Safely learning from new non-randomized treatments: Assessing the effect of remdesivir on COVID-19 mortality

David Ami Wulf[1], Brian L Hill[2], Jeffrey N Chiang[3], Onyebuchi A. Arah[1,4], David Goodman-Meza[5], and Chad Hazlett[1,6]

[1]Department of Statistics, UCLA
[2]Department of Computer Science, UCLA
[3]Department of Computational Medicine, David Geffen School of Medicine, UCLA
[4]Department of Epidemiology, Fielding School of Public Health, UCLA; Research Unit for Epidemiology, Department of Public Health, Aarhus University, Aarhus, Denmark.
[5]Division of Infectious Diseases, David Geffen School of Medicine, UCLA
[6]Department of Political Science, UCLA

Abstract

Background: Investigation of treatment effects using patient data from outside of randomized trials is common, but, even when accompanied by sensitivity analyses, can leave readers overconfident in point estimates that rely on improbable assumptions.

Methods: We analyzed the electronic health records of 136 COVID-19 positive patients hospitalized at a large university hospital system over the course of 110 days early in the pandemic. Through the stability-controlled quasi-experiment (SCQE), we utilized rapid changes in remdesivir usage over this period to provide a range of possible causal effects of remdesivir on COVID-19 mortality (within 28 days).

Results: Remdesivir use was initially high, then dropped during a no-use period (which saw 24.5% mortality), and then returned to high-usage. Remdesivir significantly ($p<0.05$) reduced mortality risk among those who took it if baseline mortality was at least 1.9 percentage points (8%) higher among patients in the combined high-use period (that is, had remdesivir not been used in this period either) than those in the middle no-use period. It could only have been harmful if baseline mortality dropped to 0% in the high-use period.

Conclusions: The assumptions required for a beneficial effect of remdesivir are plausible, but not defensible with confidence, while the assumptions required to declare it harmful are nearly impossible. More broadly, without any assumption of unconfoundedness, the SCQE reveals what inferences can be credibly supported by evidence from observational data, making it useful when randomized trials have not yet produced clear evidence or in providing estimates for different populations.

Keywords: Causality; COVID-19; Remdesivir; Inference; Observational Studies; Data Interpretation, Statistical; Models, Statistical

During the initial response to COVID-19, numerous emergent treatments were given to patients outside of randomized trials and before their safety and efficacy were established. Epidemiologists and others rightly cautioned against learning about the effects of these medical treatments from the many thousands of patients who took them outside of trials. Yet, a policy of rejecting any information obtained outside of randomized trials risks missing out on potentially life-saving findings. In practice, physicians and their patients will make their choices based on what they understand given current information, whatever its quality. How then can researchers and clinicians come to understand what can and cannot be concluded on the basis of non-randomized uses, rather than either categorically rejecting them as a source of information, or risking over-confidence in potentially biased estimates?

When randomized trials are not yet available, cannot be conducted, or apply to different populations than those of interest, investigators often undertake non-randomized, observational approaches. These typically rely on covariate adjustment to address observable differences between the treated and non-treated patients. Whether by weighting, matching, or propensity score adjustment, the validity of these methods rest on an assumption that all confounders of the relationship between treatment assignment and the outcome of interest are observed ("no unobserved confounding" or "conditional ignorability"[1,2,3]). Researchers are understandably dubious about this assumption. Many authors take care to explicitly highlight these concerns, often labeling the resulting estimates as "suggestive." Still, unmeasured confounding could easily nullify or even reverse the sign of estimated effects.

Sensitivity analyses are vital in helping to quantify the risk of unobserved confounding, but many can be difficult to interpret, asking readers to mathematically formalize the "strengths" of potential confounders.[4,5,6] Furthermore, even when a sensitivity analysis demonstrates concerning vulnerability to confounding, readers may nevertheless regard the point estimate as the best available guess at the effect.

Instead of making a "point assumption" (e.g. precisely zero confounding) and then determining the fragility of the result, another set of approaches—of which ours is an example—ask users to reason about the range that some assumed, unobserved quantity can take. The results of these "partial identification" approaches show the range of causal effect estimates that are possible given that assumption. This also allows us to state what assumptions would have to believed to make a given claim. While users may express frustration at the lack of a singular point estimate, that is the virtue of these approaches, preventing readers from becoming over-confident in a deceptively precise estimate.

In this paper we consider one such partial identification approach referred to as the stability-controlled quasi-experiment (SCQE).[7] This method asks users to reason about and put plausible bounds on how the outcome would have changed over time, absent changes in the new treatment. The basic SCQE setting involves two cohorts, one of which received none of the treatment (the "no-use cohort") and the other of which saw significant treatment usage (the "high-use cohort"). Its assumption takes the form of the expected difference between these cohorts' average outcomes *for reasons other than the treatment* — that is, the differences we would expect between the cohort-wide average outcomes had the high-use cohort not seen its treatment usage increase. Given an assumed magnitude of this "baseline trend" difference, called δ, SCQE estimates the average treatment effect among the treated patients (ATT). We do not expect to know δ, but for any range of values deemed plausible, the corresponding range of effect estimates cannot be dismissed. Conversely, if we wish to argue for or against a beneficial or harmful effect, SCQE gives us the baseline trend values we must defend as being plausible or implausible.

This approach does not require an assumption of no unobserved confounding, because it does not rest on a comparison of the treated and non-treated patient groups. In settings where little is known about how treatment is assigned, and thus little hope that confounding can be ruled out, reasoning instead about plausible baseline trends provides a transparent alternative.

We use SCQE to investigate the effect of remdesivir on inpatient mortality when used as a COVID-19 treatment, measured in a multi-center academic medical setting in the early months of the pandemic. At the time, risk factors for in-hospital COVID-19 mortality and the effectiveness of proposed treatments were not well-understood, heightening concerns about confounding in observational studies investigating the impact of any such treatment. However, during the time of study, the use of remdesivir started high, dropped to zero, and then returned to a high level. This provides a particularly useful opportunity to leverage the distinct assumptions of SCQE in asking what can and cannot be reasonably claimed about the benefits of harms of remdesivir as employed outside of randomized trials.

