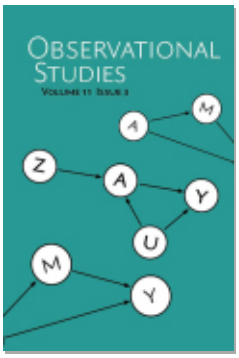
Safe inference outside of randomized trials: Application of the stability-controlled quasi-experiment to the effects of three COVID-19 therapies



David Ami Wulf, Chad Hazlett, Brian L. Hill, Jeffrey N. Chiang, David Goodman-Meza, Bogdan Pasaniuc, Onyebuchi A. Arah, Kristine M. Erlandson, Brian T. Montague

Observational Studies, Volume 11, Issue 3, 2025, pp. 301-330 (Article)

Published by University of Pennsylvania Press
DOI: *https://doi.org/10.1353/obs.2025.a973069*

Observational Studies 11.3 (2025): 301-330 Accepted Jul 09, 2025; Published Nov 2025

Safe inference outside of randomized trials: Application of the stability-controlled quasi-experiment to the effects of three COVID-19 therapies

David Ami Wulf* **amiwulf@ucla.edu**
Department of Statistics & Data Science
University of California, Los Angeles
Los Angeles, CA, USA

Chad Hazlett* **chazlett@ucla.edu**
Department of Statistics & Data Science
Department of Political Science
University of California, Los Angeles
Los Angeles, CA, USA

Brian L. Hill **blhill@ucla.edu**
Department of Computer Science
University of California, Los Angeles
Los Angeles, CA, USA

Jeffrey N. Chiang **njchiang@ucla.edu**
Department of Computational Medicine, David Geffen School of Medicine
University of California, Los Angeles
Los Angeles, CA, USA

David Goodman-Meza **dgoodman@mednet.ucla.edu**
Division of Infectious Diseases, David Geffen School of Medicine
University of California, Los Angeles
Los Angeles, CA, USA

Bogdan Pasaniuc **pasaniuc@ucla.edu**
Department of Computational Medicine, David Geffen School of Medicine
Department of Pathology and Laboratory Medicine, David Geffen School of Medicine
Department of Human Genetics, David Geffen School of Medicine
University of California, Los Angeles
Los Angeles, CA, USA

Onyebuchi A. Arah **arah@ucla.edu**
Department of Statistics & Data Science
Department of Epidemiology, Fielding School of Public Health
University of California, Los Angeles
Los Angeles, CA, USA
Research Unit for Epidemiology, Department of Public Health
Aarhus University
Aarhus, Denmark

Kristine M. Erlandson† **kristine.erlandson@cuanschutz.edu**
Department of Medicine, Division of Infectious Diseases
University of Colorado Anschutz Medical Campus
Aurora, CO, USA

*co-first authors, †co-senior authors

Brian T. Montague† **brian.montague@cuanschutz.edu**
Department of Medicine, Division of Infectious Diseases
University of Colorado Anschutz Medical Campus
Aurora, CO, USA

Abstract

When estimating the effects of medical therapies from their use outside of randomized trials, researchers often rely on assumptions that are difficult to justify and impossible to verify. The resulting estimates may thus be far from their intended causal targets, potentially making a harmful treatment appear beneficial or vice versa. We review the stability-controlled quasi-experiment (SCQE), a method suited to settings where a treatment's prevalence changes sharply over a short period, and apply it to assess the effects of remdesivir, hydroxychloroquine, and dexamethasone on COVID-19 mortality. Rather than requiring debate about the absence (or limited strength) of unobserved confounding, about "parallel trends", or other well-known strategies, the SCQE asks users to reason about a "baseline trend" assumption. In this setting, this asks "How much could COVID-19 mortality have changed over a short period, absent the treatment change in question?" Any plausible range for this assumption yields a corresponding range of plausible causal effect estimates. Conversely, SCQE clarifies what baseline trends must be defended or refuted in order to defend or refute a given conclusion about a treatment's efficacy or harm. Using data from two hospital systems early in the COVID-19 pandemic, we show that SCQE could have enabled safe yet partially informative inferences about treatment effects before clinical trial completion, producing conclusions consistent with the results of eventual randomized trials.

