# Empirical Strategies to Counter Non-Ignorable Non-Response When Estimating Coronavirus Prevalence *

**Michael A. Bailey** *Georgetown University*

Understanding the prevalence of coronavirus infections is important for understanding and responding to the trajectory of the outbreak. Unfortunately, the tendency to test the sickest people and the variation in testing rates across geographic areas makes it difficult to credibly estimate prevalence. While large-scale randomized testing is ideal, it is very expensive and imperfect compliance can make it vulnerable to non-response bias. This paper explores how to estimate prevalence using first-stage instruments that affect the probability of being tested but not the outcome of the test. First-stage instruments are indispensable when evaluating randomized testing with less than perfect compliance. They also can improve inference based on location-based testing.

*Keywords*: non-ignorable non-response, coronavirus prevalence

## Introduction

Understanding the prevalence of coronavirus in specific areas can inform policy decisions about stay-at-home orders and can help us predict future demands on the medical system. However, rates of positive tests are unlikely to be directly informative. Not only do testing rates vary substantially across time and regions, but the tests are almost always given to unrepresentative subsets of the population.

Dealing with the non-response in testing is, therefore, a central task when estimating prevalence from test results. Some efforts to estimate prevalence have assumed that differences between those tested and not tested can be completely explained by demographic factors such as age and ethnicity (Bendavid et al., 2020). If this assumption is true, non-response is "ignorable" and weighting or similar measures will adjust for non-response in expectation.

Unfortunately, it is highly likely that non-response in coronavirus testing is "non-ignorable," meaning that sick people are more likely to be tested, even after controlling for demographic characteristics.

The ideal way to overcome non-ignorable non-response problems is to implement randomized testing. However, randomized testing is expensive and compliance is likely to be imperfect. One sign of the difficulty of randomized sampling is the paucity of such efforts despite the unprecedented global focus on the pandemic (Mostashari and Emanuel, 2020).

This paper discusses a complementary and potentially alternative approach to full-scale randomized testing. The approach is based on using first-stage instruments (FSI) which are variables that affect the probability someone is tested but do not directly affect or predict whether or not that person is sick. Under a broad range of conditions, these instruments are required for statistically identifying prevalence (Miao, Ding and Geng, 2016; Wang, Shao and Kim, 2014; Sun et al., 2017; Marden et al., 2018).

With the proper instrument, the FSI approach enables the estimation of population prevalence

*DRAFT. Please do not cite without permission. Current version: May 22, 2020. Comments welcome to, and code available from Michael.Bailey@georgetown.edu. I appreciate helpful comments from John Kraemer, Lan Liu, Wang Miao and Michael Stoto. Errors are my own. Written in RMarkdown building on code provided by Steven V. Miller (http://github.com/svmiller).

even when the sample is decidedly unrepresentative. This paper focuses on two applications of the FSI approach.

First, the FSI approach is indispensable when analyzing full-scale randomization with imperfect compliance. Suppose, for example, a city identifies a random sample of 1,000 people to test but that only 500 actually submit to tests. Authorities should, quite reasonably be nervous that this sample could be either a healthier or less healthy subset of the full random sample, rendering the results subject to bias, the efforts to randomize notwithstanding. If authorities implemented a FSI strategy, however, they could estimate prevalence even with this potentially unrepresentative sample.

Second, the FSI approach can also enable prevalence estimation based on location-based testing. In particular, one could imagine local public health workers testing people at a given medical facility or a blood drive (Janes, 2020) or a retail location in an effort to ascertain trends in the community. The selection issues are more challenging than with random sampling with imperfect compliance, but given the right protocols community-wide prevalence can be identified under plausible conditions. Such location-based testing would likely cost much less than full-scale randomization and would be possible to conduct on a rolling basis over time.

This paper proceeds as follows. Section 1 discusses the challenge of estimating prevalence in terms of a non-ignorable missing-data problem. Section 2 discusses the merits and limits of the more widely recognized approaches to dealing with non-ignorable non-response, randomized testing and bounds. Section 3 describes how first-stage instruments enable the estimation of prevalence even when testing propensity is related to health status with a focus on the case of randomized testing with imperfect compliance. Section 4 discusses how the FSI logic could inform location based testing. Section 5 discusses the FSI approach in light of more general functional forms.

## Section 1: Epidemiological Testing as a Non-Ignorable Non-Response Problem

We begin with a standard two-stage model for testing. The propensity to be tested is

$$R_i^* = \gamma_0 + \gamma_1 X_i + \tau_i$$

where $\gamma_1$ is $1xk$ parameter vector, $X_i$ is a $kx1$ vector of covariates and $\tau_i$ is a mean-zero random variable. We observe $i$'s test results if $R_i^* > 0$.

The outcome of interest, $Y_i$, is whether person i has the coronavirus. $Y_i = 1$ if $Y_i^* > 0$ where

$$Y_i^* = \beta_0 + \beta_1 X_i + \epsilon_i$$

The correlation of $\epsilon_i$ and $\tau_i$ is $\rho$. Prevalence is a function of the $\beta$ parameters. For example, if $\epsilon$ is normally distributed, estimated prevalence would be the average of $\Phi(\hat{\beta}_0 + \hat{\beta}_1 X_i)$ across the population values of $X_i$ where $\Phi()$ is the CDF of a normal distribution.

We observe $Y_i|_{R_i=1}$, the test results for those who got tested.

*Ignorable non-response*

There are two sources of non-response. First, it is possible that non-response can be fully explained by measured demographics such as age, gender and education. This non-response is referred to as "ignorable" non-response because conditional on the correct covariates, this non-response can be ignored without causing bias. Relatively simple models fully account for this kind of non-response, including the weighting models commonly used in survey research.

A high profile study of COVID-19 antibody seroprevalence in Santa Clara County, California relied on weights to account for non-response (Bendavid et al., 2020). In this study, researchers

recruited participants with targeted Facebook ads. Response was voluntary and there was no randomization at any stage of the recruitment.

The sample of 3,330 people who were tested was demographically unrepresentative of the county: 63% of the sample was female (compared to 50% of the in population in the county); 8% of the sample was Hispanic (compared to 26% in of the county population) and 19% of the sample was Asian (compared to 28% in the county) (Bendavid et al., 2020, 5).

The researchers used weighting to adjust for the unrepresentative sample, a decision that had considerable effect on their conclusion. In the unweighted data, prevalence was 1.5 percent. After weighting, the estimated prevalence almost doubled, to 2.81 percent.

For weighting to be valid, the non-response needs to be ignorable. That is, conditional on covariates, the distribution of disease in the sample is the same as in the population. This is a general statistical concept that manifests itself in a simple conceptual test. For any given group (or combination of observed covariates), are the people in the sample a random sample of the population? That is, are the Hispanic respondents in the Santa Clara County data a random sample of all Hispanics in the county? If this is true for all groups, then weighting will produce an unbiased estimate of county prevalence.

However, as pointed out by many commentators on the article, the Santa Clara sample may have differed from the underlying population not only in terms of observable characteristics such as age, race, gender and zip code, but also in terms of unobserved characteristics such as health status. It is quite possible, for example, that those more likely to be sick were more likely to respond to the Facebook ads.

The authors acknowledge these potential biases ("Other biases, such as bias favoring individuals in good health capable of attending our testing sites, or bias favoring those with prior COVID-like illnesses seeking antibody confirmation are also possible." (Bendavid et al., 2020, 7)), but were not deterred from concluding that their "prevalence estimates of 2.49 percent to 4.16 percent are representative of the situation in Santa Clara."