Approach

Introduction and notation for SCQE

We begin with an idealized setting to develop intuition. Imagine two consecutive cohorts of hospitalized patients, with none of the earlier patients receiving the treatment of interest and a non-randomized 50% of the later patients being treated. Suppose, further, that the earlier cohort's 28-day cumulative mortality incidence is 20%, while the later cohort's cumulative mortality incidence is 15%. If we assumed (momentarily) that there are no differences in the expected mortality risk of the two cohorts *other than* those caused by changes in treatment, the entirety of the 5 percentage point (pp) decrease in mortality must be due to treatment introduction. That reduction occurs only though the treated 50% of patients, so their average mortality drop attributable to treatment must be 10pp.

To formalize these arguments for a population of size n, we introduce three length-n binary vectors: outcomes Y, with Y_i referring to patient i's 28-day mortality; cohort membership Z, with 0 and 1 indicating the earlier and later cohorts, respectively; and the treatment indicator D. Using the potential outcomes framework,[8,9] $Y_i(0)$ and $Y_i(1)$ refer to the outcomes we would have observed for patient i under non-treatment and treatment, respectively, regardless of their actual treatment status. Our estimand of interest is the ATT in the second cohort, $\mathbb{E}[Y(1)|Z{=}1, D{=}1] - \mathbb{E}[Y(0)|Z{=}1, D{=}1]$.

The simplifying assumption made above regarding the absence of cohort-wide differences in baseline risk can be written out as $\mathbb{E}[Y(0)|Z{=}1] - \mathbb{E}[Y(0)|Z{=}0] = 0$. However, this assumption is far too stringent in practice, and must be weakened to provide an inferentially "safe" tool. SCQE does just this, allowing the expected outcomes under non-treatment to differ between the cohorts by a prescribed amount—a "baseline trend"—which we define as $\delta \equiv \mathbb{E}[Y(0)|Z = 1] - \mathbb{E}[Y(0)|Z = 0]$.

In practice we expect three main sources of non-zero baseline trends: changes in other treatment practices between cohorts, differences in the composition of patients in the cohorts, and other trends over time (e.g. COVID-19 mutations). For any given choice of δ, whatever its source, the data tell us the logically implied treatment effect estimate. Mechanically this can be seen by separating the $\mathbb{E}[Y(0)|Z=1]$ term into an observable term and an unobservable term using the law of iterated expectations, and then solving for the unobservable term:

$$\begin{aligned}\mathbb{E}[Y(0)|Z=0] &= \mathbb{E}[Y(0)|Z=1]-\delta \\ &= \mathbb{E}[Y(0)|Z=1,D=1]\pi_1+\mathbb{E}[Y(0)|Z=1,D=0](1-\pi_1)-\delta,\end{aligned}$$

where $\pi_1 = P(D{=}1|Z{=}1)$ is the treatment rate in the high-use cohort, and thus

$$\begin{aligned}\mathbb{E}[Y(0)|Z=1,D=1] &= \frac{\mathbb{E}[Y(0)|Z=0]-\mathbb{E}[Y(0)|Z=1,D=0](1-\pi_1)+\delta}{\pi_1} \\ &= \frac{\mathbb{E}[Y|Z=0]-\mathbb{E}[Y|Z=1,D=0](1-\pi_1)+\delta}{\pi_1}\end{aligned}$$

This unobservable term, $\mathbb{E}[Y(0)|Z=1,D=1]$, is the mean potential outcome we would have seen among the treated patients in the high-use cohort, had they not received that treatment. Simply subtracting this term from the observed (treated) outcome for these patients gives the SCQE estimator for the ATT:

$$\begin{aligned}\widehat{\text{ATT}} &= \mathbb{E}[Y(1)|Z=1,D=1]-\mathbb{E}[Y(0)|Z=1,D=1] \\ &= \mathbb{E}[Y|Z=1,D=1]-\left(\frac{\mathbb{E}[Y|Z=0]-\mathbb{E}[Y|Z=1,D=0](1-\pi_1)+\delta}{\pi_1}\right)\end{aligned} \tag{1}$$

This estimate is not rooted in the more familiar comparison of treated to untreated groups, but rather of one cohort to another, and only after an adjustment. The shift in attention to the cohort rather than treatment groups avoids the myriad confounding problems that arise due to differences between those taking and not taking treatment. Instead, here the key assumption regards the potential for differences between the cohorts—specifically the difference in their average non-treatment outcomes—captured by δ. Rather than assuming only $\delta = 0$ (that there is no difference, in keeping with the "principle of concurrent control"), an estimate of the causal effect is recoverable at any postulated value of δ.

The effect estimate applies specifically to patients whose physicians deemed the treatment appropriate for them, and who agreed to take it—often a highly relevant group to learn about and possibly different from the eligible patients in an RCT. This approach is, mathematically, an extension of the instrumental variable (IV) approach in which "time" or cohort membership is used as an instrument, but allowing us to make an assumption about the baseline trend, δ, which would violate the exclusion and exchangeability restrictions required for IV (see Hazlett et al.[7] for discussion of the SCQE-IV connection and generalization of the natural experiment example).

Equation 1 provides an ATT estimate for any fixed δ. Rather than assuming a correct δ can be chosen, the range of ATT estimates corresponding to a range of δ values can be displayed without emphasizing a single preferred value. The practitioner or reader is then compelled to consider these assumptions directly in order to make any causal claims. Instead of initially making an infeasible assumption of conditional ignorability and then quantifying the impact of violations of that assumption with a sensitivity analysis, SCQE builds its sensitivity analysis into the identification process and displays the results in a way that highlights uncertainty. The range of δ values where practitioners should focus their attention — δ values we refer to as "plausible" — may be informed by domain knowledge and relevant data. For example, a simple comparison of baseline characteristics between the cohorts could suggest demographic, risk-factor, or treatment differences (not including the treatment of interest). Further, model-based prediction of non-treatment outcomes may evaluate all of these covariates at once, generating helpful risk estimates that can be compared between cohorts. Still, δ is neither observable nor identifiable from data, and we must offer a defense of any propositions about which values are plausible or implausible. Beyond the identification uncertainty paramaterized by δ, we must also consider sampling uncertainty, so we display the ATT estimate for each δ value alongside a confidence set (constructed as described in eAppendix 1).

