Ganz, Scott

Working Paper

Hypothesis testing sustained declines in COVID-19 intensity

AEI Economics Working Paper, No. 2020-07

Provided in Cooperation with:
American Enterprise Institute (AEI), Washington, DC

Suggested Citation: Ganz, Scott (2020) : Hypothesis testing sustained declines in COVID-19 intensity, AEI Economics Working Paper, No. 2020-07, American Enterprise Institute (AEI), Washington, DC

This Version is available at:
https://hdl.handle.net/10419/280620

Hypothesis testing sustained declines in COVID-19 intensity

Scott C. Ganz

American Enterprise Institute, Georgia Tech School of Public Policy, and Georgetown McDonough School of Business

AEI Economics Working Paper 2020-07

June 2020

Hypothesis Testing Sustained Declines in COVID-19 Intensity

Scott C. Ganz*

Scott.Ganz@aei.org†

This draft: June 17, 2020

First draft: May 11, 2020

Abstract

Gating indicators recommended by CDC for phased reopening following COVID-19 shutdowns include "sustained decreases" or a "downward trajectory...over a 14-day period" in public health statistics. These criteria have proven difficult to use in practice, however, because different interpretations of the data have yielded inconsistent guidance for policymakers. To standardize local reopening decisions and provide needed clarity for policymakers, I propose to define a "sustained decrease" as 14 consecutive days of declining infection rates in a region's population. I offer a hypothesis-testing framework for whether this criterion has been met that accounts for sampling variability. My test distinguishes regions experiencing a sustained decrease at modest sample sizes and substantially outperforms CDC's recommended test. I then apply the methodology to public data from 23 states for the period ending June 1 and find that only New York passes my proposed test. I note, however, that the public data to which I apply my test is unlikely to be representative of the underlying infection rate in the population. This underscores the pressing need for policymakers to implement COVID-19 testing regimes designed to measure population-level trends in regional infection intensity.

*American Enterprise Institute, Georgia Tech School of Public Policy, and Georgetown McDonough School of Business.

†The author thanks Nate Atkinson, Alex Brill, Amy Ganz, Keith Hennessey, Burke O'Brien, Michael Sarinsky, Michael Strain, and Stan Veuger for their helpful questions and comments. Special thanks to Jonathan Wand, both for introducing shape-constrained inference to me as a graduate student and for sharing the R code used in Wand (2012).

1 Introduction

Governors across the United States are watching the data on COVID-19 cases to guide decisions about the phased reopening of businesses, schools, and public spaces. Gating indicators for phased reopening recommended by the White House Coronavirus Task Force (2020) and CDC (2020) include "sustained decreases" or a "downward trajectory...reported over a 14-day period" in a series of public health statistics, such as the positive test rate, number of new cases, and number of symptomatic patients. Despite its intuitive appeal, these criteria have proven difficult to use in practice, because different interpretations of what characterizes a "sustained decrease" or "downward trajectory" have yielded inconsistent guidance for policymakers. In particular, whether a brief spike in data on COVID-19 infection rates requires that the "clock be reset" has become a major source of contention in many regions as policymakers, constituents, and the media have reached different conclusions about whether these short-term increases are the result of sampling variability or a resurgent pandemic (Blitz and Pacsale, 2020; Campbell, 2020; Pohl, 2020; Skahill, 2020; Taylor, 2020).

Guidance by the White House Coronavirus Task Force (2020) and CDC (2020) begs the question of what precisely they mean by a "sustained decrease" or "downward trajectory." This ambiguity creates two problems for policymakers. First, policymakers using the federal guidance to inform reopening decisions have been unable to clearly communicate to the public what types of data are consistent with a sustained decrease (and what types of data are not). Second, the ambiguity makes it impossible to evaluate the statistical properties of the tests proposed in CDC (2020). As a result, policymakers are unable to tell whether the criteria they are applying to the data to inform reopening decisions are achieving intended policy goals.

In this paper, I propose a clear definition for "sustained decrease" that is consistent with guidance from the White House, CDC, and other public health researchers: the rate of infection in a region's population experiences declines for 14 consecutive days. I then develop a statistical test to determine whether a region has experienced 14 days of sustained decrease in the intensity of the pandemic that takes into account sampling variability. I show that my proposed test, at modest sample sizes, successfully distinguishes regions experiencing a sustained decrease from regions that are not, and, further, that my test substantially outperforms the criteria for identifying 14-day sustained decreases

recommended in current CDC guidance. I also discuss how the framework can straightforwardly be applied to other theories about recent trends in COVID-19 positive test rates, including whether the pandemic has "near zero incidence" (CDC, 2020) or is experiencing a "rebound" (Gottlieb et al., 2020).

I then apply my test to publicly reported data from 23 states for the period ending June 1. Of these 23 states, only New York State passes the test at a 90 percent confidence level. Rather than indicating that New York is the only state experiencing sustained declines of infection intensity in the population, however, the peaks and valleys in the public data in many states are likely the result of unrepresentative sampling driven by testing programs that prioritize residents experiencing COVID-19-like symptoms or who have an elevated risk of contracting the virus. I thus conclude by emphasizing the pressing need for policymakers to invest in COVID-19 testing programs designed to produce consistent estimates of the underlying rate of infection in a region. Given the significant health and economic costs associated with reopening too soon or extending statewide shutdowns for too long, and the modest costs of additional COVID-19 testing, the benefits of widespread randomized testing on the scale necessary to apply the proposed framework outstrip the costs many times over.