Keywords: Observational studies, Real-World evidence, Partial identification, Stability-Controlled quasi-experiment, Sensitivity analysis, COVID-19

1. Introduction

During the early response to COVID-19, many patients received emergent therapies before their safety and efficacy were established, and often outside randomized trials. Epidemiologists and other health researchers rightly warned against drawing conclusions from such use, given the serious risk of confounding. Yet dismissing all information outside randomized trials risks overlooking potentially life-saving insights—while physicians and patients were forced to make difficult decisions with whatever evidence was available. How, then, can observational data be used to identify a range of plausible effects for new therapies without fostering overconfidence in potentially misleading results? We offer one answer here and illustrate how it could have yielded defensible, informative findings in the early days of the pandemic.

A common approach for making claims about the effects of therapies outside of randomized trials compares outcomes between treated and control groups, after adjusting for observable differences between these groups. Regardless of the modeling, weighting, or matching technology used to make these adjustments, for the resulting group difference to reflect only the causal effect of the treatment requires ruling out unobserved differences between the groups (unobserved confounders). This is rarely a defensible assumption, and

to the contrary, is easily violated when clinicians and patients make treatment decisions in part informed by factors not recorded in the data. The resulting bias is capable of making a beneficial treatment look harmful, or a harmful one look beneficial. Given this, and the inability to detect such biases, one may reasonably worry that estimates produced by this strategy risk being more misleading than informative.

In recognition of this danger, one approach is to begin with estimates under the no-unobserved-confounding assumption but then conduct sensitivity analyses to examine how results would vary in the presence of unobserved confounders of postulated strengths (e.g. Lin et al., 1998; Cinelli and Hazlett, 2020; D'Agostino McGowan, 2022). Another strategy is to recognize the difficulties of the no-unobserved-confounding assumption, and seek to place assumptions on alternative quantities that can yield (partial or full) identification. For instance, researchers may opt to adopt a parallel trends assumption in the context of difference-in-differences (DID), make assumptions on the relations of confounding to placebo treatment or outcome variables (as in proximal causal inference), or posit that an observed variable influences the probability of treatment without affecting or predicting outcomes, as is the case with instrumental variables (IV). These assumptions can also be relaxed by suitable sensitivity analyses, or constructed as partial identification strategies that bound estimates given argued bounds on assumed quantities. Across these options, including our chosen approach, the effectiveness of any method in a given context depends on two key factors: (i) whether the context provides the necessary data structure and conditions for implementation (e.g., multiple time periods and the required group or panel structure for DID), and (ii) whether the required assumptions align with those the investigator can plausibly evaluate or bound based on the setting and their expertise. This "right tool for the job" perspective encourages the development of new strategies based on assumptions that researchers are best positioned to assess in specific scenarios.

In this spirit, we review a relatively new approach known as the "stability controlled quasi-experiment" or SCQE (Hazlett, 2019; Hazlett et al., 2020; Wulf, 2021). This framework is designed for contexts where a swift transition occurs in the proportion of units (patients) given a treatment, though we note other use cases below as well. The key unknown to reason about in the SCQE is the "baseline trend": the (unobservable) change in average outcomes we would have seen between successive cohorts of admitted patients, had no such treatment usage change occurred. Any postulated value of the baseline trend, combined with the observed data, implies a treatment effect. Thus, the approach always conveys the relationship between values of the baseline trend and the treatment effect, and is increasingly informative about the treatment effect to the degree that investigators can defensibly bound the baseline trend.

In this paper we use the SCQE framework to analyze the effects of three COVID-19 therapies—dexamethasone, hydroxychloroquine, and remdesivir—used early in the pandemic, prior to the completion of randomized controlled trials (RCTs). Each of these treatments saw sudden increases or decreases in usage between cohorts we define. The baseline trend assumption to reason about is then how much COVID-19 mortality rates among hospitalized patients plausibly could have changed between these subsequent cohorts, had this sudden change in usage not occurred. That is, how might the average outcome have changed between cohorts, for reasons other then the shift in treatment usage we are studying? This is unknown, and users are not expected to defend precise quantities of this unobservable

value. Rather, any range of plausible values they suggest points to a defensible range of causal effect estimates, and values they deem implausible imply estimates they can rule out. These plausibility judgments can be aided (but not replaced) by ancillary data, such as differences between the two cohorts in disease severity, cohort composition, hospital crowding, usage of other therapies, or other treatment practices. Like other partial identification approaches, SCQE presents its conclusions as a range of estimates rather than centering on a single point claim. This guards against overconfidence in results that depend on narrow or indefensible assumptions. We further show that DID and IV can be viewed as special cases of SCQE: while not always applicable, they correspond to specific assumed values of the baseline trend. In this way, SCQE both generalizes these approaches and provides a framework for sensitivity analysis in terms of their implied baseline trends.

