as high as 40% (e.g. Woloshin et al., 2020, Arevalo-Rodriguez et al., 2020). Based on this, we assign a normal prior to q, with a mean of 20%, a standard deviation of 7%, truncated to the interval $[0, 1 - \mu_t]$ (this ensures that π_t does not exceed 1). Comparing rows 1 and 2 in Table 1, we see that assigning a prior to q while fixing $\gamma_t = 1$, shifts the posterior of π_t to the right (as predicted by (3)) and increases its dispersion. The posterior mean increases by more than 26% to 8.90%, whereas the posterior standard deviation is nearly 40 times larger. As the value of γ_t decreases, rows 3–5 of Table 1 show that the posterior mean, median and standard deviation of π_t all decrease. For example, when getting a test is four times less likely in the non-infected population compared to the infected population ($\gamma_t = 0.25$), the posterior mean for the infection rate is only 2.38% and much lower than the posterior mean of μ_t. Thus, even with a

potentially large rate of false negatives, selectivity in testing yields estimates of the infection rate that are relatively low.

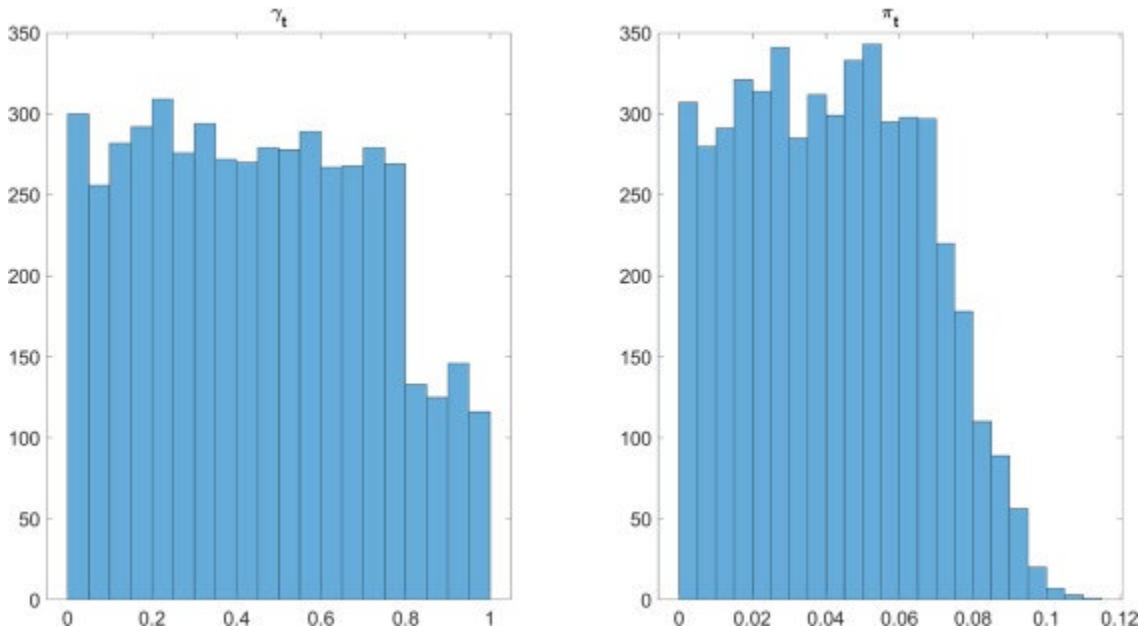

Fig. 1. Uniform mixture prior for γ_t (left) and posterior of π_t (right).

Finally, instead of fixing its value, we now assign a prior distribution to $\gamma_t = P(T_t = 1|C_t = 0)/P(T_t = 1|C_t = 1)$. Recent evidence suggests that as many as 30% of individuals infected with SARS-CoV-2 may be asymptomatic (Nishiura et al., 2020). Assuming that this group is very unlikely to receive a test, $P(T_t = 1|C_t = 1)$ may be around 0.7. CDC data through May 2020 shows that around 14% of infected individuals end up being hospitalized (Stokes et al., 2020). This suggests that the infected population may be roughly partitioned into an asymptomatic group (30%), a group with severe symptoms that require hospital care (14%), and a residual group with "moderate" symptoms (56%). Regarding the numerator of γ_t, suppose that in the non-infected population only those with moderate symptoms receive a test, and that severe symptoms requiring hospitalization are extremely unlikely. If the occurrence of at least moderate symptoms is less likely in the non-infected population than the infected population, then γ_t is bounded from above by $0.56/(0.14 + 0.56) = 0.80$.