*Non-ignorable non-response*

The more insidious version of non-response occurs when non-response is related to the outcome the test is trying to measure. This is referred to as "non-ignorable" non-response because we cannot simply re-weight the data to account for this type of non-response.

Papers focusing on non-ignorable non-response in epidemiology include Marden et al. (2018), Sun et al. (2017), Wang, Shao and Kim (2014) and Miao, Ding and Geng (2016). Papers focusing on non-ignorable non-response in survey research include Bailey (2019) and Peress (2010). A foundational paper in this literature is Heckman (1979). Vella (1998) reviews applications in economics.

Meng (2018) provides a unified framework for thinking about non-response, establishing among other things that bias is a function of an interaction between the extent of non-ignorability in non-response and the extent of non-response. When a sample is small relative to the target population, a small amount of non-ignorable non-response can produce more bias than a large amount of non-ignorable non-response from a sample that is a larger proportion of the overall population.

Non-response in virus testing is likely to be non-ignorable because those getting tested are more apt to be sick. Hence, as is widely recognized, the observed rates at which people test positive for the coronavirus are not indicative of the actual infection rates in the population.

Figure 1 shows an example. The prevalence is 20 percent. However, those getting tested are more likely to be sick ($\rho = 0.6$), producing a situation in which the tested population has a 60 percent chance of testing positive.

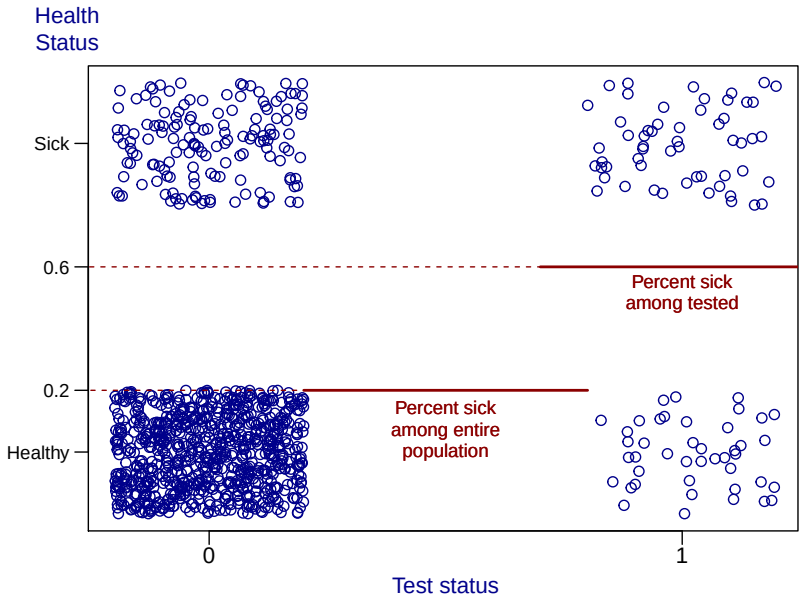


Figure 1: Test results do not reflect prevalence in population

Similar problems arise in other contexts such as HIV testing where people's willingness to test may be affected by their HIV status (Marra et al., 2017; McGovern, Canning and Barnighausen, 2018).

Two problems arise when testing non-response is non-ignorable. First, testing becomes less informative about prevalence as the relationship between the probability of getting a test and the probability of being ill increases. Figure 2 shows the relationship between prevalence and observed positive test results when 10 percent of the population is tested and the errors are assumed to follow a bivariate normal distribution with various levels of $\rho$, the correlation in errors in the two equations.

The blue line at the top of Figure 2 shows the relationship between actual and observed test results when the propensity to get tested is strongly related to the propensity to be sick ($\rho = 0.9$). Point A indicates that for this level of $\rho$, the test results will be positive 60 percent of the time when the prevalence in the general population is 8 percent.

The red line in Figure 2 shows this relationship when there is a weaker (but still strong) relationship between getting tested and being sick ($\rho = 0.6$). Point B indicates that for this level of $\rho$, the test results will be positive 60 percent of the time when the prevalence in the general population is 20 percent. When there is no relationship between getting tested and being sick ($\rho = 0$), a 60 percent positive test rate is simply associated with a 60 percent prevalence (point D).

Looking at the lines for various values of $\rho$, it is clear that the disconnect between observed test results and population prevalence is larger when $\rho$ is higher.

*Identification problems*

A second problem is more daunting for those trying to model prevalence: the model is generally unidentified, meaning that multiple combinations of parameters can explain the data equally well (Miao, Ding and Geng, 2016).

We can see the essence of the identification problem in Figure 2. The points labelled A, B, C and D all explain the observed outcome of 60 percent positive tests equally well. That is, a

prevalence of 8 percent with a $\rho$ of 0.9 (point A) explains the data as well as a prevalence of 60 percent with a $\rho$ of 0.0 (point D).[1]

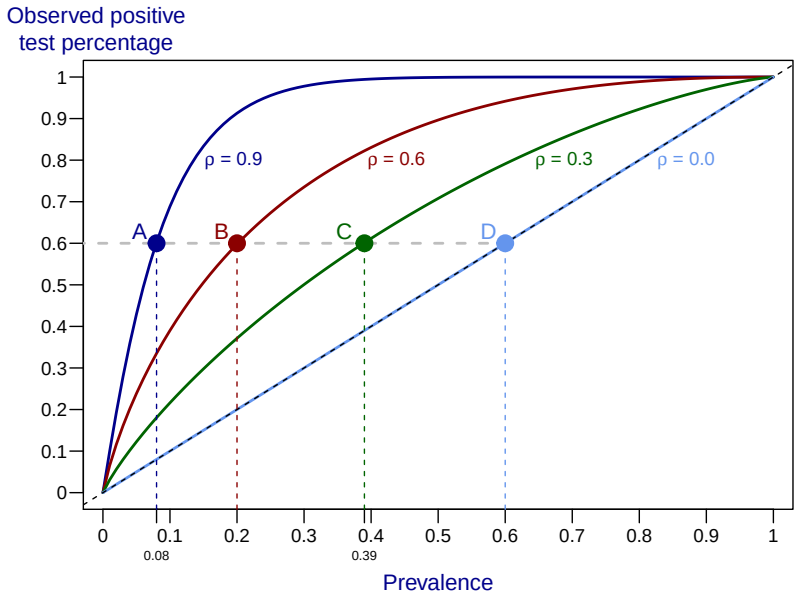


Figure 2: Relationship between general prevalence and test results for various levels of $\rho$

The identification problem is exacerbated by differences in testing rates. Figure 2 showed a situation in which the testing rate was 10 percent. However, as testing rates change, the mapping of results back to prevalence becomes more complex and depends on $\rho$. Given substantial differences in testing rates across regions and across time, this makes it impossible to treat information such as positive percent rates as either directly informative about prevalence or comparable across regions without knowing $\rho$.

Figure 3 illustrates how testing rates and identification problems interact. In the top panel, expected rates of positive tests (out of those being tested) is plotted as a function of prevalence for two testing regimes for $\rho = 0.7$, an environment in which sick people are much more likely to be tested. The blue line on the top is for a low testing regime in which only 1 percent of the population is tested. As one would expect, when tests are rare and sick people are more likely to get tested, the positive test rate is quite high. Point A indicates prevalence is 2 percent when the expected positive rate is 40 percent and 1 percent of the population for $\rho = 0.7$.

The red line in the top panel of Figure 3 plots the positive test rate for a regime in which testing is vastly expanded (to 20 percent of the population). While the positive test rate is still much higher than the prevalence, it is lower than when 1 percent were being tested as many more people are being tested. Point B in the top panel of Figure 3 indicates that prevalence is 5 percent when the expected positive rate is 20 percent and 20 percent of the population is tested.