Applying the SCQE to data from two hospital systems over several months early in the pandemic, we find that remdesivir and dexamethasone could have plausibly had a beneficial effect, and that it was nearly impossible that they were harmful to an extent that would produce statistically significant results in this study. By contrast, under a range of baseline-trend assumptions we argue to be plausible, hydroxychloroquine would have had a harmful or non-significant effect, while it would be very difficult to strongly believe the assumptions that imply it had benefit. This is consistent with the results of eventual randomized trials (see e.g. Lamontagne et al., 2023; COVID-19 Treatment Guidelines Panel, 2023). For patients and physicians forced to make treatment decisions before RCTs on these therapies were complete, our analysis would have provided informative guidance, without resting on indefensible assumptions that can produce misleading results. Beyond this specific case, such an approach may be useful in other contexts where trial results are not yet available, but also in cases where no randomized trial is likely to be conducted or when examining real-world effectiveness after drug approval.

2. Background: The stability-controlled quasi-experiment (SCQE)

We first describe the SCQE and the assumption on which it rests. Imagine two consecutive cohorts of hospitalized patients, with none of the earlier patients receiving the treatment of interest and a non-random 50% of the later patients being treated. Suppose also that in the earlier ("low-use") cohort, 20% of patients died within 28 days of admission, while in the later ("high-use") cohort, 15% died. 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. This gives an intuition for the approach, which we refine below.

More formally, for a sample 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 (Neyman, 1923; Rubin, 1974), $Y_i(0)$ 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 in-

terest is the average treatment effect among treated patients (ATT) in the second cohort, $\mathbb{E}[Y(1)|Z=1, D=1] - \mathbb{E}[Y(0)|Z=1, D=1]$.

Rather than assuming no cohort-to-cohort baseline risk differences as above, SCQE allows average expected outcomes under non-treatment to differ between the cohorts by a prescribed amount, i.e. $\mathbb{E}[Y(0)|Z = 1] - \mathbb{E}[Y(0)|Z = 0]$. We call this the "baseline trend", because it represents what the between-cohort change in mean outcome would be, absent (additional adoption of) treatment. The key result driving SCQE is that for any given choice of this baseline trend (which we label δ), the data tell us the logically implied treatment effect estimate.

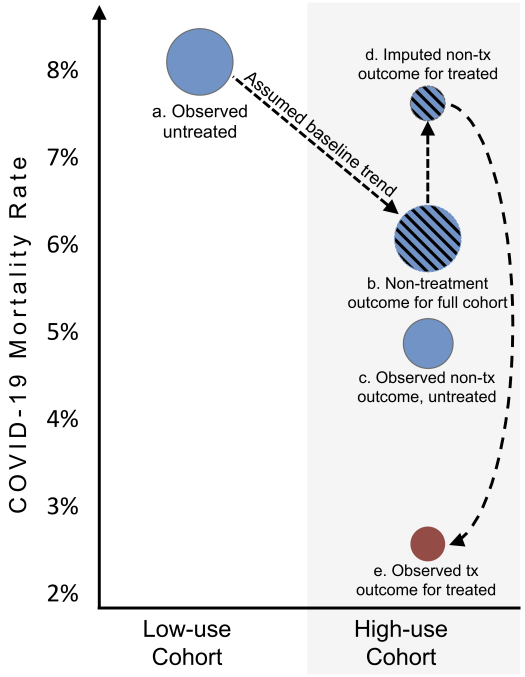


Figure 1: **Graphical explanation of SCQE.** Each circle represents a group, with height representing that group's average hypothetical outcome (mortality rate). In the low-use/no-use cohort (left), suppose we observe average mortality (8%) under non-treatment. We then postulate (as an example) an assumption regarding how the non-treatment outcome would have changed from one cohort to the next. Here this is postulated to be a 2pp drop, putting the average non-treatment outcome for the high-use cohort at 6% (b). The value of (b) is a weighted sum of the average non-treatment outcomes for those who were not treated (c) and those who were treated. We can thus solve algebraically for the average non-treatment outcome that would have been experienced by the treated (d). Comparing the observed average outcome for the treated (e) to this imputed average non-treatment outcome for the treated (d) produces the average treatment effect for the treated. No assumption regarding comparability of the treated and control (c and e) is made, only an assumption on the trend in the average non-treatment outcome.