Setting and data extraction

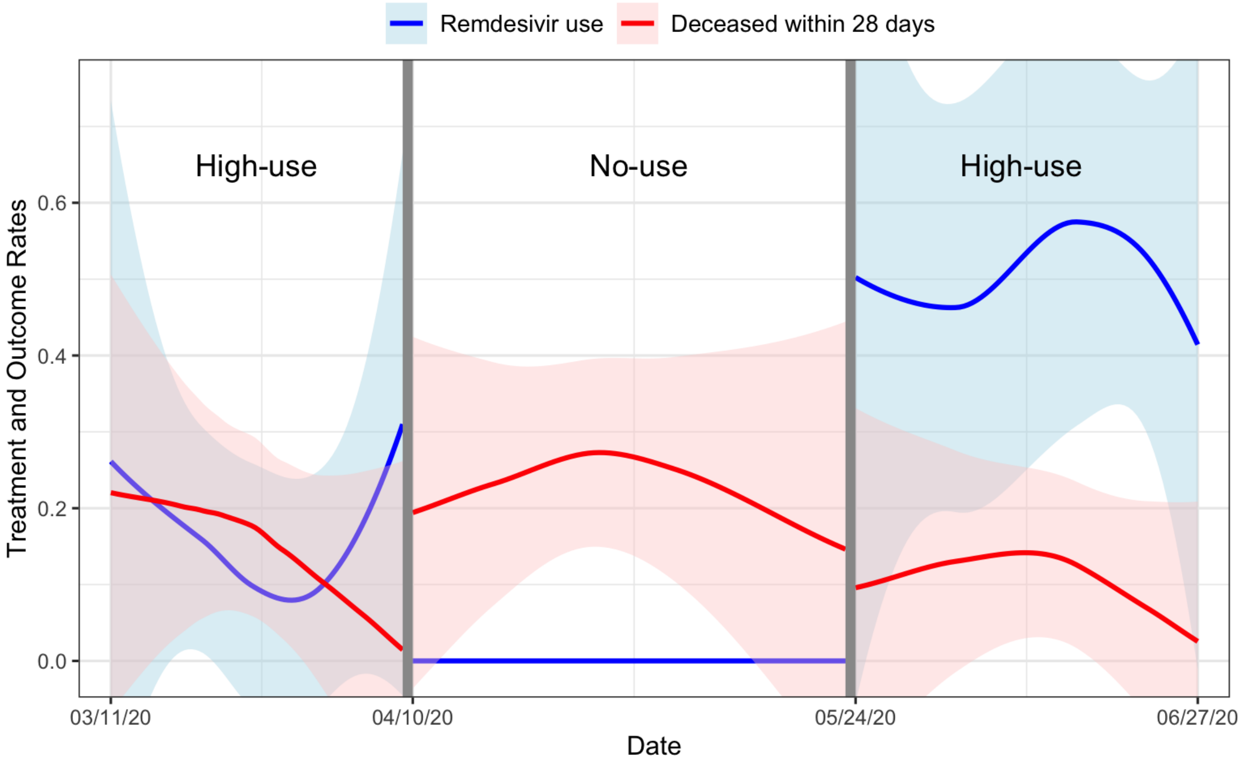
Remdesivir was used to treat COVID-19 in the United States as early as January 25, 2020 on a "compassionate-use" basis, delivered as a 200mg dose for one day, followed by 9 days of 100mg doses.[10] Shortly afterwards, randomized trials for remdesivir began recruitment.[11] On May 10, 2020, the Federal Drug Administration (FDA) granted an Emergency Use Authorization (EUA) for Remdesivir to treat COVID-19, which was followed by adjustments to that EUA and an eventual approval as treatment apart from the EUAs.[12] On November 20, 2020, the World Health Organization (WHO) issued a "conditional recommendation against the use of remdesivir in hospitalized patients with COVID-19," citing "low-certainty" evidence in suggesting "that remdesivir has possibly no effect on mortality".[13]

Our study population is comprised of inpatients with positive PCR tests for SARS-CoV-2 infection and an ICD-10 code U07.1 for COVID-19 at two hospitals at an academic medical center. Data extracted from the electronic heath record database included both pre-admission covariates such as demographics and comorbidities, as well as early hospitalization measures such as initial laboratory values and vital signs (within 24 hours of admission) and other COVID-19 treatments received (within 7 days of admission). We include remdesivir treatment initiated at any point in a hospitalization, though all but two remdesivir patients began treatment within 3 days of admission. We excluded patients transferred from another acute care facility as their prior remdesivir usage was unknown. Our outcome was 28-day mortality.

In constructing the cohorts, the "A-B-A" (high-use, no-use, high-use) structure we observe is useful in mitigating differences in treatment practices or patient characteristics that may shift roughly monotonically over time (e.g. hydroxychloroquine use, which generally decreased over time, and dexamethasone use, which increased over time). Remdesivir was not administered to any of the 53 COVID-19 patients admitted between April 10, 2020 and May 24, 2020, making this a suitable no-use cohort. The first part of the high-use cohort begins on 11 March 2020, the date of the first admitted patient in the dataset. The end date (the last date of the second part of the high-use cohort) was chosen to be June 27, 2020. This date avoided including a set of several patients with high dexamethasone usage who were admitted soon afterwards. The resulting combined high-use cohort is comprised of 83 patients, 46 admitted between March 11 and April 9 and 37 admitted between May 24 and June 27.

Results

Figure 1: Trends in remdesivir usage and 28-day mortality, indexed by date of hospital admission. LOESS curves and the corresponding 95% confidence intervals are shown.



In the no-use period, the 28-day cumulative mortality incidence was 24.5%. In the combined high-use period, 31.3% of patients were treated with remdesivir, and the cumulative mortality incidence was 13.3%: 15.4% among the treated patients and 12.3% among the non-treated. Figure 1 shows the smoothed treatment rates and cumulative mortality incidences over the study period.