In the scenario depicted in the top panel of Figure 3, it is actually bad news to observe the positive test rate falling from 40 percent to 20 percent when we increase the testing percent from 1 to 20 percent. That is, going from point A to point B suggests that prevalence has gone from 2 percent to 5 percent, even as the positive test rate has fallen markedly. In order to infer the prevalence is the same or falling, the increase in testing would need to produce a positive rate lower

[1]Instead of using the percent of tests that are positive as input data, some media reports report the number of positive tests as a percentage of total population. This measure is flawed as it so directly is affected by testing rates.

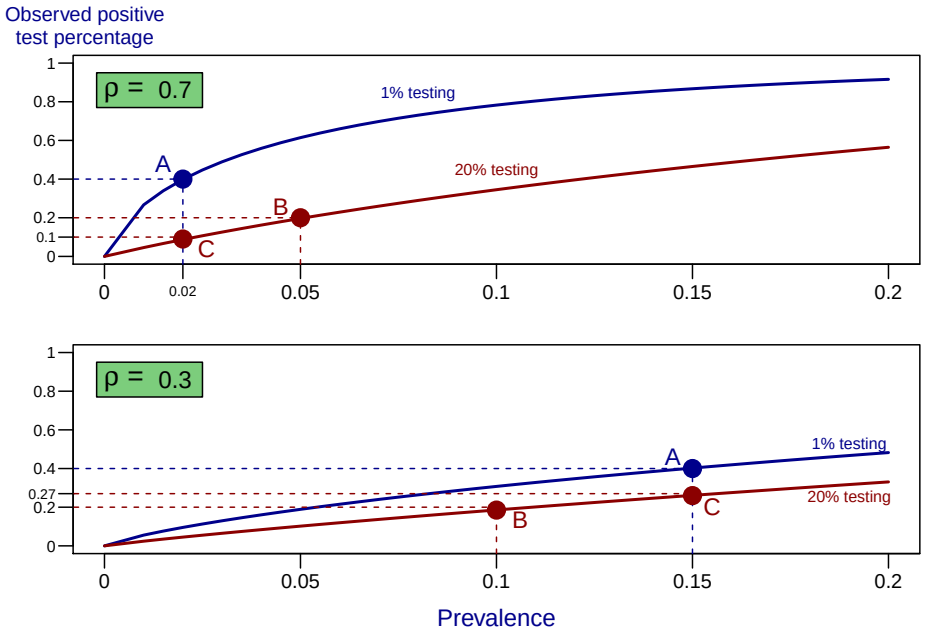


Figure 3: Relationship between general prevalence and test results for various levels of $\rho$ and testing rates

than that indicated by point C, which is about 10 percent. In other words, when the sick are much more likely to be tested, increasing testing needs to lead to a dramatic fall in positive test rates in order to suggest a fall in prevalence.

The lower panel of Figure 3 shows a similar plot but for a situation in which the relationship between testing and being sick is much lower ($\rho = 0.3$).

In this bottom panel, point A indicates that the expected positive rate is 40 percent when prevalence is 15 percent, 1 percent of the population is tested and $\rho = 0.3$. Compared to the top panel, there is a weaker relationship between being sick and testing and a given positive test rate is associated with a higher prevalence.[2] Point B indicates that the expected positive rate is 20 percent when prevalence is 10 percent and 20 percent of the population is tested.

In the scenario depicted in the bottom panel of Figure 3, a drop from 40 percent positive rate to a 20 percent positive rate is good news. That is, going from point A to point suggests prevalence has fallen from 15 percent to 10 percent. When $\rho = 0.3$ any drop below the point indicated by point C (27 percent positive rate) is consistent with a fall in prevalence.

While the scenarios underlying Figure 3 may be somewhat extreme, the underlying problem pervades any effort to translate positive test rates into prevalence estimates when non-response is non-ignorable and testing rates vary.

The problem is not necessarily insurmountable, however. In Figures 2 and 3, knowing $\rho$ would allow us to map observable test percentages back to prevalence in light of testing rates. Therefore, identifying methods that can account for $\rho$ is a crucially important step in estimating prevalence when non-response is non-ignorable.

This information could be used to create concordance tables that could allow comparison implied prevalence given positive test rates for different testing environments. Appendix 2 provides examples for case of bivariate normal errors and two different values of $\rho$.

[2] In the limit, when $\rho = 0$ the positive test rate is, in expectation, the prevalence.

## Section 2: Standard Approaches to Accounting for Non-Ignorable Non-Response

There are two widely recognized approaches to accounting for $\rho$ when there is non-ignorable non-response. The first is to structure the data so as to guarantee that $\rho = 0$ in expectation. This is done via fully randomized testing (Mostashari and Emanuel, 2020; Ioannidis, 2020). In fully randomized testing everyone in a target population has an equal probability of being tested regardless of health status and the observed proportion of people testing positive will on average equal the community prevalence. In terms of Figure 2, randomization ensures that $\rho = 0$, which in turn identifies the observed prevalence as an unbiased estimate of the actual prevalence.

Such efforts are rare, but not unheard of. Iceland, for example, recently randomly identified 6,782 Icelanders between the ages of 20 and 70 to be tested (Gudbjartsson et al., 2020). Many other regions are considering or beginning implementation of such programs.

A fully randomized testing program is difficult. In the short term for any given outbreak it is likely that test kits may be in high demand and medical professionals may be unavailable. Hopefully, this problem will become less relevant over the course of a pandemic.

Two more fundamental problems limit reliance on full-scale randomization. The first is that compliance is a challenge as not everyone chosen to be tested will in fact submit to a test. It is hard to reach everyone selected to be tested and it is quite plausible that people will vary with regard to their willingness to be tested, perhaps because of varying levels of trust of outsiders or concern about health or other commitments in their life.

In the Iceland study, for example only 33.7 percent of those randomly chosen to be tested had actually been tested as of publication of their analysis (Gudbjartsson et al., 2020, 2). The techniques discussed below can be used to address non-compliance in randomized sampling.[3]

Non-compliance is a potentially serious problem. In the Iceland study, it is possible that sickest third of randomly target people showed up to be tested. In that case, the estimate would be an obvious overestimate. Or, it is possible that the healthiest third showed up, rendering the results a clear underestimate.

As a general matter, the observed prevalence in a sample will be biased if compliance is less than 100 percent and $\rho \neq 0$. Hence, at a minimum, any random sample based testing protocol should test whether $\rho = 0$, something that requires a first-stage instrument as discussed below.

Randomized sampling is not only subject to non-ignorable non-response, it is also expensive. Prevalence surveillance requires not just a single national test for a given time period, but ongoing assessment from week to week in many locales. It may simply be too costly for many locales to carry out high quality fully-randomized studies for each time period of interest. Hence, expanding the testing toolkit to include methods that are less costly, but still useful in the presence of non-ignorable non-response is very important.

A second approach to dealing with non-ignorable non-response is to calculate bounds that do not depend on assumptions about ignorability or parametric assumptions. The generality of this approach makes it very attractive.

While bounds are very attractive theoretically, they may not always be useful practically. Manski and Molinari (2020) apply bounds analysis for high impact regions in early April 2020 and find that infection rates are bounded between 0 and 50 percent for Illinois and New York and between 0 and 64 percent for Italy.

[3] Identifying the appropriate sampling frame may be difficult as well. In the Iceland study the sampling frame seems unbalanced for reasons that are unclear as only 41 percent of those invited were male.