2 "Sustained Decrease" as a Monotonicity Constraint

A good definition for "sustained decrease" in the intensity of the COVID-19 pandemic is that the rate of infection in a region's population declines for 14 consecutive days. This definition is consistent with the requirement of "14-days of consecutive downward slope" in CDC (2020). It is also in line with the recommendations of public health researchers that states should resist transitioning to the next phase of reopening until they observe "a sustained reduction in cases for at least 14 days" and should reinstate stay-at-home restrictions if "there is a sustained rise in new cases" (Gottlieb et al., 2020). This definition is also reflected implicitly in the common practice among analysts to rely on rolling averages both to estimate the infection rate and to visually examine whether the smoothed curve has been in decline for a sufficient number of consecutive days (Collins and Leatherby, 2020; Johns Hopkins University, 2020; Silver, 2020; Wisniewska et al., 2020). Finally, this definition is supported by epidemiological models and related public health guidance that define

pandemics in phases, often characterized by a peak phase followed by a period of monotonic decline in disease activity (see, e.g., Best and Boice, 2020; Fukuda and World Health Organization, eds, 2009; Holmdahl and Buckee, 2020).

Due to sampling variability and other data imperfections, however, 14 days of decreasing infection rates may not produce 14 consecutive days of declining public health indicators. A good statistical test for whether rates in the population are experiencing a sustained decrease needs to take into account that the test is being conducted on a limited sample. A straightforward way to test the theory that the data are consistent with 14 consecutive days of declining infection rates, then, is to apply a monotonicity constraint to the daily infection rate and evaluate the extent to which the constrained model departs from the observed data. In other words, the best-fitting model in which the infection rate on one day is restricted to be less than or equal to the infection rate the day prior should not do too much violence to the data.

The statistical methodology proposed here *directly* tests the hypothesis that the prior 14 days of data is consistent with a monotonically decreasing function. The basic structure, developed in Wand (2012) and built upon the "model confidence set" framework developed in Hansen et al. (2011), is as follows: for a given public health indicator — e.g., the positive COVID-19 test rate — identify a series of "shape constrained" models (Silvapulle and Sen, 2005; Wand, 2012) that reflect possible theories about trends in the prior two weeks of data. Then, sequentially cull the proposed models one-by-one until the data cannot meaningfully differentiate between the models that remain, based on a predefined confidence level. The remaining models are called the "model confidence set" (MCS).

Policymakers can use this method to determine which hypotheses about recent trends are supported by the data. For example, the test might show that public health indicators are best characterized by a function with an inverse u-shape (indicating that the peak in cases occurred in the prior two weeks) or by a constant line (which would be consistent with a prolonged plateau). If instead the monotonic decreasing model is included in the MCS and all other models that are inconsistent with a monotonic decrease have been culled, then policymakers should feel confident that the 14-day sustained decrease criterion is satisfied.

CDC guidance for policymakers evaluating whether a region is experiencing "sustained decline" in positive test rates, in contrast, is based on a linear regression model fit to the prior 14 days of

data. Specifically, CDC (2020) recommends that regions evaluate positive test rates according to whether they observe "14 consecutive days of downward trend," and then adds additional criteria intended to rule out u-shaped or inverse u-shaped (IUS) patterns in the data. Hypothesis tests based on linear models, however, are ill-suited to this task. They risk being overly restrictive because infection rates are unlikely to follow a linear progression. Depending on the shape of the peak and the speed of the pandemic recovery, epidemiological models suggest the post-peak phase could be linear, convex, concave, or reverse-sigmoid. And, statistical tests based on slopes of linear models that use traditional significance levels are likely to be too conservative for policymakers, because critical values are set in order to decrease the likelihood of false positives at the cost of high false negative rates. Slope parameters fit to data sampled from populations with consistent but slow declines in infection intensity, therefore, are unlikely to be statistically distinguishable from zero at a high confidence level.

Hypothesis tests based on linear models also risk being overly permissive, because they fail to differentiate data with a slow but consistent decline from data with an inverse u-shape or a u-shape. If the infection rate declines at a faster rate in the post-peak phase than it increases during the peak phase or if the growth and decline are symmetric around a peak that occurred in the first half of the observation window, a linear model will falsely identify an inverse u-shape as a 14-day sustained decrease. Data consistent with steep declines from a peak in the beginning of the observation window followed by a rebound at the end of the observation window may also be incorrectly identified as a sustained decrease by a linear model. As I show later in this paper, even after taking into account the additional criteria introduced in CDC (2020) to reject non-monotonic trends, CDC's proposed test over-rejects data that is consistent with a 14-day monotonic decrease and under-rejects data consistent with inverse u-shaped or u-shaped trends.

3 Shape-constrained Inference and the Model Confidence Set Framework

The methodology proposed here builds on the framework developed in Wand (2012), which draws heavily on Hansen et al. (2011), which I describe here briefly. Readers interested in large-sample statistical properties or computational aspects of the model confidence set framework are encouraged

to refer to Hansen et al. (2011). Those interested in the application of MCS to shape-constrained models should refer to Wand (2012). One important area where this paper differs from prior work is that I adapt the algorithm in Hansen et al. (2011) to take advantage of nesting in the hypothesized models, which improves the power of the test for small samples.

3.1 Defining the Model Confidence Set

Hansen et al. (2011) develops an algorithm to identify the set of best models $\mathcal{M}^*$ from a set of candidate models $\mathcal{M}^0$. The method evaluates the null hypothesis that all candidate models fit the data equally well, that is $\mathcal{M}^* = \mathcal{M}^0$. If the null hypothesis is rejected, then the worst-fitting model is eliminated from $\mathcal{M}^0$, a new set of models $\mathcal{M} \in \mathcal{M}^0$ is evaluated, and the process is repeated until the hypothesis $\mathcal{M}^* = \mathcal{M}$ cannot be rejected.