We offer an in-depth intuition for this result in Figure 1. For a brief mathematical explanation, following Hazlett (2019), we start with the definition of δ above and note that the average non-treatment outcome over everybody in the second cohort, while unobserved, can be broken up according to the law of iterated expectations:

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

where $\pi_1 = P(D{=}1|Z{=}1)$ is the treatment rate in the second cohort. We can solve for the average non-treatment outcome among the treated in the second cohort,

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

which represents how these treated patients in the second cohort would have done had they not been treated. Subtracting this counterfactual from these patients' observed (treated) outcomes gives us the ATT,

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

We employ the analog estimator for 1 across postulated values of δ, replacing expectations and proportions with their analogous sample means. Confidence intervals can be constructed for each such estimate, quantifying statistical uncertainty conditional on that δ value (see Appendix A for details).

This 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. Thus, while we avoided assumptions that compare treated and untreated patients (and thus the risks from unmeasured confounding), instead assuming degrees by which entire cohorts may differ, we still arrived at a targeted, treatment-specific effect estimate. As Equation 1 shows, this effect is identified for any assumed value of the baseline trend, δ.

2.1 Average treatment effect for whom?

We make here three remarks on the estimand. First, our focus on the ATT arises due to the nature of the identification procedure described above. Whether the ATT is more or less desirable than the ATE or other usual quantity depends on the context and inferential aims. The ATT is most appropriate when we wish to know "what was the effect of the treatment in those who received it, or in others like them who might receive it in the future?" In cases where the treatment could be expanded to new populations in the future, we must be concerned with how that population might differ from the treated population examined.

Second, one concern with randomized trials is that the eligible population for those trials is often a highly restricted one, differing widely from the population that would likely try the treatment in question if it is approved or made more widely available. This is one important motivation for seeking "real world evidence" for the effectiveness of therapies even when trial results are already available. SCQE provides another tool in the observational research toolkit to aid in producing credible real world evidence of this kind.

Finally, on a more technical point, in the exposition above the "low-use" cohort is actually a "no-use" cohort, meaning there is one-way non-compliance (those in the first cohort cannot take the treatment). This convenience leads to identification of the ATT under postulated shifts in $\mathbb{E}[Y(0)]$ between cohorts. However, in some settings there may

be some treated units in the low-use cohort. In that case, the estimand is an average effect over "compliers": those units who would have been treated if they appeared in the high-use cohort but would not be if they appeared in the low-use cohort. This complier average treatment effect (CATE) will be a familiar one from the perspective of instrumental variable analysis, which we discuss below. The δ to reason about in that setting is not the expected shift in outcomes absent treatment, but rather the shift in outcomes had the change in treatment usage not occurred. In two of the applications below, 3%-6% of patients in the low-use cohort are treated. For simplicity of interpretation and discussion, rather than switching between the ATT and CATE, we use the ATT throughout. This makes little practical difference, since such a small proportion (3%-6%) of the population would be always-takers rather than compliers; even if these units experienced the maximally different reaction to treatment, it can introduce only a very small difference between the CATE and ATT. However, one can use the SCQE in cases where there is more substantial two-sided non-compliance, e.g. if 20% of units in the low-use period took treatment, compared to a 50% rate in the high-use period. In those cases, it would be important to interpret the result as the complier average effect and to understand δ as the change in outcomes had the treatment assignment strategy remained constant (see Wulf, 2021 for further details).

2.2 Using SCQE for inference under unknown δ

We recommend two modes of argumentation investigators can employ using this mapping between δ (baseline trend) values and the logically implied ATT estimates in a sample. The first approach begins with an *ex ante* claim about the plausible limits of δ, asking users to declare (and explain) their proposed range of plausible δ values. SCQE combines this with the observed data to produce the consequent range of plausible ATT estimates, alongside statistical uncertainty around each such δ-conditional estimate. The "SCQE plot" used below displays these conditional estimates and confidence intervals across different δ values. The range of δ deemed plausible may be informed by domain knowledge, relevant data, and even natural limits on the trend (e.g. mortality rates must fall between 0 and 1), as we discuss in the context of our specific cases below. The premise is not that the user could or should defend a particular value of δ, but rather that they must accept as possible the entire range of estimates implied by the range of δ they cannot rule out. In this interval approach, no single focal estimate is produced, avoiding overconfidence based on an indefensible point assumption.