**Section 3: First-stage Instruments**

As a general matter non-ignorable non-response models do not work when the same variables are used to explain whether someone gets tested and the result of the test. In other words, non-ignorable non-response models generally require that we must have at least one variable that explains whether people get tested and does not explain whether they test positive.[4]

One of the insights of recent epidemiological research is that prevalence can be estimated with data short of a full-scale randomization (Sun et al., 2017; Wang, Shao and Kim, 2014; Miao, Ding and Geng, 2016). Specifically, if we have data that predicts testing propensity but not health status, then prevalence is statistically identifiable under a broad range of assumptions.

Without instruments, we have an identification problem as illustrated in Figure 2. With first-stage instruments, we are able to pin down the value of $\rho$, which in turn allows to hone in on the correct prevalence.

The appendix explains the intuition in detail for a specific parametric model, but the general idea is relatively simple. The existence of a first-stage instrument implies that there are high and low probability of being tested groups. If $\rho > 0$, the propensity to be tested is related to the outcome of the test. The low probability group in this case would have a high probability of testing positive. The high probability group will include not only the types who would have gotten tested if they had been in the low probability group, but also a group of people who would not have gotten tested in the low probability group. If $\rho > 0$ these people will have lower probability of testing positive because they have a lower propensity of being tested. Hence, the difference in the proportion who test positive across these two groups is informative about the value of $\rho$.

A metaphor may be helpful. Imagine a hospital with two doors. Two hundred people show up in order of how sick they are (with the most sick people showing up first). They line up at these two doors, with the choice of doors being completely random. The people did not know this when they lined up, but it turns out that the first 20 people at door A are tested and the first 80 people at door B are tested. If the proportion who are sick from these two groups is the same, then we have evidence that eagerness of getting tested is not related to actually being sick (which suggests $\rho = 0$). That is, if door B keeps getting the same positive rate even though they are getting a portion of the population who was less eager to be tested than the door A group, then eagerness of being tested does not seem to be related to the probability of testing positive.

On the other hand, if there is a much higher proportion of people testing positive at door A than at door B, then we have evidence that eagerness to get tested is related to testing positive. The specific difference in proportions of positive tests at the two doors will be a function of $\rho$. In general we have seen this as the percent who test positive has declined as the number of tests has increased.

The exact steps to estimating prevalence depend on model assumptions. This section focuses on a baseline case in which there is a first-stage instrument and in which the errors in the response and outcome equations are bivariate normally distributed. This approach has been used previously to estimate, among other things, HIV prevalence (Barnighausen et al., 2011) and public opinion (Bailey, 2019). In a later section I discuss more general modeling approaches.

The model is as follows.

$$R_i^* = \gamma_0 + \gamma_1 T_i + \tau_i$$

[4]Due to the non-linearity of the equations, a selection model that assumes the errors are distributed bivariate normally can be identified without a first-stage instrument. As a practical matter, models relying only on functional form perform poorly (Bailey, 2019; Puhani, 2000; Stolzenberg and Relles, 1997). Miao, Ding and Geng (2016) and Sun et al. (2017) discuss additional identification conditions even when a first-stage instrument exists.

where $\tau_i$ is a mean-zero random variable and $T_i$ is the first-stage instrument. For simplicity, we assume $\gamma_1 > 0$ and that there are no other covariates. (Adding covariates typically enhances precision.) We observe $i$'s test results if $R_i^* > 0$.

As before, the outcome of interest, $Y_i$, is whether person i has the coronavirus. $Y_i = 1$ if $Y_i^* > 0$ where

$$Y_i^* = \beta_0 + \epsilon_i$$

The key characteristic of the first-stage instrument, $T_i$, is that it has no direct effect on Y.

First-stage instruments come in two flavors. First, they may be observational. For example, it is possible that we believe that distance to a testing site affects whether someone is tested but does not directly affect whether someone has the disease. This resembles the use of distance to hospitals as an instrument to identify the effect of neonatal intensive care units, for example (Lorch et al., 2012).

In general, such observational first-stage instruments are rare and subject to doubts about the assumption that they are unrelated to the outcome of interest. For example, it could be possible that those living near a hospital have different infection rates, even after controlling for covariates.

The second flavor of first-stage instrument is based on a randomization. This could be treatment that affects the probability of testing, but does not affect probability of testing positive. For example, one could identify a random sample of people to be tested (as, for example, done in Iceland) and then also divide these people into a group who is contacted once and a group that is contacted multiple times. Being in the multiple contact group will not directly affect the likelihood of being sick but will likely increase participation, an expectation that is easily tested empirically.

This approach is extremely useful for a large-scale randomized sample approach when non-trivial non-compliance is expected. While the extent of non-compliance is an empirical question and may vary from place to place, recent experience with survey research indicates that it can be very hard to get people to respond when they are randomly chosen to participate in a study (Kennedy and Hartig, 2019; Dutwin and Lavrakas, 2016). Contact tracers have, at least in some places, been frustrated by response rates lower than 50 percent (Siegel, Abdelmalek and Bhatt, 2020).

*Estimation for 3 states*

To illustrate the estimation process, I simulate data and then use the model discussed above to estimate prevalence for three hypothetical states.

We assume for that each state has identified a random sample of 3,000 for testing. We also assume a reasonably strong relationship between the testing and outcome equations ($\rho = 0.7$).

The states vary across several dimensions. State 1 has a prevalence of 0.1 with a baseline 50 percent response rate for random testing. State 2 has a prevalence of 0.2 with a baseline 30 percent response rate for random testing. State 3 has a prevalence of 0.3 with a 30 percent response rate for random testing. The response rates are chosen to be roughly in line with Iceland's experience with randomized testing.

| State | Prevalence | Baseline response rate | Testing approach |
|---|---|---|---|
| State 1 | 10% | 50% | First-stage instrument |
| State 2 | 20% | 30% | First-stage instrument |
| State 3 | 30% | 30% | No randomization |

Importantly, states 1 and 2 implement a first-stage randomization protocol such that 20 percent of those selected initially (those for whom $T_i = 1$) are subject to more extensive outreach such that

they have 20 percentage point higher probability of being tested. Since these treatment groups are randomly chosen, these increases in teseting propensity are unrelated to individual health status. State 3 does no such randomization.

Given the assumptions of the model, we estimate prevalence using the following information:

- Data on whether someone got tested ($R_i$)
- Data on test results for those who got tested ($Y_i|_{R_i=1}$)
- Data on whether someone was in the treatment group that had a higher probability of getting tested ($T_i$)

Figure 4 shows the results based on 50 simulations for each state.[5] The blue bars in the panel on the left show the observed positive rates among those tested. The thick red line shows the prevalence in the state and the black lines show the range from the minimum to the maximum across the simulations. This panel illustrates the widely recognized point that positive test rates are not informative about prevalence when sick people are more likely to be tested.

The blue bars in the panel on the right in Figure 4 show estimated prevalence results using the estimation approach outlined here. Again, the thick red line shows the prevalence in the state. The black lines show the range from 5th to 95 percentile of prevalence estimates across the simulations for each state.

The prevalence estimates for states 1 and 2 are, on average, quite accurate. In state 1 the blue bar shows that the average prevalence estimate across the simulations is 10.7 percent, which is close to the true prevalence of 10 percent. In state 2 the average prevalence estimate across the simulations is 20.03 percent, which is the very close to the true prevalence.

The estimates are, like any statistical measure, noisy. The prevalence estimates range from 9 percent to 14 percent for state 1 and from 15 percent to 25 percent for state 2. But the estimates are always better than using the raw test result data and on average are quite accurate. The accuracy of the estimates can be increased by increasing the sample size, the size of the treatment group or the magnitude of the effect of the randomization on the probability of being tested.