SCQE estimates, like those using time as an instrument or the idealized natural experiment, begin with a comparison between cohorts rather than between treated and untreated groups. Whether the lower cumulative mortality incidence in the high-use period suggests a beneficial effect of the treatment depends on how mortality would have dropped in the high-use cohort for reasons other

Table 1: Baseline Characteristics of the No- and High-Use Cohorts

Covariate	No-Use	High-Use
Age	69.0	63.5
Male	0.51	0.59
White	0.34	0.43
Latinx	0.32	0.30
BMI	26.4	27.9
Admitted through ED	1.00	0.88
Diabetes	0.11	0.08
Asthma	0.04	0.02
CHD	0.02	0.04
Hyperlipidemia	0.09	0.00
Hypertension	0.13	0.17
Within 24 hours of admission,		
CRP (if measured)	10.3	10.4
No CRP measured	0.19	0.30
WBC	8.65	8.76
ANC	6.93	6.41
ALC	0.94	1.30
Creatinine	1.96	1.26
SpO_2	92.7	93.8
Oxygen Flow Rate (if used)	6.16	3.57
No Supplemental Oxygen	0.19	0.28
Ventilator use	0.13	0.19
ICU admission	0.34	0.29
Mortality	0.02	0.01
Within 7 days of admission,		
Dexamethasone	0.00	0.08
Other Corticosteroid[a]	0.15	0.19
Proning	0.08	0.06
Convalescent Plasma	0.02	0.00
Hydroxychloroquine	0.13	0.22
Heparin	0.47	0.53
Enoxaprin	0.57	0.54
Azythromycin	0.57	0.54
Ceftriaxone	0.70	0.55
Leronlimab trial[b]	0.25	0.05
Tocilizumab	0.11	0.06

No-Use cohort: N=53. High-Use cohort: N=83 and remdesivir treatment rate=31.3%. [a] Prednisone, Methylprednisolone, or Hydrocortisone. [b] Values for Leronlimab trial indicate enrollment in a still-blinded clinical trial, with a 2:1 treatment:placebo design. Thus we expect actual treatment rates of roughly 0.17 in the no-use cohort and 0.03 in the high-use cohort, for a difference of roughly 0.14.

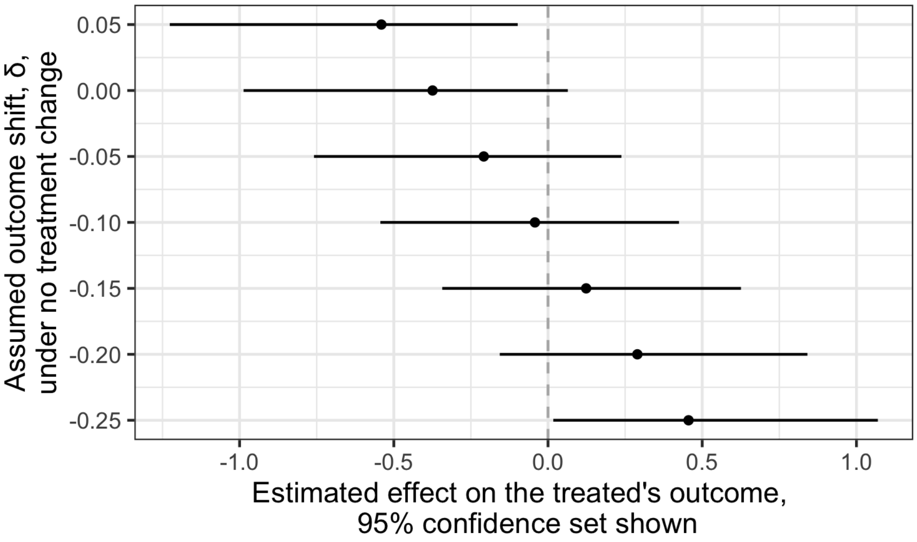
than remdesivir usage, i.e. δ. To select plausible values of δ we initially look to observed baseline covariates, which provide useful insights about baseline risk (we discuss our definition of "baseline covariates," and test this consideration's robustness to an alternate definition, in eAppendix 2).

The first way to do so is to compare covariates at the cohort level. In Table 1, we see that patients in the high-use cohort were younger and less likely to have a history of hyperlipidemia, and within 24 hours of admission they showed more normal lab results and received supplemental oxygen less often and at lower levels, which suggest lower mortality risk.[14,15,16,17] They were more frequently male, however, suggesting higher risk.[18] Notable treatment differences in the first 7 days for this cohort included higher use of dexamethasone and hydroxychloroquine and lower use of ceftriaxone, leronlimab, and tocilizumab. Of these, dexamethasone and tocilizumab are the only treatments shown to lower mortality in randomized trials.[19,20] Several differences between these groups suggest baseline mortality could have been lower in the high-use cohort, most notably the lower use of supplemental oxygen within 24 hours of admission. We may consider how strongly a given covariate could potentially relate to the outcome. For example, if patients on room air had between a 10pp and 30pp lower probability of 28-day mortality compared to those requiring supplemental oxygen, the 9pp increase in such patients in the high-use cohort alone would imply between a 0.9pp and 2.7pp drop in expected mortality.

Considering the collective movement in these covariates between cohorts prior to seeing any results, a domain expert may be able to propose a range of δ they consider plausible. One of the authors, an infectious disease specialist on the care team for many patients in our cohort, judged that these differences may suggest a reduction in baseline mortality in the high-use period relative to the no-use period, conservatively suggesting δ values between -5pp and -15pp.

We also take a more formal approach to reasoning about changes in baseline mortality risk, through modeling at the patient level. Here, we use a flexible, nonlinear, machine learning model, extreme gradient boosting[21] (XGBoost), to generate estimates of $Y(0)$ as a function of our covariates, among the no-use cohort. We then generate these XGBoost risk estimates for every patient in each cohort regardless of treatment status, and compare the average predicted non-treatment risk between the cohorts. This procedure generates estimated average risks of 24.5% in the no-use cohort and 17.8% in the high-use cohort, suggesting that the *observable* differences alone between the cohorts could plausibly contribute -6.7pp to the value of δ. Though we discuss threats to this

Figure 2: Estimated average treatment effect of remdesivir on the 28-day mortality of those patients receiving it. Estimates and confidence sets displayed across values of δ, the counterfactual non-treatment outcome shift between cohorts under no remdesivir use. Plausible δ values are those ranging from +0.05 to -0.15, which support claims of either beneficial or null effects of remdesivir.



modeling strategy in eAppendix 3, the goal here is not to generate a point estimate for δ, but to thoroughly examine the range of values δ could most plausibly take.

As these empirical approaches to inform a δ range only assess observed covariates, we must remain conservative to account for possible cohort differences in unobserved covariates. Combining these considerations, we suggest that δ values ranging from +5pp to -15pp are *plausible* or difficult to rule out. Consequently, we declare any (non-treatment) 28-day mortality in the high-use cohort outside of the 29.5% to 9.5% range to be implausible. We also know that δ could not possibly be any lower than -24.5pp because, considering the outcome rate of 24.5% in the no-use cohort, a more extreme baseline trend would imply a negative counterfactual cumulative mortality incidence.