If the same significance level α is used in all tests, then the set of surviving models for which the hypothesis that $\mathcal{M}^* = \mathcal{M} \in \mathcal{M}^0$ cannot be rejected, i.e., the model confidence set $\hat{\mathcal{M}}^*_{1-\alpha}$, satisfies the following property:

$$\lim_{n\to\infty} P(\mathcal{M}^* \subset \hat{\mathcal{M}}^*_{1-\alpha}) \geq 1 - \alpha$$

Further, if $\mathcal{M}^*$ contains just one model, then a stronger result holds:

$$\lim_{n\to\infty} P(\mathcal{M}^* = \hat{\mathcal{M}}^*_{1-\alpha}) = 1$$

Hansen et al. (2011) also proposes a method for comparing the fit of each candidate model with the data, based on the Kullback-Leibler Information Criterion (KLIC). Define $Q(Z, \theta_j)$ as twice the negative log-likelihood function for the data Z evaluated at θ_j, i.e., $Q(Z, \theta_j) = -2\ell(\theta_j)$, where $j \in 1, ..., m$ indexes models in $\mathcal{M}$. The null hypothesis that the data cannot differentiate between the models in $\mathcal{M}$ states that $E[Q(Z, \theta_{0i}) - Q(Z, \theta^*)] - [Q(Z, \theta_{0j}) - Q(Z, \theta^*)]] = 0$ for all $i, j \in \mathcal{M}$, where θ_{0i} and θ_{0j} represent pseudo-true population parameters associated with models i and j, respectively, and θ^* represents the population parameters for a correctly specified model.

In order to test the null hypothesis, Hansen et al. (2011) introduces a range statistic $\mathcal{T}_{\mathcal{M}}$, where

$$\begin{aligned}\mathcal{T}_{\mathcal{M}} &= \max_{i,j \in \mathcal{M}} |E[Q(Z, \theta_{0i}) - Q(Z, \theta^*)] - E[Q(Z, \theta_{0j}) - Q(Z, \theta^*)]| \\ &= \max_{i,j \in \mathcal{M}} |E[Q(Z, \theta_{0i})] - E[Q(Z, \theta_{0j})]|\end{aligned}$$

The hypothesis that all of the models in $\mathcal{M}$ fit the data equally well is then tested by comparing $\mathcal{T}_{\mathcal{M}}$ against its null distribution using the significance level α.

3.2 Populating $\mathcal{M}^0$ using Shape Constraints

The advantage of the MCS framework is that any model can be included in the set of candidate models $\mathcal{M}^0$, including the monotonic decreasing model. Unlike a bivariate linear regression model, in which the coefficients represent best-fitting intercept and slope estimates, the parameters estimated in a monotonic decreasing model are associated with the estimated mean of the data for each day, under the restriction that the parameter value at time $t+1$ is less than or equal to the parameter value at time t. If the data actually reflect a monotonically decreasing pattern, then the constraint on the parameters will not bind and the parameter values will closely resemble the observed means. If the data does not exhibit a monotonically decreasing pattern, then the constraint will be binding and the parameter values will depart from the observed means, perhaps considerably.

However, I also include four other shape-constrained regression models in $\mathcal{M}^0$. Let θ_{jt} represent the parameter estimating the relevant health statistic for model j on a given day, where $t \in 1, ..., 14$ indexes dates.

1. Constant Model: The parameter value is the same on each day. $\theta_{jt} = \theta_{jt+1}$ for all t.

2. Monotonic Decreasing: The parameter value on a focal day is less than or equal to the parameter value on the preceding day. $\theta_{jt} \geq \theta_{jt+1}$ for all t.

3. Inverse U-shaped: The parameter values increase monotonically from day 1 to a peak day t^{peak}, then decrease monotonically from $t^{peak}+1$ through day 14. $\theta_{jt} \leq \theta_{jt+1}$ for all $t \leq t^{peak}$ and $\theta_{jt} \geq \theta_{jt+1}$ for all $t > t^{peak}$. Note that the monotonic decreasing function is equivalent to an inverse u-shaped function where the peak is on day 1.

4. Unrestricted: The parameter values can take any value. $\theta_{jt} \gtreqless \theta_{jt+1}$ for all t.

Importantly, these four models are nested and are ordered from the most-restrictive model (constant) to the least-restrictive model (unrestricted). An important implication of the nested hypotheses is that, under the null, the likelihood calculated from estimates of the pseudo-true parameters based on data generated by *the most restrictive model* is equal for all models in $\mathcal{M}$. I use this property of nested comparisons in order to improve the power of the testing procedure relative to the implementation proposed in Hansen et al. (2011).

3.3 MCS Implementation

Estimating $Q(Z, \theta_j)$ from the data requires estimating a correction for the overfitting of the sample estimate of the pseudo-true parameter $\hat{\theta}_j$. It also requires estimating the null distribution of $\mathcal{T}_{\mathcal{M}}$. Hansen et al. (2011) recommends a bootstrap implementation for both of these tasks. The methodology used here differs slightly from the implementation in Hansen et al. (2011) to take advantage of the nesting of the models in $\mathcal{M}^0$.

Specifically, given a set of models $\mathcal{M}$, I simulate B parametric bootstrap samples generated using the most restrictive model, where an individual bootstrap sample is defined as $Z_b^*(\mathcal{M})$. Under the null, $Q(Z_b^*(\mathcal{M}), \hat{\theta}_i) = Q(Z_b^*(\mathcal{M}), \hat{\theta}_j)$ for all $i, j \in \mathcal{M}$. Following Shibata (1997), I estimate the overfitting estimate under the null given the models in $\mathcal{M}$ for boostrap samples $b = 1, ..., B$:

$$\hat{k}_j^*(\mathcal{M}) = B^{-1} \sum_{b=1}^{B} \left[Q(Z_b^*(\mathcal{M}), \hat{\theta}_j) - Q(Z_b^*(\mathcal{M}), \hat{\theta}_{b,j}^*) \right]$$

where $\hat{\theta}_{b,j}^*$ are parameter values estimated from bootstrap sample b. With this in hand, the range statistic is calculated:

$$\hat{\mathcal{T}}_{\mathcal{M}} = \max_{i,j \in \mathcal{M}} \left| [Q(Z, \hat{\theta}_i) + \hat{k}_i^*(\mathcal{M})] - [Q(Z, \hat{\theta}_j) + \hat{k}_j^*(\mathcal{M})] \right|$$