State 3 presents a very different story. Recall that for state 3 there is no randomization that affected the propensity to show up for testing. The blue bar shows that the average estimate of prevalence for state 3 is 48 percent, far from the true value of 30 percent. And the range of estimates is large: from 18 percent to 89 percent. The poor estimate for state 3 occurs because, as illustrated in Figure 2, the observed test results can be explained by many different combinations of $\rho$ and prevalence when there is no variable that affects the decision to test, but not the results of the test. In statistical terms, the model is essentially unidentified for state 3.

The practical implications for randomized testing efforts are clear and can be summarized as follows.

- Estimating $\rho$ is imperative. If not everyone randomly selected to be tested actually gets tested, it is clearly unwise to simply analyze the data as if we knew $\rho = 0$. As we saw in Figures 2 and 3, prevalence estimates generated if we assume $\rho = 0$ may be far from the true prevalence estimates if $\rho \neq 0$.
- It is generally infeasible (and highly imprecise at best) to estimate $\rho$ without a first-stage instrument.

[5] I use maximum likelihood via the optim function in R (for the likelihood, see, e.g., Dubin and Rivers (1989)).

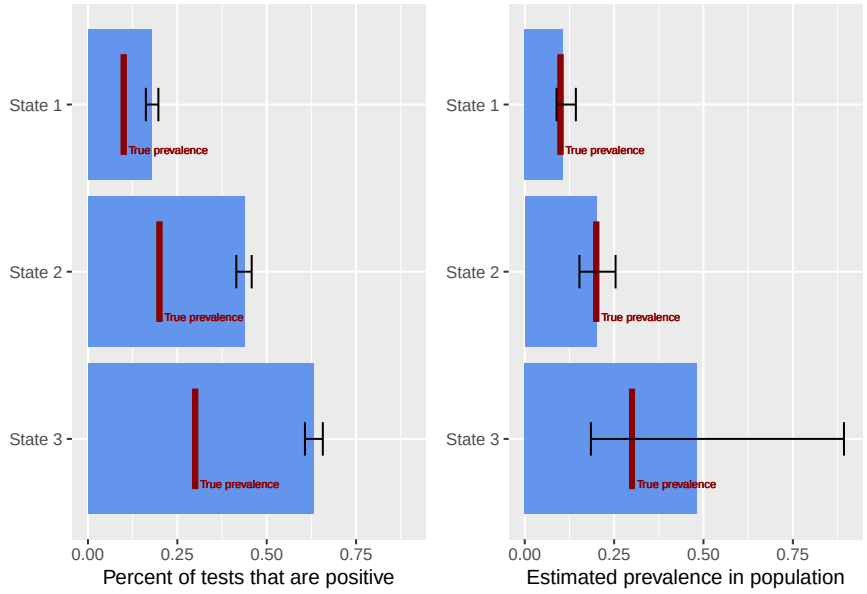


Figure 4: Results for three states in the simulation study

- Implementing a first-stage instrument is not difficult and adds relatively little cost to a randomization effort as it involves randomizing the selected sample into low and high contact groups, on the (easily testable) assumption that the high contact group will be more likely to respond.

**Section 4: Location-based testing**

Adding a first-stage instrument to large-scale randomized testing is effective, but may still be quite onerous for many locales which do not have the resources to implement such testing once, let alone across multiple time periods. From a policymaker's perspective, a more attractive approach to sampling would target some population that has already shown up at some location. For example, one could imagine policymakers implementing a testing program for people who show up at a given medical facility or a retail store or government office.

This approach creates clear challenges for estimation of population prevalence as the type of people who show up at any one of these locales is likely unrepresentative of the broader population in many respects, including most importantly in terms of their propensity to be sick.

This section explores if and when first-stage instruments may enable generalization from such testing protocols. This is both a constructive and critical exercise. We present a reasonable model in which generalization is feasible and also present a more general statement of how first-stage instruments can enable identification of prevalence and discuss the limits to such an approach.

*Testing at a Medical Facility*

In the first model we consider, we add another selection stage to the model consider so far. That is, suppose that there is an initial selection stage in which people decide to go to the hospital based on their perceived symptoms, which we label $\tau_i$. The latent propensity to appear at the hospital is

$$H_i^* = \kappa_0 + \tau_i$$

We observe $H_i = 1$ if $H_i^* > 0$, which implies that $H_i = 1$ if $\tau_i > -\kappa_0$.

At the hospital, patients are randomly divided into treatment ($T_i = 1$) and control ($T_i = 0$) groups. Whether someone is tested is determined by a latent propensity, similar to the equation used earlier:

$$R_i^* = \gamma_0 + \gamma_1 T_i + \tau_i$$

where $\tau_i$ is a mean-zero random variable and $T_i$ is the first-stage instrument. As before, we assume $\gamma_1 > 0$ and do not include other covariates for simplicity of exposition. We observe $i$'s test results if $R_i^* > 0$.

The randomization could be literally based on a random process or based on an essentially random process. Suppose, for example, there are two triage nurses and that patients see whomever is available next. If triage nurse 1 orders tests rarely while triage nurse 2 orders tests more aggressively, we would have a pseudo-randomization that affects propensity to be tested, but does not directly predict health status.

Importantly, we assume for this model that the error in the hospital equation is the same error that affects propensity to be tested. Going to the hospital and being tested are two separate processes, but the underlying symptoms and personality characteristics drive both.

Because the proportion of people who show up at the hospital is (by necessity) greater than or equal to the proportion tested, $\gamma_0 \geq \kappa_0$. Combining this fact with $\gamma_1 > 0$, the $H_i = 1$ requirement is not binding because any $\tau_i > -\gamma_0 - \gamma_1 T_i$ (the condition for $R_i = 1$) will also be larger than $-\kappa_0$ (the condition for $H_i = 1$). This implies that the model looks quite similar to the model without the hospital step as the key condition for getting tested remains $R_i = 1$ if $\tau_i > -\gamma_0 - \gamma_1 T_i$.

It is straightforward to incorporate these elements into a maximum likelihood model. There are four types of observations: {H=0}, {H=1, R=0}, {H=1, R=1, Y=0}, {H=1, R=1, Y=1} and the likelihood for each is well defined and identified given the model. Simulations similar to what was presented earlier provide results that again appear unbiased and reasonably precise.

The reason we can estimate prevalence in this model is that the difference in observed positive test rates among the treatment and control groups is informative about $\rho$. As we have seen above, estimating $\rho$ facilitates identification of prevalence.

*More general considerations regarding location based testing*

The above model leveraged an assumption that the processes of showing up at the hospital and getting tested were both driven by a common factor of individual level symptoms.

Such a model does not describe much of the testing being undertaken. For example, the Centers for Disease Control and Prevention is planning to test 325,000 people in 25 metropolitan areas at blood donation centers by the fall of 2021 (Janes, 2020). The state of New York tested 15,000 people at grocery stores and community centers in April (Cuomo, 2020).

It is quite likely that the processes the lead people to show up and agree to testing at these sites produce samples that are unrepresentative not only demographically, but also in terms of their propensity to be (or have been) sick. To the extent this happens, we cannot be confident in generalizing from the results to the general population.

One rejoinder is that such as approach could at least provide a fixed tracking point allowing researchers to identify trends in infection rates, at least for the subpopulation who shows up at these locations. However, this is only true if $\rho$, the relationship between being sick and propensity of being tested, is constant over time. This could be true, but need not be as, perhaps, over time sick people become more (or perhaps less) likely to show up at these sites. Without building in

a capcity to assess $\rho$ we cannot be sure if changes in infection rates are due to population level changes or to changes in who shows up at the testing sites.