Effect of remdesivir

Figure 2 shows the ATT estimates that we would obtain under various possible values of δ. Over the range of [0.05, -0.15] we deem plausible, remdesivir had either a beneficial effect (at the $\alpha = 0.05$ level) or no significant effect, but not a harmful one. At the $\delta =$ -0.15 end, the ATT is a non-significant 11.9pp [95% CI: -0.33, 0.60]. At the $\delta = 0.05$ end, remdesivir would prove significantly beneficial, reducing mortality among the treated by 52.0pp [-1.16, -0.09].

The point estimate falls in the beneficial direction if we assume the baseline risk rose, stayed flat, or fell by as much as 11.3 pp ($\delta \geq$ -0.113) in the high-use cohort relative to the no-use cohort, which includes much of the range we deem plausible as well as the value suggested by our modeling exercise above. Because of the relatively small sample size, confidence intervals exclude zero only for very large point estimates. Here, we do not achieve a statistically significant benefit of remdesivir on mortality until we reach a point estimate of a 42pp benefit, which implies the observed 16% mortality among the treated would have been 58% if not for remdesivir. To sustain this result, we would have to assume that baseline risk was 1.9pp higher in the no-use period than in the high-use period ($\delta = +0.019$). This falls within the range we deemed plausible, though we have described reasons to expect a lower, rather than higher, baseline risk in the high-use period is more likely. In short, these considerations imply the point estimate could plausibly fall in the beneficial direction, if not significantly.

We can make a more definitive statement about whether remdesivir had a harmful effect with a confidence interval excluding zero. This would require that the baseline risk shift was -24.5pp, which would imply a non-treatment cumulative mortality incidence of (coincidentally) 0.0 for the high-use cohort. This δ value is clearly indefensible, and nearly impossible. In eAppendix 4 we run sensitivity analyses assessing variations on our analysis, none of which produced substantively differing results.

Comparison to covariate-adjustment with sensitivity analysis

For comparison, we consider a more conventional analytical approach based on covariate adjustment, but augmented with sensitivity analysis. We use a linear probability model for comparable interpretation, regressing mortality on remdesivir use and five covariates that prior research has suggested may be important confounders: age, diabetes, hypertension, and ventilation and transfer to the ICU within 24 hours. The estimated difference in risk associated with remdesivir use is 0.018 [-0.127, 0.164].

To take this estimate at face value would require presuming precisely zero unmeasured confounding, which cannot be credibly argued in this case. Rather than defending this claim, a careful researcher could instead ask whether unmeasured confounding could have been sufficiently strong to change the research conclusion, i.e. that there is no substantive or statistically-detectable effect on mortality. Applying the sensitivity analysis of Cinelli and Hazlett [6], we find that estimate could have actually been a significant, beneficial one if unobserved confounders explain (for example) 23% of the residual variance in both remdesivir use and mortality. Likewise, the estimate would have been significant and harmful if omitted confounding acting in the opposite direction explains 20% of the residual variation in both the treatment and the outcome. Given that we know little about determinants of COVID-19 mortality and how treatment decisions were made, it is not clear that we could rule out that remdesivir was beneficial, harmful, or neither. We discuss how this and other results compare to those of SCQE below.

Discussion

Using the SCQE methodology we demonstrated that, over a reasonable range of assumptions, we observe null or beneficial effects of remdesivir use on 28-day mortality in treated patients hospitalized for COVID-19. Further, we rule out that remdesivir could have caused a statistically significant increase in 28-day mortality among treated patients in this sample.

Comparing what we can credibly conclude from SCQE to what we could credibly conclude using covariate-adjustment approaches points to three ways in which SCQE can complement existing approaches. First, while the range of credible estimates under SCQE remains wide, it still tells us that (i) either null or beneficial effects are plausible, and (ii) a (statistically significant) harmful effect is implausible. By comparison, examining the sensitivity of covariate adjusted results to unmeasured confounding offered no means to confidently rule-out confounders strong enough to imply that the true effect had been harmful. Of course, SCQE may produce either more or less informative results in other applications, and should be viewed as an alternative approach worth considering rather than as a superior method in every setting.

The second difference is in our ability to understand and reason about the assumptions required by each approach. To consider the risks of unobserved confounding under covariate-adjustment techniques requires us to reason in terms of a chosen method's paramaterazation of "the strength of the confounding-treatment and confounding-outcome relationships." Whatever the parameterization, we expect that practitioners can better understand questions about δ, "the change we would see in the outcome over time, absent changes in treatment."

The third difference is more behavioral than statistical, but is perhaps the most important for reducing the risk of learning the wrong lessons from an analysis. Covariate-adjustment approaches are usually used to present a single point estimate, and readers will often use that estimate to guide their decision-making, regardless of caveats or sensitivity analyses even when they are provided. A benefit of presenting results as we do in Figure 2 or with arguments about "what you need to assume to reach a given conclusion" is that they provide no particular focal point or default estimate that may be (mis)understood as a "best guess." While this will surely be frustrating for many readers, it is also a safety feature, forcing us to confront rather than conceal the limits of what can be concluded while still learning what we can.

New treatments attempted during the COVID-19 pandemic offer a "tough case" for SCQE, because mortality risk could vary substantially over time owing to changes in the disease, the population arriving at the hospitals in question, and other treatments being attempted. This in turn leads to far wider credible ranges on the baseline trend, δ, than what we would expect for longer-running diseases that see fewer or slower changes in the standards of care and stable average outcomes over months or years, and thus represent ideal applications for SCQE. Still, this setting is an even tougher case for bounding the strength of unmeasured confounders in a covariate-adjusted estimate, as evidenced by the narrower conclusions SCQE allows us to defend by utilizing the rapid changes in remdesivir usage we do see. There are many settings in which the opposite is true and SCQE will offer less precise insights; a cautious assessment of unmeasured confounding may provide the most narrow defensible range of plausible effects. Different scenarios lend themselves to reasoning about the credibility of different parameters, and thus consideration of both approaches will lead practitioners to narrower believable conclusions than if they only ever thought to use one or the other.