Next, I compare $\hat{\mathcal{T}}_{\mathcal{M}}$ to the range statistic generated under the null hypothesis. Because $Q(Z_b^*(\mathcal{M}), \hat{\theta}_i) = Q(Z_b^*(\mathcal{M}), \hat{\theta}_j)$ for all $i, j \in \mathcal{M}$ under the null, the joint distribution

$$\{Q(Z, \hat{\theta}_1) + k_1^* - E[Q(Z, \theta_{01})], ..., Q(Z, \hat{\theta}_m) + k_1^* - E[Q(Z, \theta_{0m})]\}$$

can be estimated by the empirical distribution of

$$\{Q(Z_b^*(\mathcal{M}), \hat{\theta}_{b,1}^*) + \hat{k}_1^*(\mathcal{M}), ..., Q(Z_b^*(\mathcal{M}), \hat{\theta}_{b,m}^*) + \hat{k}_m^*(\mathcal{M})\}$$

Based on this empirical distribution, I generate a range statistic under the null hypothesis for each bootstrap sample b

$$\hat{\mathcal{T}}_{b,\mathcal{M}}^* = \max_{i,j \in \mathcal{M}} \left| [Q(Z_b^*(\mathcal{M}), \hat{\theta}_{b,i}^*) + \hat{k}_i^*(\mathcal{M})] - [Q(Z_b^*(\mathcal{M}), \hat{\theta}_{b,j}^*) + \hat{k}_j^*(\mathcal{M})] \right|$$

which I use to generate the distribution of $\hat{\mathcal{T}}_{\mathcal{M}}$ under the null. If the observed $\hat{\mathcal{T}}_{\mathcal{M}}$ exceeds the quantile of the simulated null distribution defined by the confidence level $1 - \alpha$, then the null hypothesis is rejected, the worst fitting model is removed from $\mathcal{M}$ and the algorithm is repeated. If not, then the models in $\mathcal{M}$ are the model confidence set $\hat{\mathcal{M}}_{1-\alpha}^*$.

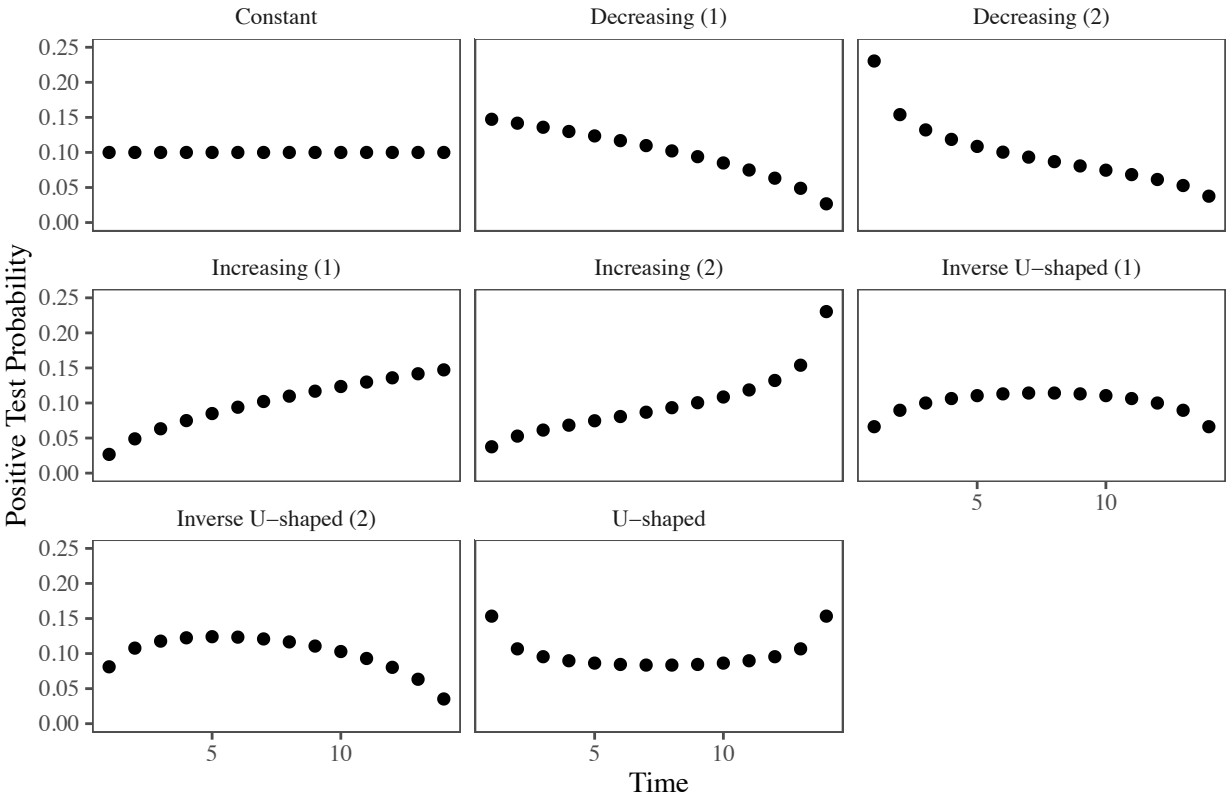
3.4 Small-sample Properties

I apply my test to simulated data in order to evaluate its small-sample properties. I generate data from eight populations characterized by different potential trends in COVID-19 infection intensity. Positive test rates for the eight populations are illustrated in Figure 1. Across all simulations, positive tests are 10 percent of total tests. Only two of the populations — Decreasing (1) and Decreasing (2) — are consistent with a 14-day sustained decrease. The others are consistent with a prolonged plateau [Constant], growing pandemic [Increasing (1) and Increasing (2)], peak phase [Inverse U-shaped (1) and Inverse U-shaped (2)], or a rebound [U-shaped].[1] I simulate data from each population 250 times and, for each synthetic dataset, estimate $\hat{\mathcal{M}}_{0.9}^*$ based on 250 bootstrap samples.