A FSI approach in this context can enable researchers to assess – and therefore account for – $\rho$. The approach would be relatively straightforward to implement: testing authorities would pick a random subset of the population to encourage to visit the testing site. Perhaps this would simply be a text message, email, phone call or letter. Or perhaps the encouragement could include payment or a coupon. The key is that there would be an identifiable group for who the probability of visiting the testing site is higher for reasons unrelated to their propensity to be sick.

Note that the FSI does not require, or even produce, a representative sample. Instead, the approach allows data produced by a potentially unrepresentative sample to have the properties that make identification of $\rho$ possible. And, as above, having an estimate of $\rho$ will enable estimation of prevalence even when the sample is unrepresentative. Specifically, $\rho$ is identified because we have two groups: one group that shows up naturally, without encouragement (the $T_i = 0$ group) and one group that includes people both people who would have shown up without encouragement and people who were pushed from "almost" showing up to showing up by the encouragement. If this marginal group who showed up in response to the encouragement are no healthier than those who showed up without encouragement, we have evidence that health is not related to propensity to test.

However, if the marginal group who showed up in response to the encouragement are healthier than those who showed up naturally, we have evidence that propensity to test is related to health. In the limit, if the positive test rate declines as we pull in more and more people, we have evidence that testing is related to health status.

This approach is quite feasible as it simply involves identifying a testing location and encouraging a random selection of people to show up for testing in addition to testing those who are not randomly encouraged. In this sense, the approach is a hybrid between the organically occurring testing we currently have in most places and a full-scale randomized test, something that will likely be much cheaper than a full-scale randomized test with near perfect compliance.

There are also reasons to be cautious about this approach. The estimate of $\rho$ is "local" in the statistical sense in that it is estimated only for those for whom the testing encouragement is sufficient to push from not testing to testing. It is possible that the relationship between testing propensity and health status is different for different groups.

However, the results can be useful even in light of this concern. First, they can be treated as a test of whether $\rho \neq 0$ for any subgroup. It is possible to rejecting this null hypothesis even with a local estimate. A failure to reject the null should not be taken as broad confirmation that $\rho = 0$ for all possible subgroups.

Given the relative low costs, it may be possible for testing authorities to implement use this approach to examine multiple treatments and subpopulations. If there is evidence that $\rho \neq 0$ for a given testing site, it would make sense to try other types of encouragement and other testing locations in order to get a sense of the heterogeneity in the population. Or, ideally, a full scale randomized sample could be run concurrently (with a first-stage instrument to deal with non-response) and results for a given location could be, at least tentatively, calibrated to the broader results.

On the other hand, if evidence is generally consistent with either $\rho = 0$ or with a fixed $\rho$ across treatments and locations, testing authorities may be more confident in generalizing.

### Section 5: More general functional forms

So far in this paper, the conceptual discussions have been couched in general terms and the models have been presented as an extension of the canonical bivariate normal Heckman model to the case in which the outcome variable is dichotomous (Heckman, 1979).

Much of the recent literature on non-ignorable non-response has focused on extending the FSI approach beyond models that assume errors are bivariate normally distributed. This section presents an overview of this progress.

The critical point here is that all of these advances require a first-stage instrument. Once a first-stage instrument has been built into the data collection process, these and other models can be estimated, providing analysts with a sense of how much estimates depend on modeling assumptions. Without a first-stage instrument, however, the only options are to use bounds (which are too wide to be useful) or to assume away non-ignorable non-response (which is unrealistic and creates large potential biases).

There are three major strands in the literature. First, non-ignorable non-response models with first-stage instruments can be identified under different parametric assumptions. For example, Miao, Ding and Geng (2016) show that a standard $t$ distribution with degrees of freedom equal to $\nu$ can be identified. McGovern et al. (2015) and Gomes et al. (2019) use copulas to explore a broad range of possible joint distributions of the error terms in the selection and outcome equations, selecting the final model based on model fit.

Sun et al. (2017) use doubly robust methods to estimate prevalence. In this approach one formulates parametric models for the probability of response as a function of covariates and health status and for probability of testing positive as a function of covariates and response propensity. If either of the two parametric models is correct, prevalence can be estimated. The attraction here is that method requires only one, but not necessarily both, of the models to be correct.

Second, Das, Newey and Vella (2003) present a nonparametric approach in which they first estimate a propensity to respond based on the response data and then use a polynomial function of that propensity score as a control variable for a continuous outcome variable.

Finally, first-stage instruments can improve the performance of bound estimators. Marden et al. (2018) present an approach that uses a first-stage instrument to produce bounds on prevalence. This approach does not produce a point estimate for prevalence or the effects of covariates, but does provide information about presence, direction and magnitude of selection bias that does not depend on parametric assumptions.

### Conclusion

Coronavirus testing is useful for many reasons. Some of the reasons have little to do with estimating prevalence. Most obviously, knowing if a person has corona virus can inform treatment and can focus contact tracing resources.

Coronavirus testing as a means to understand community prevalence is important as well. This knowledge can inform policy decisions about stay at home orders and help us predict future demands on the medical system.

The challenge is that it is difficult to generalize about disease prevalence in the general population from test results when those who are sick are more likely to get tested and different regions test at different rates.

Using weights and other standard tools in survey research does not solve the problem. In statistical terms, weights require non-response to be ignorable, something that is unlikely to be

true for testing as implemented in most health care systems because the people being tested are typically much more likely to sick than the population in general.

When non-response in testing is non-ignorable, there is no statistical magical bullet that will allow us to convert the testing results we currently have into credible estimates of prevalence. We saw, for example, that prevalence, testing rates and the degree of non-ignorability in the data interact to produce observed positive test rates. Many combinations of those three factors can explain the same observed test results.

Therefore, in the spirit of Rubin (2008)'s admonition that "design trumps analysis" we need to design our data collection to provide the information that makes analysis feasible.

The most obvious way to generate useful information for prevalence estimation is to implement large-scale randomized testing. Ongoing randomized testing is expensive, however, making it impractical for many communities and time periods. In addition, such testing will potentially suffer from non-ignorable non-response.

Large-scale randomized testing is not, strictly speaking, necessary in order to statistically identify community prevalence. A considerable and growing body of research indicates that non-ignorable, non-response models can estimate prevalence if we have a variable that predicts likelihood of getting tested, but does not predict the result of a test. These randomizations can be done much more cheaply than full-scale randomized testing, potentially allowing them to be implemented for specific communities and time periods.

The statistical analysis of data from first-stage randomizations is not trivial and depends in part on assumptions about functional form. But the analysis is feasible and can yield useful information. If there is less than 100 percent response rates for a randomized testing effort, a first-stage instrument is relatively easy to implement and produces the information necessary to account for potential non-ignorable non-response. For location based testing, testing authorities can use first-stage instruments to enable prevalence estimates based on unrepresentative samples under plausible assumptions. Given the relative low cost and flexibility of first-stage instruments, it is possible to use use multiple approaches in order to explore some of the assumptions underlying this approach to testing.

First-stage randomizations do not solve all testing challenges. The sensitivity and specificity of testing technology needs to be accounted for, as do the practical challenges of implementing any testing approach. The appeal of the approach, however, is clear as they can produce data that enables us to understand the trajectory of a disease outbreak across communities and time.

## Appendix 1: Intuition

*Observed positive test rates depend on correlation of errors*

The top panel of Figure 5 shows the expected value of positive test rates for the treatment (T=1) and control (T=0) groups.