While RCT results for remdesivir have limited comparability to those from SCQE (due to the differing populations for whom they estimate treatment effects), we note that our results do not contradict the most recent RCTs and recommendations from the WHO, though we do suggest a larger possibility of significant benefit.[13]

A further practical benefit of SCQE is that it can be employed without access to individual level data. In applications where the treatment, time period, and outcome are all binary, as they are here, SCQE requires only a set of summary statistics: 8 counts (4 subgroup sample sizes and the number of deaths within each) and a δ range. These numbers are sufficient to produce estimates and confidence sets like those seen here. This enable readers, reviewers, and other researchers to plausibly apply the SCQE approach to existing studies based on only summary statistics that may be available in print, or could easily be shared without raising privacy concerns (a web application for making these computations is available at [removed for blinded review]).

When RCTs are available, they provide better information than observational studies of any sort, assuming both study the same treatments and sufficiently similar populations. Still, observational studies will continue to be part of the research landscape, and would ideally provide valuable information before RCTs are completed, where randomization is not an option, or over populations different from those eligible for RCTs. To achieve this positive use of observational data, however, requires tools that can extract credible conclusions, without generating over-confidence in claims that may be wrong or dangerous. SCQE is one approach to achieving this, in the context of newly attempted treatments, by asking practitioners and readers to consider a range of effect estimates exactly as narrow as can be warranted by the assumptions they are able to defend. This and other partial identification approaches will sometimes—perhaps often—illustrate the wide uncertainty in what we can conclude. However, particularly where harm may be done, revealing such uncertainty and coping with the limitations of our knowledge is far preferable to strategies that risk over-confidence in seemingly precise but indefensible estimates.

References

[1] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[2] Joshua D Angrist and Jörn-Steffen Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2008.

[3] Judea Pearl. *Causality*. Cambridge university press, 2009.

[4] Paul R Rosenbaum and Donald B Rubin. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society: Series B (Methodological)*, 45(2):212–218, 1983.

[5] Tyler J VanderWeele and Onyebuchi A Arah. Unmeasured confounding for general outcomes, treatments, and confounders: bias formulas for sensitivity analysis. *Epidemiology (Cambridge, Mass.)*, 22(1):42, 2011.

[6] Carlos Cinelli and Chad Hazlett. Making sense of sensitivity: Extending omitted variable bias. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(1):39–67, 2020.

[7] Chad Hazlett, Werner Maokola, and David Ami Wulf. Inference without randomization or ignorability: A stability controlled quasi-experiment on the prevention of tuberculosis. *Statistics in Medicine*, 39:4169–4186, 2020. doi: 10.1002/sim.8717.

[8] Jerzy S Neyman. On the application of probability theory to agricultural experiments. essay on principles. section 9. translated and edited by dm dabrowska and tp speed, statistical science (1990), 5, 465-480. *Annals of Agricultural Sciences*, 10:1–51, 1923.

[9] Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

[10] Jonathan Grein, Norio Ohmagari, Daniel Shin, George Diaz, Erika Asperges, Antonella Castagna, Torsten Feldt, Gary Green, Margaret L Green, François-Xavier Lescure, et al. Com-

passionate use of remdesivir for patients with severe covid-19. *New England Journal of Medicine*, 382(24):2327–2336, 2020.

[11] John H Beigel, Kay M Tomashek, Lori E Dodd, Aneesh K Mehta, Barry S Zingman, Andre C Kalil, Elizabeth Hohmann, Helen Y Chu, Annie Luetkemeyer, Susan Kline, et al. Remdesivir for the treatment of covid-19. *New England Journal of Medicine*, 383(19):1813–1826, 2020.

[12] FDA Press Release. *FDA Approves First Treatment for COVID-19.* [Press Release] www.fda.gov/news-events/press-announcements/fda-approves-first-treatment-covid-19, 2020.

[13] WHO Living Guideline. Therapeutics and covid-19: living guideline, 20 november 2020. World health organization, 2020.

[14] CDC COVID-19 Response Team, Stephanie Bialek, Ellen Boundy, Virginia Bowen, Nancy Chow, Amanda Cohn, Nicole Dowling, Sascha Ellington, et al. Severe outcomes among patients with coronavirus disease 2019 (covid-19)—united states, february 12–march 16, 2020. *Morbidity and mortality weekly report*, 69(12):343–346, 2020.

[15] Nanshan Chen, Min Zhou, Xuan Dong, Jieming Qu, Fengyun Gong, Yang Han, Yang Qiu, Jingli Wang, Ying Liu, Yuan Wei, et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in wuhan, china: a descriptive study. *The lancet*, 395 (10223):507–513, 2020.

[16] Chaolin Huang, Yeming Wang, Xingwang Li, Lili Ren, Jianping Zhao, Yi Hu, Li Zhang, Guohui Fan, Jiuyang Xu, Xiaoying Gu, et al. Clinical features of patients infected with 2019 novel coronavirus in wuhan, china. *The lancet*, 395(10223):497–506, 2020.

[17] Dawei Wang, Bo Hu, Chang Hu, Fangfang Zhu, Xing Liu, Jing Zhang, Binbin Wang, Hui Xiang, Zhenshun Cheng, Yong Xiong, et al. Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus–infected pneumonia in wuhan, china. *Jama*, 323(11):1061–1069, 2020.

[18] Hannah Peckham, Nina M de Gruijter, Charles Raine, Anna Radziszewska, Coziana Ciurtin, Lucy R Wedderburn, Elizabeth C Rosser, Kate Webb, and Claire T Deakin. Male sex iden-

tified by global covid-19 meta-analysis as a risk factor for death and itu admission. *Nature communications*, 11(1):1–10, 2020.

[19] RECOVERY Collaborative Group. Dexamethasone in hospitalized patients with covid-19. *New England Journal of Medicine*, 384(8):693–704, 2021.

[20] RECOVERY Collaborative Group. Tocilizumab in patients admitted to hospital with covid-19 (recovery): a randomised, controlled, open-label, platform trial. *The Lancet*, 397(10285): 1637–1645, 2021.