I estimate the probability that each of the four models in $\mathcal{M}^0$ – constant, monotonic decreasing, inverse u-shaped, and unrestricted — is included in $\hat{\mathcal{M}}_{0.9}^*$ for samples with 250, 500, 1000, and 2000 daily tests. I also estimate the probability that the monotonic decreasing model is included and the constant model is excluded from $\hat{\mathcal{M}}_{0.9}^*$, which is my proposed criterion for reopening. The

[1] The exact probabilities are derived from the following cumulative distribution functions. Constant: Beta(1,1); Decreasing (1): Beta(1, $\frac{3}{2}$); Decreasing (2): Beta($\frac{3}{4}$, $\frac{5}{4}$); Increasing (1): Beta($\frac{3}{2}$, 1); Increasing (2): Beta($\frac{5}{4}$, $\frac{3}{4}$); Inverse U-shaped (1): Beta($\frac{5}{4}$, $\frac{5}{4}$); Inverse U-shaped (2): Beta($\frac{5}{4}$, $\frac{3}{2}$); U-shaped: Beta($\frac{3}{4}$, $\frac{3}{4}$).

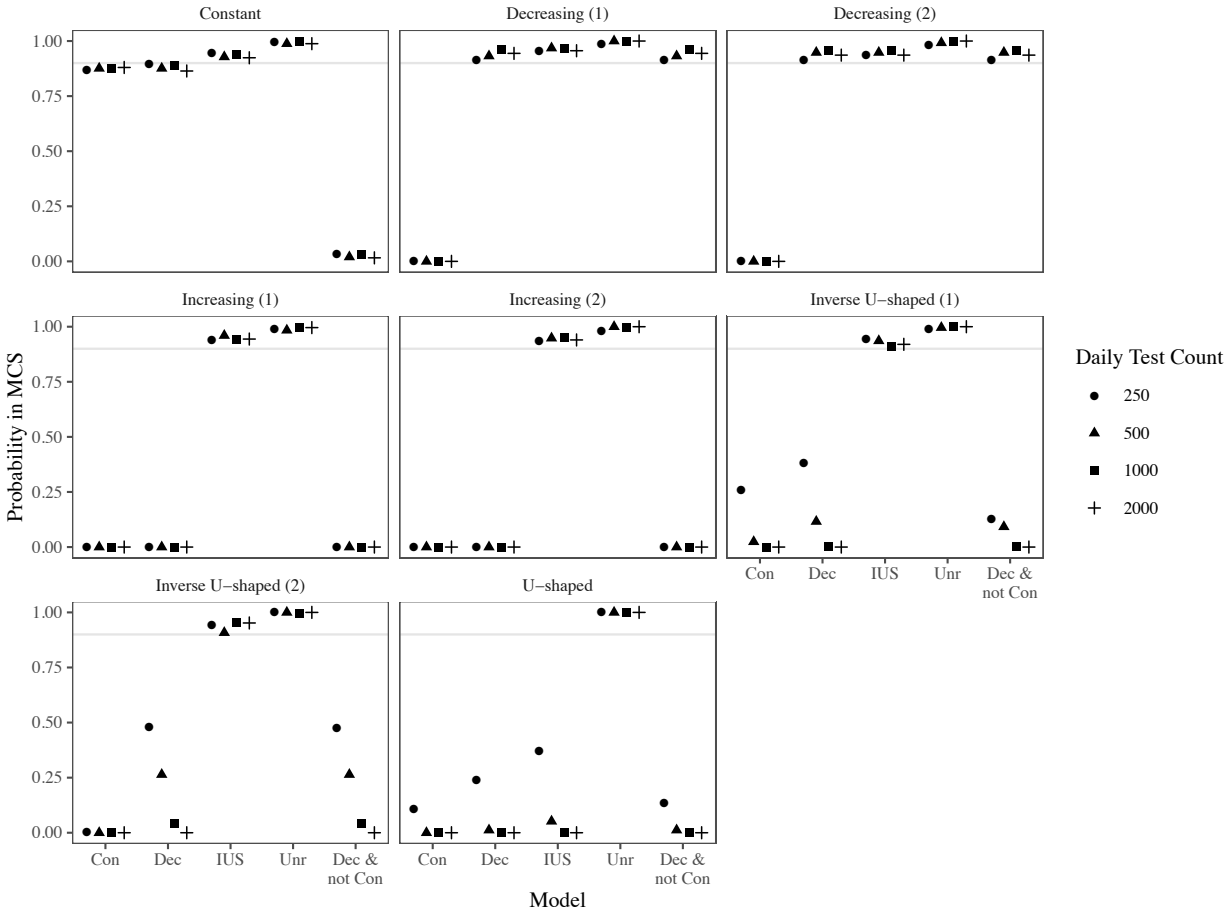
Figure 1: Population-level Positive Test Rates in Simulations



results are displayed in Figure 2. For the Constant, Decreasing (1), Decreasing (2), Increasing (1), and Increasing (2) populations, the probability of inclusion in $\hat{\mathcal{M}}^*_{0.9}$ reaches its target significance level (represented on the graphs by the light gray line) at 250 daily tests. In other words, the test achieves its intended Type II error rates when the average daily count of positive tests is just 25.