- The blue line at the top shows positive rates for the control group. When $\rho = 0$, which implies the sick and healthy are equally likely to be tested, the observed testing results will be the prevalence. As we move away from this unlikely scenario and $\rho$ increases, the observed rates of positive tests among those tested increases as testing is relatively rare and sick people are more likely to be tested.

- The green line in the top panel of Figure 5 shows positive test results for those tested in the treatment group. When $\rho = 0$, the expected observed testing results will be the prevalence

because those tested are still a true random sample of the population.[6] As $\rho$ rises, the positive rates among those tested rise as well, but not as much as for the control group because the treatment group includes people who are less sick.

The bottom panel of Figure 5 shows $\Delta$, the difference in the expected positive rates for the treatment and control groups as a function of the correlation of errors. Given the bivariate normality assumption and the parameters relating to testing propensity, this rises in a predictable way.

Specifically, bigger differences in the percent positive among those tested in the treatment and control groups are associated with higher values of $\rho$, the parameter that characterizes the relationship between how sick a person is and their likelihood of being tested.

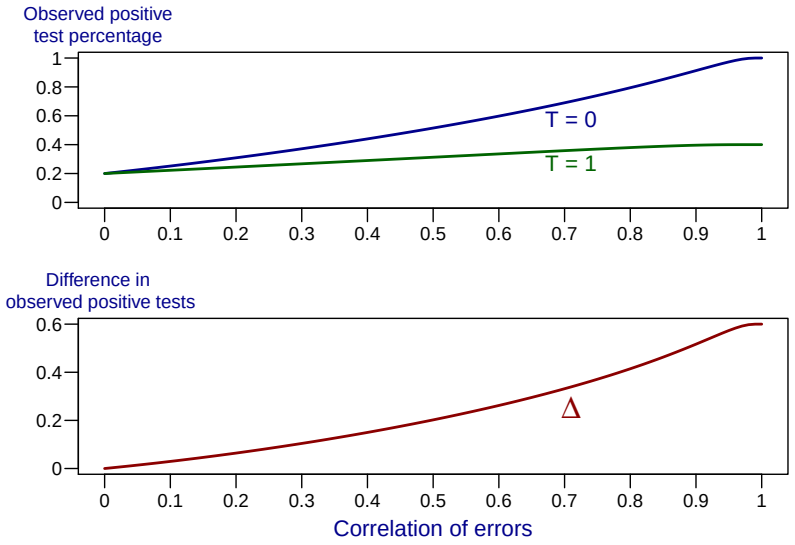


Figure 5: Positive testing rates as a function of correlated errors

*The observed difference in positive test rates is related to the correlation of errors.*

Figure 6 reverses the axes from the bottom panel of Figure 5 to illustrate that for (almost) any given difference in positive test rates among those tested in treatment and control groups, there is a $\rho$ that corresponds to it.

Using observed differences in positive test results for the treatment and control groups allows us to learn about the correlation of errors which, in turn, allows us to isolate the true prevalence. For example, if we knew $\rho$ for the case illustrated in Figure 2, we could back out the prevalence from observed positive test rates for most scenarios. This is true even when those getting tested are more likely to be sick.

The explanation in this appendix is an effort to provide intuition about how small-scale randomization can provide useful information. The actual statistical estimation process is quite different. In particular, the above figures depend on a specific value of prevalence, which is, of course, unknown. In the estimation processes, the prevalence and correlation are estimated simultaneously.

[6] When $\rho = 0$ the control group will be a random sample of, say 10 percent of the population while the treatment group will be a random sample of, say 15 percent of the population. Since both are random samples both should produce the same average test results even though those in the treatment group had a higher probability of being chosen.

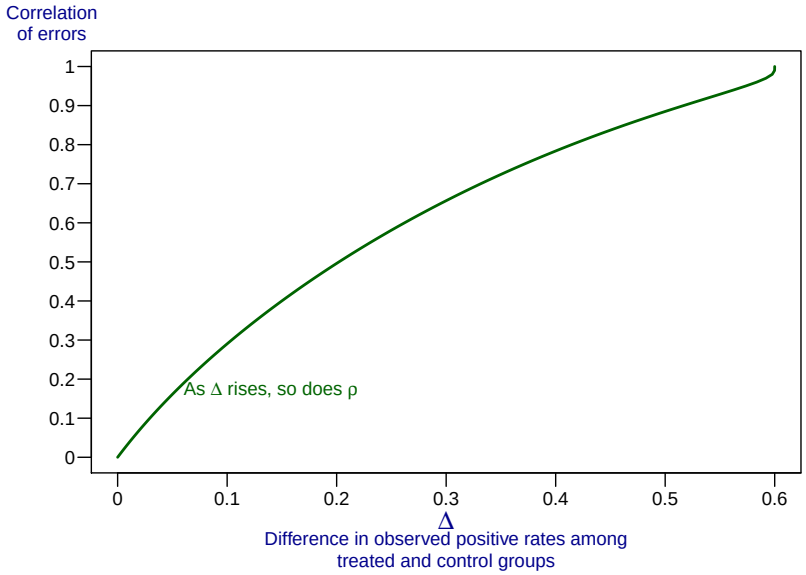


Figure 6: Relationship between correlated errors and observed differences in test results among those tested

## Appendix 2: Concordance tables

Table 2: Expected positive test rates for $\rho = 0.7$ and given testing rates and prevalence

| Test rate | Prevalence 0.02 | 0.04 | 0.06 | 0.08 | 0.1 |
|---|---|---|---|---|---|
| 0.02 | 0.31 | 0.47 | 0.57 | 0.65 | 0.71 |
| 0.04 | 0.23 | 0.37 | 0.47 | 0.55 | 0.61 |
| 0.06 | 0.19 | 0.31 | 0.41 | 0.49 | 0.55 |
| 0.08 | 0.16 | 0.28 | 0.37 | 0.44 | 0.51 |
| 0.10 | 0.14 | 0.25 | 0.33 | 0.40 | 0.47 |

Tables 2 and 3 show the expected positive test rates that would be observed for given prevalence and testing rate combinations given an assumption that errors are distributed with a bivariate normal distribution.

These concordance tables should be used cautiously. The parametric assumption may not hold and the tests used could vary in sensitivity across regions. They may, however, provide at least a rough guide to relating test results across regions. For example, Table 2 indicates that when $\rho = 0.7$, a region testing 2 percent of its population and observing 47 percent positive tests would have the same expected prevalence (4 percent) as a region testing 10 percent of its population and observing 25 percent positive tests.

Table 3: Expected positive test rates for $\rho = 0.3$ and given testing rates and prevalence

| Test rate | Prevalence 0.02 | 0.04 | 0.06 | 0.08 | 0.1 |
|---|---|---|---|---|---|
| 0.02 | 0.08 | 0.14 | 0.19 | 0.24 | 0.28 |
| 0.04 | 0.07 | 0.13 | 0.17 | 0.21 | 0.25 |
| 0.06 | 0.06 | 0.11 | 0.16 | 0.20 | 0.24 |
| 0.08 | 0.06 | 0.11 | 0.15 | 0.19 | 0.23 |
| 0.10 | 0.06 | 0.10 | 0.14 | 0.18 | 0.22 |

## References


Bailey, Michael A. 2019. "Designing Surveys to Account for Non-Ignorable Non-Response." Manuscript, Georgetown University.

Barnighausen, Till, Jacob Bor, Speciosa Wandira-Kazibwe and David Canning. 2011. "Correcting HIV Prevalence Estimates for Survey Nonparticipation Using Heckman-type Selection Models." *Epidemiology* 22(1):27–35.