[21] Tianqi Chen and Carlos Guestrin. Xgboost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, 2016. doi: 10.1145/2939672.2939785. URL `http://dx.doi.org/10.1145/2939672.2939785`.

eAppendices

1 Robust confidence sets for SCQE

In Hazlett et al.[1], the connection between SCQE and IV is formalized through the creation of a pseudo-outcome $\tilde{Y} = Y - \delta Z$. This $\tilde{Y}$ represents the outcome after adjustment for some cohort-to-cohort shift δ, effectively removing differences between cohorts other than those caused by the treatment difference of interest. For the correct value of δ, implementing an IV approach with $\tilde{Y}$ rather than $\tilde{Y}$ as the outcome guarantees that the exogeneity assumption holds and lets us borrow inferential tools from the IV literature. Although we cannot know the correct value of δ, we express our estimates conditionally on δ, allowing for valid inference within the partial identification framework.

In the paper, the confidence sets we use are Anderson-Rubin confidence sets,[2] which have correct coverage even in the face of weak instruments (in the SCQE case we face a weak instrument when the change in treatment usage between cohorts is too small). Implementation with code in the SCQE context is a two-step process: For a given value of δ we adjust the outcome data to form $\tilde{Y}$, and then we construct the confidence sets conditional on that δ value using a standard IV library like ivmodel in R.[3] When SCQE can be implemented without unit-level data (as noted in the discussion section), Anderson-Rubin confidence sets can also be implemented.[4] Interpretation of these confidence sets in the presence of weak instruments is provided elsewhere.[5]

2 "Baseline" covariate inclusion consideration

When assessing possible values of δ by investigating covariate differences between cohorts, we must choose inclusion criteria for those covariates. In the text, by "baseline" we mean covariates that are strictly or plausibly "pre-treatment", i.e. those that cannot likely be affected by treatment itself. The demographic and comorbidity factors we consider in Table 1 are strictly pre-treatment. Measures taken in the first 24 hours could in principle occur after a decision to give remdesivir, but we do not anticipate remdesivir could have had any effect within this time, and thus these measures are plausibly pre-treatment. Treatments given within 7 days, however, are more likely to be affected by remdesivir use. The risks of considering (and modeling) post-treatment variables when informing plausible choices of δ are similar to the risks involved in conditioning on post-treatment variables in standard covariate adjustment techniques. For example, if a given covariate (e.g. dexamethasone use) is affected by remdesivir use, using cohort-to-cohort differences in its presence to inform δ will induce bias, as those differences may represent indirect effect pathways (dexamethasone use could be a mediator of remdesivir's effect on mortality) rather than violations of the exogeneity assumption. Additional related introductions and sources of bias are discussed in Wulf[4].

In order to test the sensitivity of this δ-evaluation procedure to our definition of "baseline" (and the biases it could produce), we reran the analysis while limiting alternate treatments to those given within 48 hours. We reiterate that while such covariate-driven analysis is informative, it is not precise enough to trust a narrow δ range, and any range it produces should only be treated as one suggestion to consider. Still, if the results of this reanalysis were to produce meaningfully different results, a reasonable response would be to widen the range of δ values we consider plausible. In fact, table 1 was substantively unchanged after this change, with only slightly smaller cohort-to-cohort differences in dexamethasone, hydroxychloroquine, and leronlimab use, and a slightly larger difference in tocilizumab use. Correspondingly, the XGBoost modeling procedure suggested a δ value of -6.8pp, nearly identical to the value of -6.7pp suggested in the original analysis.

3 Modeling δ suggestions – threats to validity

In the text, we utilize a modeling procedure to suggest how changes in only pre-treatment observables might plausibly contribute to δ. The procedure models the outcome as a function of baseline covariates among those patients in the no-use cohort. The model is then applied to all patients, and the average predicted non-treatment outcomes are compared between cohorts in order to estimate the baseline observable-driven difference to be addressed using δ. This procedure faces the same issues discussed in eAppendix 2 – post-treatment covariates that are included in the model may introduce bias. It makes an additional assumption, however: we assume that controlling for covariates renders non-treatment outcomes independent of cohort membership, $Y(0) \perp\!\!\!\perp Z \mid X$, which implies $\mathbb{E}[Y(0)|X,Z] = \mathbb{E}[Y(0)|X]$. When this holds, we can establish the relationship between the covariates and non-treatment outcomes in the no-use cohort, apply that learned relationship in the high-use cohort, and interpret the difference in mean $Y(0)$ between cohorts as the contribution of observables to δ. The interplay between this assumption and the pre-treatment restriction from eAppendix 2 is discussed in Wulf[4]. Note that this assumption is likely far more defensible than the standard conditional ignorability assumption, which can be written $Y(0) \perp\!\!\!\perp D \mid X$.

4 Sensitivity analyses

In this appendix, we provide the results of several sensitivity analyses, each of which take the form of an adjustment to the choices we made in the original analysis. Finding minimal differences between the conclusions they produce and those from the original analysis would not guarantee those results are correct, only that they are reasonably robust to differences in those particular choices.

Table 2 displays the patient counts, outcome and treatment proportions, important baseline characteristics, covariate modeling results, SCQE results, and inferential cutoffs across six such changes. Across each of these adjustments, the ATTs corresponding to a plausible range of δ values ranged from beneficial to not meaningfully different from 0, as in the original analysis. More importantly, the δ values required to argue for a harmful effect of remdesivir at the α=0.05 level remained implausible, and in some cases mathematically impossible, continuing to rule out the remdesivir could have had a statistically significant harmful effect among those who took it in this sample.

The first change re-defines the outcome as 14-day mortality rather than 28-day mortality. Although the latter was the more commonly assessed cutoff in randomized trials of remdesivir, some used 14-day mortality as well.[6]. This change lowers the baseline mortality in the no-use cohort, narrows the resultant δ range, but supports the same conclusions – that plausible δ values suggest either beneficial or null ATT estimates, and that defending claims of a harmful ATT estimate is impossible.

The second change adjusted the treatment definition to be remdesivir use within 3 days rather than at any time. Only two patients began remdesivir treatment after this cutoff, and they were excluded from this adjusted analysis. In order to align the data-informed δ procedures with this change, we also limited the considered covariates to those measured within 2 days rather than 7 days (as we did in eAppendix 2). This treatment definition change similarly maintained the same conclusions.