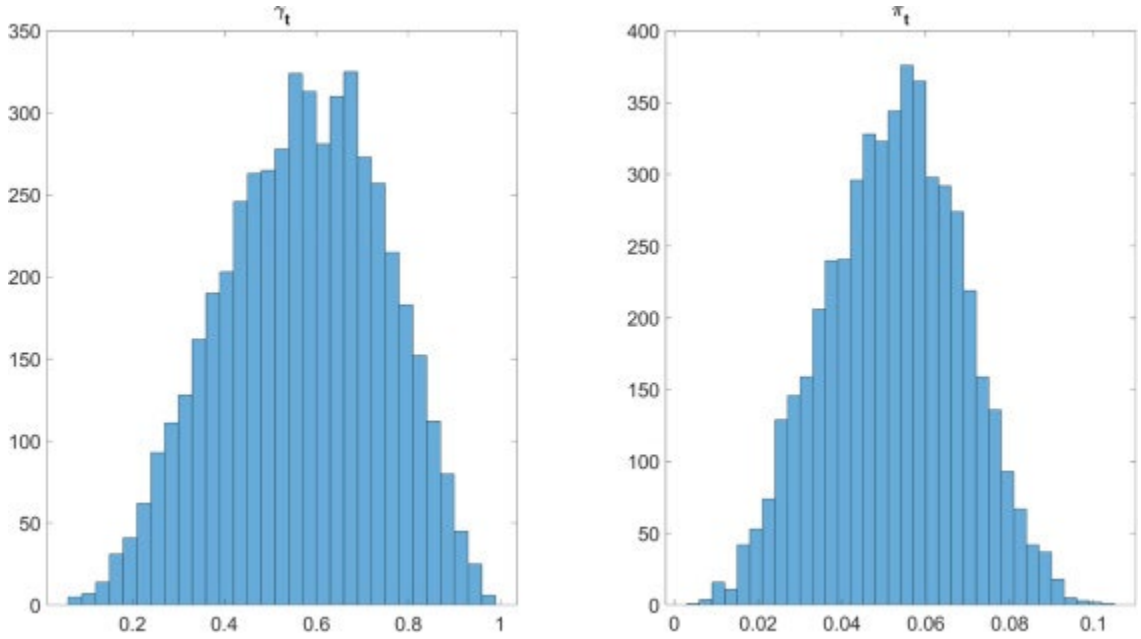

Fig. 2. Beta prior for γ_t (left) and posterior of π_t (right).

A prior that reflects this reasoning would place most of its probability mass on values below 0.80. We consider two such priors. The first is a uniform mixture, where $0 \leq \gamma_t \leq 0.8$ with probability 90%, and $0.8 < \gamma_t \leq 1$ with probability 10%. Our second prior is a beta distribution with parameters (4,3). This distribution has a mean around 0.6, is slightly left-skewed and assigns a probability of around 90% to values below 0.8. The posteriors of π_t resulting from these priors are shown in the last two rows of Table 1, and in Fig. 1, Fig. 2.

Allowing for uncertainty about γ_t leads to a substantial increase uncertainty about π_t, as evident from the standard deviations and 95% HPD intervals. For example, moving from $\gamma_t = 0.75$ to a uniform mixture prior, the posterior standard deviation of πt increases about fourfold from 0.0062 to 0.0243. The beta prior for γ_t is more informative than the uniform mixture; this leads to a slightly more concentrated posterior of π_t, with the standard deviation decreasing from 0.0243 to 0.0162. The 95% HPD interval becomes narrower, which is mainly due to an increase in the lower bound. Finally, Fig. 1, Fig. 2 show that the shapes of the prior of γ_t and the posterior of π_t are similar. This results from the mathematical relationship in (3).

4. Conclusion

In the current phase of the SARS-CoV-2 pandemic, tracking infection rates over time remains critically important. Much of the publicly available information in the U.S., and indeed in countries all over the world, takes the form of cumulative testing data. Estimating the infection rate by the fraction of positive test results is problematic because individuals are not tested randomly and the tests themselves may yield false negative results.

In this paper, we present a mapping from the test positive rate to the infection rate that is parameterized in terms of a measure of testing selectivity and the false negative rate. This mapping subsequently forms the basis of a Bayesian model to conduct inference about the

(unidentified) rate of infection. As an illustration, we use recent data from North Carolina. Our method, however, can be applied to any entity (e.g., counties, states, countries) for which case count data is available.