It is worth pointing out that random error in positive test rates will tend to produce over-rejection of the null hypothesis, leading a correctly-specified model to be excluded from $\hat{\mathcal{M}}^*_{0.9}$ with increased probability and resulting in inflated Type II error rates. This differs somewhat from hypothesis testing in a linear regression framework, where random error in the dependent variable reduces the probability of rejecting the null hypothesis. Therefore, in the testing framework proposed here and when testing the statistical significance of a slope parameter in a linear regression, adding random error to the outcome variable makes the test more conservative. Practically, this means that the number of tests required to achieve the target significance level is increasing in the error rate associated with individual COVID-19 tests.

Figure 2: Small-sample Simulation Results



For 250 daily tests, the Type I error rate is elevated for Inverse U-shaped (1), Inverse U-shaped (2), and U-shaped populations. For Inverse U-shaped (1), the test incorrectly identifies the data as consistent with the constant model 26 percent of the time and with the monotonic decreasing model 38 percent of the time, leading to the sustained decrease criterion being satisfied 13 percent of the time. The test has even higher rates of Type I error for Inverse U-shaped (2), because $\hat{\mathcal{M}}^*_{0.9}$ correctly excludes the constant model over 99 percent of the time but includes the monotonic decreasing model 48 percent of the time, leading to the criterion being satisfied in nearly 48 percent of the simulated datasets. The method also has elevated false-positive rates for the U-shaped simulations because $\hat{\mathcal{M}}^*_{0.9}$ includes the constant and monotonic decreasing models in 10 and 24 percent of simulated datasets, respectively, leading to a Type I error rate of 14 percent.

However, the Type I error rate declines quickly as the daily test count increases beyond 250 tests. For the Inverse U-shaped (1), Inverse U-shaped (2), and U-shaped (2) populations, the false-

positive rates are 1, 9, and 26 percent, respectively, for 500 daily tests; and less than 5 percent across all three populations for 1000 and 2000 daily tests. Based on these results, I recommend that policymakers using this methodology collect samples that are large enough so that the expected number of positive tests each day is greater than 100. For a 10 percent positive test rate, this means that one thousand daily tests are required. For a 1 percent daily test rate, I recommend daily samples of at least ten thousand tests. These testing volumes are in line with the current testing intensity in many states and are consistent with recommendations from other public health experts (Bigley, 2020).

The guidance in CDC (2020) for identifying 14 days of sustained decrease in positive test rates performs substantially worse. CDC (2020) recommends the following criteria to test for a downward trajectory in positive test rates: (1) "14 consecutive days of downward trend with up to 2-3 consecutive days of a grace period due to data inconsistency" and (2) "the 14th day must be lower than 1st day." Thus, the CDC criteria is satisfied if the data have:

1. A downward-sloping best-fit regression line.

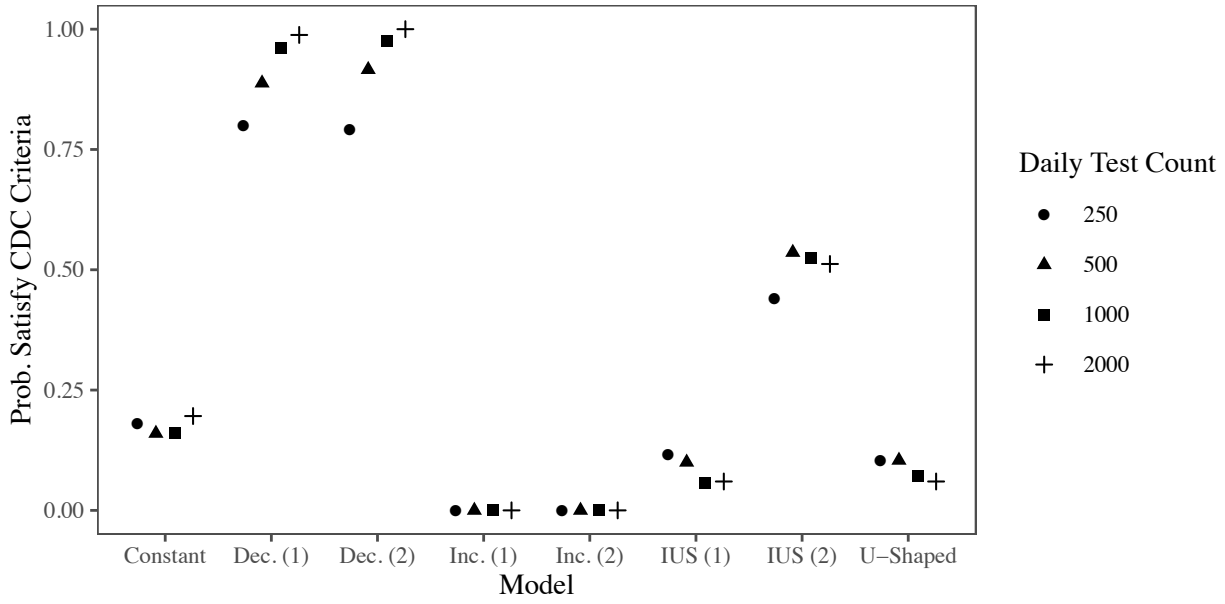
2. No more than one instance with 2 or 3 consecutive days of increasing positive test rates.

3. No instances of more than 3 consecutive days of increasing positive test rates.

4. Positive tests rates on day 14 that are lower than on day 1.

Figure 3 illustrates Type I and Type II error rate for the same simulated datasets using the CDC guidance. The Constant simulations experience Type I error rates between 16 and 20 percent for all four daily test counts. The Inverse U-shaped (1) simulations have false-positive rates between 6 and 11 percent. The Inverse U-shaped (2) simulations have the worst performance, with false-positive rates ranging from 44 to 54 percent. Finally, the U-shaped simulations incorrectly recommend reopening between 6 and 10 percent of the time. In addition to elevated Type I error rates, the CDC guidance also has elevated Type II error rates in small samples. The probability of satisfying the CDC criteria for the Decreasing (1) and Decreasing (2) populations is just 80 percent for 250 daily tests. However, in contrast to the Type I error rates, the probability of falsely recommending that a region remain closed is decreasing in the daily number of tests.