Bendavid, Eran, Bianca Mulaney, Neeraj Sood, Soleil Shah, Emilia Ling, Rebecca Bromley-Dulfano, Cara Lai, Zoe Weissberg, Rodrigo Saavedra, James Tedrow, Dona Tversky, Andrew Bogan, Thomas Kupiec, Daniel Eichner, Ribhav Gupta, John Ioannidis and Jay Bhattacharya. 2020. "COVID-19 Antibody Seroprevalence in Santa Clara County, California." *medRxiv* .
**URL:** *https://www.medrxiv.org/content/early/2020/04/17/2020.04.14.20062463*

Cuomo, Andrew. 2020. "Amid Ongoing COVID-19 Pandemic, Governor Cuomo Announces Results of Completed Antibody Testing Study of 15,000 People Showing 12.3 Percent of Population Has COVID-19 Antibodies." *www.governor.ny.gov* (May 2).

Das, Mitali, Whitney K. Newey and Francis Vella. 2003. "Nonparametric Estimation of Sample Selection Models." *The Review of Economic Studies* 70:33–58.

Dubin, Jeffrey A. and Doug Rivers. 1989. "Selection Bias in Linear Regression, Logit and Probit Models." *Sociological Methods & Research* 18:360–390.

Dutwin, David and Paul J. Lavrakas. 2016. "Trends in Telephone Outcomes, 2008 - 2015." *Survey Practice* 9.

Gomes, Manuel, Rosalba Radice, Jose Camarena Brenes and Giampiero Marra. 2019. "Copula Selection Models for Non-Gaussian Outcomes that are Missing Not at Random." *Statistics in Medicine* 38:480–496.

Gudbjartsson, Daniel F, Agnar Helgason, Hakon Jonsson, Olafur T Magnusson, Pall Melsted, Gudmundur L Norddahl, Jona Saemundsdottir, Asgeir Sigurdsson, Patrick Sulem, Arna B Agustsdottir, Berglind Eiriksdottir, Run Fridriksdottir, Elisabet E Gardarsdottir, Gudmundur Georgsson, Olafia S Gretarsdottir, Kjartan R Gudmundsson, Thora R Gunnarsdottir, Arnaldur Gylfason, Hilma Holm, Brynjar O Jensson, Aslaug Jonasdottir, Frosti Jonsson, Kamilla S Josefsdottir, Thordur Kristjansson, Droplaug N Magnusdottir, Louise le Roux, Gudrun Sigmundsdottir, Gardar Sveinbjornsson, Kristin E Sveinsdottir, Maney Sveinsdottir, Emil A Thorarensen, Bjarni

Thorbjornsson, Arthur Love, Gisli Masson, Ingileif Jonsdottir, Alma Moller, Thorolfur Gudnason, Karl G Kristinsson, Unnur Thorsteinsdottir and Kari Stefansson. 2020. "Spread of SARS-CoV-2 in the Icelandic Population." *medRxiv* .
**URL:** *https://www.medrxiv.org/content/early/2020/03/31/2020.03.26.20044446*

Heckman, James J. 1979. "Sample Selection Bias As a Specification Error." *Econometrica* 47:153–162.

Ioannidis, John P.A. 2020. "A Fiasco in The Making? As the Coronavirus Pandemic Takes Hold, We Are Making Decisions Without Reliable Data." *StatNews.com* (March 17).
**URL:** *www.statnews.com/2020/03/17/a-fiasco-in-the-making-as-the-coronavirus-pandemic-takes-hold-we-are-making-decisions-without-reliable-data/*

Janes, Chelsea. 2020. "How many people are infected with the coronavirus? A major study will attempt to provide an answer." *Washington Post* (May 20).

Kennedy, Courtney and Hannah Hartig. 2019. "Response Rates in Telephone Surveys Have Resumed Their Decline." *Pew Research.org* .
**URL:** *http://www.pewresearch.org/fact-tank/2019/02/27/response-rates*

Lorch, Scott A., Michael Baiocchi, Corinne S. Ahlberg and Dylan E. Small. 2012. "The Differential Impact of Delivery Hospital on the Outcomes of Premature Infants." *Pediatrics* pp. 270–278.

Manski, Charles F. and Francesca Molinari. 2020. "Estimating the COVID-19 Infection Rate: Anatomy of an Inference Problem." Manuscript, Northwestern University.
**URL:** *https://arxiv.org/abs/2004.06178*

Marden, Jessica R., Linbo Wang, Eric J. Tchetgen Tchetgen, Stefan Walter, M. Maria Glymour and Kathleen E. Wirth. 2018. "Implementation of Instrumental Variable Bounds for Data Missing Not at Random." *Epidemiology* 29:364–368.

Marra, Giampiero, Rosable Radice, Till Barnighausen, Simon N. Wood and Mark E. McGovern. 2017. "A Simultaneous Equation Approach to Estimating HIV Prevalence with Nonignorable Missing Responses." *Journal of the American Statistical Association* 112:484–496.

McGovern, Mark, David Canning and Till Barnighausen. 2018. "Accounting for Non-Response Bias Using Participation Incentives and Survey Design." *CHaRMS Working Papers 18-02, Centre for Health Research at the Management School (CHaRMS)* .
**URL:** *https://ideas.repec.org/p/qub/charms/1802.html*

McGovern, Mark, Till Barnighausen, Giampiero Marra and Rosable Radice. 2015. "On the Assumption of Bivariate Normality inSelection Models: A Copula Approach Applied to Estimating HIV Prevalence." *Epidemiology* 26:229–237.

Meng, Xiao-Li. 2018. "Statistical Paradises and Paradoxes in Big Data (1): Law of Large Populations, Big Data Paradox, and the 2016 Presidential Election." *The Annals of Applied Statistics* 12:685–726.

Miao, Wang, Peng Ding and Zhi Geng. 2016. "Identifiability of Normal and Normal Mixture Models with Nonignorable Missing Data." *Journal of the American Statistical Association* 111:1673–1683.

Mostashari, Farzad and Ezekiel J. Emanuel. 2020. "We Need Smart Coronavirus Testing, Not Just More Testing." *StatNews.com* (March 24).

Peress, Michael. 2010. "Correcting for Survey Nonresponse Using Variable Response Propensity." *Journal of the American Statistical Association* 105:1418–1430.

Puhani, Patrick. 2000. "The Heckman Correction for Sample Selection and Its Critique." *Journal of Economic Surveys* 14:53–68.

Rubin, Donald B. 2008. "For Objective Causal Inference, Design Trumps Analysis." *Annals of Applied Statistics* 2:808–840.

Siegel, Benjamin, Mark Abdelmalek and Jay Bhatt. 2020. "Coronavirus contact tracers' nemeses: People who don't answer their phones." *abcnews.go.com* (May 15):494–507.

Stolzenberg, Ross M. and Daniel A. Relles. 1997. "Tools for Intuition about Sample Selection Bias and Its Correction." *American Sociological Review* 62:494–507.

Sun, BaoLuo, Lan Liu, Wang Miao, Kathleen Wirth, James Robins and Eric J. Tchetgen Tchetgen. 2017. "Semiparametric Estimation with Data Missing Not at Random Using an Instrumental Variable." Manuscript, Harvard University.
**URL:** *https://arxiv.org/abs/1607.03197*

Vella, Francis. 1998. "Estimating Models with Sample Selection Bias: A Survey." *Journal of Human Resources* 23:127–169.

Wang, Sheng, Jun Shao and Jae Kwang Kim. 2014. "An Instrumental Variable Approach for Identification and Estimation with Nonignorable Nonresponse." *Statistica Sinica* 24:1097–1116.