A critical ingredient in the Bayesian model is the prior probability distribution for the testing selectivity parameter and the false negative rate. We motivated our choice of prior by referring to external studies with relevant direct and indirect information. Of course, we do not claim that our prior is "correct". Indeed, other sources of prior information could be brought to bare in this analysis. For example, information on cold and flu prevalence and associated symptoms may provide evidence on symptom rates. As new tests are adopted, different information on misclassification rates will become available (Toulis, 2020). Or, as different populations are tested, new information on selection rates could be gained (Sacks et al., 2020). Over time, as more relevant information about the unknown parameters becomes available, prior distributions can be further adjusted, the scope for disagreement about priors will narrow and the accuracy of inference will improve.

Acknowledgements

We would like to thank Jeremy Bray for helpful discussions during the course of this research, and an anonymous referee for constructive comments that helped improve the paper.

References

Arevalo-Rodriguez, I., Buitrago-Garcia, D., Simancas-Racines, D., et al., 2020. False-negative results of initial rt-pcr assays for covid-19: A systematic review. http://dx.doi.org/10.1101/2020.04.16.20066787, medRxiv.

Basu, A., 2020. Estimating the infection fatality rate among symptomatic covid-19 cases in the United States. Health Aff. 39, 1229–1236. http://dx.doi.org/10.1377/hlthaff.2020.00455.

Brown, T.S., Walensky, R.P., 2020. Serosurveillance and the covid-19 epidemic in the US: Undetected, uncertain, and out of control. JAMA http://dx.doi.org/10.1001/jama.2020.14017.

Havers, F.P., Reed, C., Lim, T., et al., 2020. Seroprevalence of antibodies to sars-cov-2 in 10 sites in the United States, March 23-May 12, 2020. JAMA Int. Med. http://dx.doi.org/10.1001/jamainternmed.2020.4130.

Manski, C.F., 2020. Bounding the predictive values of COVID-19 antibody tests. Working Paper 27226. National Bureau of Economic Research.

Manski, C.F., Molinari, F., 2020. Estimating the covid-19 infection rate: Anatomy of an inference problem. J. Econometrics http://dx.doi.org/10.1016/j.jeconom.2020.04.041.

Moon, H.R., Schorfheide, F., 2012. Bayesian and frequentist inference in partially identified models. Econometrica 80, 755–782.

Nishiura, H., Kobayashi, T., Miyama, T., et al., 2020. Estimation of the asymptomatic ratio of novel coronavirus infections (covid-19). Int. J. Infect. Dis. 94, 154–155. http://dx.doi.org/10.1016/j.ijid.2020.03.020.

Patel, M.M., Thornburg, N.J., Stubblefield, W.B., et al., 2020. Change in antibodies to sars-cov-2 over 60 days among health care personnel in Nashville, Tennessee. JAMA http://dx.doi.org/10.1001/jama.2020.18796.

Peracchi, F., Terlizzese, D., 2020. Estimating the Prevalence of the Covid-19 Infection, with an Application to Italy. Covid Economics, CEPR, pp. 19–41.

Poirier, D., 1998. Revising beliefs in nonidentified models. Econom. Theory 14, 483–509.

Sacks, D.W., Menachemi, N., Embi, P., Wing, C., 2020. What can we learn about SARS-CoV-2 prevalence from testing and hospital data? arXiv:2008.00298, [econ, stat], version: 1.

Stokes, E.K., Zambrano, L.D., Anderson, K.N., et al., 2020. Coronavirus disease 2019 case surveillance - United States, January 22-May 30, 2020. MMWR. Morb. Mortal. Wkly. Rep. 69, 759–765. http://dx.doi.org/10.15585/mmwr.mm6924e2.

Stoye, J., 2020. Bounding disease prevalence by bounding selectivity and accuracy of tests: The case of COVID-19. arXiv:2008.06178, [econ, stat].

Stringhini, S., Wisniak, A., Piumatti, G., et al., 2020. Seroprevalence of anti-sars-cov-2 igg antibodies in Geneva, Switzerland (serocov-pop): a population-based study. Lancet http://dx.doi.org/10.1016/S0140-6736(20)31304-0, (London, England).

Toulis, P., 2020. Estimation of Covid-19 prevalence from serology tests: A partial identification approach. arXiv:2006.16214, [econ, stat].

West, C.P., Montori, V.M., Sampathkumar, P., 2020. Covid-19 testing: The threat of false-negative results. Mayo Clin. Proc. 95, 1127–1129.

Wiersinga, W.J., Rhodes, A., Cheng, A.C., et al., 2020. Pathophysiology, transmission, diagnosis, and treatment of coronavirus disease 2019 (covid-19): A review. JAMA http://dx.doi.org/10.1001/jama.2020.12839.

Woloshin, S., Patel, N., Kesselheim, A.S., 2020. False negative tests for sars-cov-2 infection - challenges and implications. New Engl. J. Med. http://dx.doi.org/10.1056/NEJMp2015897.