Figure 3: Simulation Results Based on CDC Criteria



3.5 Applying Other Shape Constraints

Other relevant shape constraints are easily added to the proposed framework. In this way, the method can accommodate maximum allowable positive test rates (Collins, 2020) or the virus reaching a "a near-zero plateau" (CDC, 2020). It can also be used to test for a "rebound" (Gottlieb et al., 2020). However, more daily tests would likely be required for the proposed framework to be able to successfully differentiate between more models or non-nested models.

In order to gauge the extent to which the power of the test is weakened by application to non-nested models, in particular, I calculate the Type I error rate for the Inverse U-shaped (2) simulations using the non-nested algorithm in Hansen et al. (2011). For daily samples of size 250, the false-positive rate is 60 percent using the non-nested algorithm, which is approximately 15 percentage points higher than the error rate for the nested algorithm and for the CDC guidance. The Type I error rate for the non-nested test declines to 8 percent for daily samples of 1000 and less than 1 percent for daily samples of 2000, which suggests that around two times as many tests would be required to accommodate non-nested models. Further power analyses would be required to determine the minimum number of daily tests needed in order to accommodate other potential models in $\mathcal{M}^0$.

4 Test Applied to U.S. Data

I next apply the test to positive COVID-19 test rates from 23 states, selected based on the quality of the public data made available through the COVID Tracking Project. I include only those states that have reported at least one test, at least one positive test, and at least one negative test in at least 24 of the 28 days prior to June 1. In addition, I require that each of the last 14 such days have more than 100 positive tests and a positive test rate less than 50 percent.

Before presenting the results of the test, I describe the data in two ways. First, in Figure 4, I graph positive test rates (in black) along with the best-fitting monotonically decreasing curve (in red) for the last 14 days with acceptable data. Although a few states to appear to exhibit a general downward trend, only New York has data that is well-modeled by the monotonically decreasing shape constraint.

Next, I calculate $E[Q(Z, \theta_{0j})] = Q(Z, \hat{\theta}_j) + \hat{k}_j^*$ for all four models in $\mathcal{M}^0$ in each state, where lower values of $E[Q(Z, \theta_{0j})]$ indicate better fit with the data. For simplicity, I estimate $\hat{k}_j^*$ based on bootstrap samples generated from $\hat{\theta}_j$ as in Hansen et al. (2011) so that the degrees-of-freedom estimates do not vary with $\mathcal{M}$. The results, presented in Table 1, are consistent with the visualizations. The value of $Q(Z, \hat{\theta}_j) + \hat{k}_j^*$ for the unrestricted model is considerably lower than the other three models for all states other than New York. For New York, in contrast, the monotonic decreasing, inverse u-shaped models, and unrestricted models fit the data equally well.

Given the data visualizations and the estimates of $Q(Z, \theta_{0j})$, it should not be surprising that only New York satisfies the 14-days sustained decrease criterion. For New York, the monotonic decreasing, inverse u-shaped, and unrestricted models are all in $\hat{\mathcal{M}}_{0.9}^*$. For all the other states, the unrestricted model is the only model in $\hat{\mathcal{M}}_{0.9}^*$.

5 Discussion and Conclusion

Observers might be surprised that the test identifies only New York as having satisfied the sustained decrease criterion. Among U.S. states, New York has experienced by far the most serious effects of the pandemic. Many other states less affected by COVID-19, meanwhile, have already begun phased reopenings.

That only a single state satisfies the proposed test is primarily an indictment of the quality of

Table 1: Model Fit Comparisons: $Q(Z, \hat{\theta}_j) + \hat{k}_j^*$ for $j \in \mathcal{M}^0$

State	Constant	Decreasing	IUS	Unrestricted
AZ	41009.77	40999.38	40888.10	40803.89
CA	265337.61	265174.68	264034.86	263219.03
CT	40514.15	40034.90	39988.03	39384.26
FL	94366.77	94368.75	92009.24	91741.33
GA	74271.56	74274.28	70910.83	69046.95
IA	32474.73	32307.78	32164.16	32048.67
IL	173419.24	172177.67	172080.20	171854.81
IN	45789.69	45741.64	45740.29	45705.40
LA	47327.35	47237.37	47141.46	46849.48
MA	92290.61	91805.66	91805.66	91679.27
MD	81426.86	79842.91	79842.91	79099.26
MN	59557.80	59159.57	59159.59	58997.23
MO	21494.53	21459.51	21227.09	21125.57
NC	76906.32	76888.25	76228.82	75516.10
NE	24108.26	24089.58	24005.59	23878.37
NJ	112285.25	106762.87	106596.76	105890.69
NY	182480.23	180821.76	180821.76	180820.12
OH	57058.30	56865.03	56773.79	56482.93
PA	73664.69	73475.91	73328.65	73165.58
TN	46328.70	46324.43	45959.02	45690.99
TX	127159.81	126964.60	125731.83	123411.48
VA	89566.93	89333.22	89005.33	87474.91
WI	46276.36	46068.33	46033.37	45923.02

the public data. Because testing regimes in the United States currently emphasize mitigating the spread of the virus, rather than measuring the underlying intensity of the pandemic in a region, the data on positive test rates do not offer consistent daily estimates of the current infection rate. As a result, no statistical test can draw accurate inferences about the virus' recent spread among regional populations. New York State's aggressive efforts to have all residents tested, whether symptomatic or not, provides some assurance that the data is consistent with underlying population trends. But, concerns about biased sampling due to self-selection or changes to the availability of inexpensive tests in New York still remain.

In order to account for errors caused by issues with data collection, many analysts have employed certain adjustments, including rolling averages, lags in inter-day comparisons, linear trends, and more sophisticated smoothing algorithms. Note that these same adjustments can be straightforwardly accommodated in the proposed framework. But policymakers should not have to rely on ad hoc adjustments to the data in order to apply objective criteria to future reopening decisions.