The third adjustment included transfer patients, those patients whose medical record indicated their admission source was a different acute care facility. We excluded these patients because we could not rule out the possibility of remdesivir treatment prior to admission. Still, the (likely) substantial differences between transfer and non-transfer patients could mean their exclusion hides a subset of patients for whom remdesiver is far more or less beneficial. After including transfer patients, the larger sample had a higher baseline cumulative mortality incidence (requiring a wider δ range) and the δ modeling procedure strongly suggested negative values. Still, the resultant range included null and beneficial ATT estimates, and harmful ATT estimates were highly implausible (though not mathematically impossible).

The fourth adjustment was to not truncate the later high-use period (which was done originally to avoid an influx of patients treated with dexamethasone, which would increase the cohort imbalances), allowing that period to extend to August 25th rather than stopping at July 4th. The increased imbalance in dexamethasone usage between cohorts (0% in the no-use period and 22% in the high-use period, compared to only 8% in the primary analysis' high-use period) suggests that we carefully account for its potential effects when assessing the range of δ values to consider plausible. In addition to the larger dexamethasone difference, a lower rate of ventilator use and ICU admission within 24 hours of admission in the high-use period suggest a lower δ value. Conversely, this change also lowered the usages of hydroxychloroquine, heparin, azythromycin, and tocilizumab in the high-use cohorts, which, if any of those were helpful, would suggest a higher δ value. When considered together and alongside the data-driven model estimates of δ, a widened and lower δ range is reasonable. This range, like each of those previously considered, suggests a null or beneficial effect of remdesivir and rules out a harmful effect.

The final two changes investigate the impact of excluding either the earlier or the later high-use cohorts. Larger differences in baseline characteristics between cohorts in each of these analysis (alongside smaller sample sizes) require widened δ plausibility ranges. When excluding the earlier high-use cohort, we observe even larger differences in supplemental oxygen use, dexamethasone use, and ICU admission within 24 hours, and a smaller difference in ventilation within 24 hours. Given the direction of these differences, this change implies lower δ values, even with a difference in tocilizumab use possibly suggesting higher δ values. A reasonable δ range taking these changes into account suggest, once again, a null or beneficial effect of remdesivir.

Finally, when excluding the later high-use cohort, several changes to the cohort-to-cohort comparison are clear. Most suggest comparatively higher risk in this high-use cohort than in the original analysis' high-use cohort (more male patients, more early transfers to the ICU and ventilation, higher supplemental oxygen use, less dexamethasone use), though some were either unclear or suggested lower risk (higher tocilizumab, hydroxychloroquine, and leronlimab use, as well as more white patients). Overall, a higher and wider δ range is warranted, which maintains the same conclusions reached in the prior adjustments and the original analysis.

Table 2: Sensitivity analyses results

	Original	$Y_{\max}$ 14d	$D_{\max}$ 3d	w/transfers	no date trim	no earlier high-use	no later high-use
N no-use	53	53	53	65	53	53	53
N high-use	83	83	81	95	197	37	48
$\overline{Y}$ no-use	0.245	0.189	0.245	0.277	0.245	0.245	0.245
$\overline{D}$ high-use	0.313	0.313	0.296	0.305	0.391	0.514	0.167
Notable changes to baseline cohort comparisons (in %)	reference	identical	identical	very minor changes	latinx 32 vs 42 ICU24hr 34 vs 24 vent24hr 13 vs 10 hydroxy 13 vs 10 dexameth 0 vs 22 heparin 47 vs 37 azythro 57 vs 45 tociliz 11 vs 3	O_2 miss 19 vs 41 ICU24hr 34 vs 19 vent24hr 13 vs 11 dexameth 0 vs 14 tociliz 11 vs 0	male 51 vs 67 white 34 vs 50 latinx 32 vs 23 O_2 miss 19 vs 19 ICU24hr 34 vs 35 vent24hr 13 vs 25 hydroxy 13 vs 38 dexameth 0 vs 4 leronli 25 vs 2 tociliz 11 vs 10
XGBoost δ estimate	-0.049, [-.131,.032]	-0.014, [-.080,.051]	-0.008, [-.074,.058]	-0.080, [-.156,-.005]	-0.105, [-.175,-.035]	-0.090, [-.177,-.003]	-0.045, [-.134,.043]
Plausible δ range[a]	[-.15, .05]	[-.125,.075]	[-.15, .075]	[-.2, .025]	[-.2, .05]	[-.2, .05]	[-.15, .1]
Min beneficial δ[b]	0.019	0.025	0.024	-0.028	-0.068	0.029	0.058
Max harmful δ[c]	-0.245	-0.21*	-0.243	-0.273	-0.26*	-0.303*	-0.257*

[a] across all adjustments, these ranges produced ATTs ranging from beneficial (at the $\alpha=.05$ level) to null. [b] δ values greater than these produce ATTs that are beneficial at the $\alpha=.05$ level. [c] δ values less than these produce ATTs that are harmful at the $\alpha=.05$ level, and those with asterisks are algebraically impossible given the outcome rate in the no-use period.

Appendix Bibliography

[1] Chad Hazlett, Werner Maokola, and David Ami Wulf. Inference without randomization or ignorability: A stability controlled quasi-experiment on the prevention of tuberculosis. *Statistics in Medicine*, 39:4169–4186, 2020. doi: 10.1002/sim.8717.

[2] Theodore W Anderson, Herman Rubin, et al. Estimation of the parameters of a single equation in a complete system of stochastic equations. *Annals of Mathematical statistics*, 20(1):46–63, 1949.

[3] Hyunseung Kang, Yang Jiang, Qingyuan Zhao, and Dylan Small. *ivmodel: Statistical Inference and Sensitivity Analysis for Instrumental Variables Model*, 2021. URL `https://CRAN.R-project.org/package=ivmodel`. R package version 1.9.0.

[4] David A Wulf. *Causal Inference Outside of Randomized Trials with the Stability-Controlled Quasi-Experiment: Extensions and Considerations.* PhD thesis, UCLA, June 2021. URL `https://escholarship.org/uc/item/4ms7f0sq`. ProQuest ID: Wulf_ucla_0031D_19932. Merritt ID: ark:/13030/m5z95dsb.

[5] Isaiah Andrews, James H Stock, and Liyang Sun. Weak instruments in instrumental variables regression: Theory and practice. *Annual Review of Economics*, 11:727–753, 2019.

[6] WHO Living Guideline. Therapeutics and covid-19: living guideline, 20 november 2020. World health organization, 2020.