Instead, I strongly support proposals for collecting data that more accurately gauge underlying trends in the population (Greenstone, 2020; Padula, 2020). Getting representative samples large enough to describe these trends requires a public investment that pales in comparison to the benefits associated with not reopening too soon and not remaining closed for too long. If each state tested ten thousand residents per day at an average cost of $100 per test (Fehr et al., 2020) for the next 18 months, which is when experts predict that a vaccine will be widely available (Quinn, 2020), the total cost would be $27.4 billion. To put that number in context, the cost to the economy of a national shutdown is $20 billion *per day* (Mulligan et al., 2020). The cost of a widespread randomized testing program would be less than one percent of $3 trillion in total federal commitment to date for COVID relief (Snell, 2020). By bringing good data to the testing framework proposed here, policymakers will be able to make informed decisions about phased reopening based on clear and objective criteria, an investment with an expected payoff many, many times greater than the associated costs.

References

Best, Ryan and Jay Boice, "Where The Latest COVID-19 Models Think We're Headed – And Why They Disagree," *FiveThirtyEight*, May 2020.

Bigley, Sharon, "Many States Short of Covid-19 Testing Levels Needed for Reopening," *STAT*, April 2020.

Blitz, Matt and Jordan Pacsale, "D.C. Was Set For 14th Day Of Declining Cases, But New Data Reset Clock By Three Days," *DCist*, May 2020.

Campbell, Jon, "Coronavirus: New York shifts metrics, allowing Buffalo and Albany to open sooner," *Rochester Democrat and Chronicle*, May 2020.

CDC, "CDC Activities and Initiatives Supporting the COVID-19 Response and the President's Plan for Opening America Up Again," Technical Report, U.S. Department of Health and Human Services Centers for Disease Control and Prevention May 2020.

Collins, Keith, "Coronavirus Testing Needs to Triple Before the U.S. Can Reopen, Experts Say," *New York Times*, April 2020.

_ and Lauren Leatherby, "Most States That Are Reopening Fail to Meet White House Guidelines," *The New York Times*, May 2020.

Fehr, Rachel, Karen Pollitz, and Jennifer Tolbert, "Five Things to Know about the Cost of COVID-19 Testing and Treatment," *KFF*, May 2020. Library Catalog: www.kff.org.

Fukuda, Keiji and World Health Organization, eds, *Pandemic Influenza Preparedness and Response: A WHO guidance document*, Geneva: World Health Organization, 2009.

Gottlieb, Scott, Caitlin Rivers, Mark B. McClellan, Lauren Silvis, and Crystal Watson, "National Coronavirus Response: A road map to reopening," Technical Report, American Enterprise Institute, Washington, DC March 2020.

Greenstone, Michael, "The U.S. Should Pay People to Get Tested for COVID-19," *Washington Post*, May 2020.

Hansen, Peter R., Asger Lunde, and James M. Nason, "The Model Confidence Set," *Econometrica*, 2011, *79* (2), 453–497.

Holmdahl, Inga and Caroline Buckee, "Wrong but Useful – What Covid-19 Epidemiologic Models Can and Cannot Tell Us," *New England Journal of Medicine*, May 2020. preprint.

Johns Hopkins University, "New Cases of COVID-19 In World Countries," *Coronavirus Resource Center*, May 2020.

Mulligan, Casey B., Kevin M. Murphy, and Robert H. Topel, "Some Basic Economics of COVID-19 Policy," *Chicago Booth Review*, April 2020. Library Catalog: review.chicagobooth.edu.

Padula, William V., "Why Only Test Symptomatic Patients? Consider Random Screening for COVID-19," *Applied Health Economics and Health Policy*, April 2020, pp. 1–2.

Pohl, Jason, "Remote California County Abruptly Halts Reopening Plans, Citing First COVID-19 Cases," *The Sacramento Bee*, May 2020.

Quinn, Melissa, "Fauci Predicts U.S. Should Have "Couple of Hundred Million Foses" of Coronavirus Vaccine by New Year," *CBS News*, June 2020. Library Catalog: www.cbsnews.com.

Shibata, Ritei, "Bootstrap Estimate of Kullback-Liebler Information for Model Selection," *Statistica Sinica*, 1997, *7* (2), 375–394.

Silvapulle, Mervyn J. and Pranab Kumar Sen, *Constrained Statistical Inference: Inequality, order, and shape restrictions* Wiley series in probability and statistics, Hoboken, N.J: Wiley-Interscience, 2005.

Silver, Nate, "Coronavirus Cases Are Still Growing In Many U.S. States," *FiveThirtyEight*, April 2020.

Skahill, Patrick, "COVID-19 Hospitalizations Jump After Two Week Decline as State Reports 77 Additional Deaths," *The CT Mirror*, May 2020.

Snell, Kelsey, "Here's How Much Congress Has Approved For Coronavirus Relief So Far And What It's For," *NPR.org*, May 2020. Library Catalog: www.npr.org.

Taylor, Janelle Irwin, "Does Thursday's Coronavirus Spike Reset the Clock on Gov. DeSantis' Plan to Reopen the Economy?," *Florida Politics*, April 2020.

Wand, Jonathan, "Testing Competing Theories with Shape Contrained Inference," April 2012. mimeo.

White House Coronavirus Task Force, "Guidelines for Opening Up America Again," Technical Report, White House, Washington, DC April 2020.

Wisniewska, Aleksandra, Alan Smith, Max Harlow, David Blood, Steven Bernard, John Burn-Murdoch, and Cale Tilford, "Coronavirus Tracked: the latest figures as countries fight to contain the pandemic," *Financial Times*, February 2020.

Figure 4: Observed Data vs. Monotonic Decreasing Model in 23 States