The second, *ex post*, approach inverts this. Since the SCQE plot shows ATT estimate corresponding to any postulated value of δ, we can then look at the threshold δ values required to reach any given conclusion about the sign of the point estimate or a statistically significant result of either sign. This allows the researcher—and reader—to ask whether they are willing and able to defend the values of δ required to defend a given research conclusion. This approach can be applied whether or not one has made, or believes, an *ex ante* plausibility range for δ. For example, consider regions of the SCQE plot in which investigators would conclude the treatment was beneficial to some chosen degree of statistical or substantive significance. Investigators (and readers) can then argue whether δ is "certainly", "probably", "possibly", "probably not", or "certainly not" contained in this range. These conclusions may not be as satisfying as users would hope in most cases, but (i) can

be meaningfully different conclusions for decision makers and to guide future investigation, and (ii) reflect honest uncertainty about what conclusions can be reached or refuted.

Finally, one idea to consider is to combine statistical uncertainty and identification uncertainty into one ultimate quantity by requiring a distributional belief on δ, then computing the ATT and its uncertainty interval marginalized over this distribution. Such a marginalized estimate may be of interest in some settings. Here, however, we do not advocate this approach for two reasons. First, for researchers to construct a distribution over δ is far more challenging and fraught than proposing and defending values of δ that cannot be realistically exceeded. Second, for purposes of medical studies and decision-making, we argue that presenting results in a marginalized estimate and interval would not facilitate safe inference, and might invite abuse. In a given study, δ has a particular but unknown value, so the estimate "in expectation" over the distribution of what the investigator believes δ could have been provides no assurance about the effect of the treatment in the case under consideration. More concretely, if we find a harmful ATT estimate at a value of δ the investigator cannot convincingly refute, for our purposes this should be equally concerning regardless of what the treatment effect would be at other values, or how much weight is put on them in the marginalized estimate. We worry that presenting a marginalized estimate distracts from this and risks encouraging speculation or misunderstanding about what conclusions can be defended.[1] However, we recognize tastes may vary on this question and depending on the treatment studied and the audience of the analysis; we refer to Wulf (2021) where this is explored in greater depth.

2.3 Connections to other approaches

The SCQE can also be understood as a generalization or relaxation of IV and of DID, in which (i) the identification assumptions of these approaches are reformulated as equivalent assumptions on the baseline trend, and (ii) those assumptions are relaxed by postulated degrees. SCQE can thus be understood as a means to add sensitivity analysis to, or a partial identification variant of, IV and DID. We elaborate on each of these connections below.

Time as an instrument. SCQE can be viewed as a form of "broken IV" in which time serves as the instrument: moving from one cohort to the next changes the probability of treatment (Hazlett et al., 2020; Wulf, 2021). Conventional IV fails in this setting because the exclusion restriction and exogeneity assumption— which jointly ensure that any association of the instrument (cohort) and the outcome is the result of the instrument's effect on treatment uptake—is violated by any nonzero baseline trend (δ). One way to account for such violations is to introduce two sensitivity parameters: one describing how strongly an unobserved variable U relates to time (cohort identity), and another describing how strongly U relates to $Y(0)$ (see Cinelli and Hazlett, 2021). SCQE effectively captures the

1. Relatedly, in sensitivity analyses for other approaches (e.g. covariate adjustment), sensitivity parameters represent postulated violations of the assumption (e.g. no unobserved confounding). It is not typical practice to marginalize over an elicited distributional belief about these sensitivity parameters. Rather, as here, these sensitivity parameters are used to determine what conclusions can be defended based on what values of these parameters can arguably be excluded, or to indicate what would have to be believed about these parameters to sustain a particular conclusion (see e.g. Cinelli and Hazlett, 2020).

joint influence of these parameters through the baseline trend, which in many cases we may be able to reason about directly. The SCQE estimand can thus be interpreted as a δ-adjusted—or "exclusion restriction violation–adjusted"—Wald estimand. Conventional IV estimates, with their point estimate and confidence interval, correspond exactly to SCQE with $\delta = 0$. By embedding this as a special case within a broader range of δ values, SCQE makes explicit both the fragility of conventional IV's assumptions and the spectrum of defensible results that follow when those assumptions are relaxed.

Difference-in-differences. The SCQE may also bring the DID approach to mind. One major difference is that DID requires groupings for units so that members of the "treated group" will be untreated if they appear in the first cohort, but treated if they appear in the second cohort. We do not have such a structure in this example: for an individual who appears in the first cohort, we cannot label them as being in the group that would be treated (or not) in the second period. Thus, the application here is an example of a setting where the DID setup is not possible. In other settings where both DID and SCQE are possible (i.e. panel-structured data), SCQE again provides a reconceptualization of the identification assumption together with a sensitivity analysis. Specifically, the parallel trends assumption required of DID—that the over-time trend (in $\mathbb{E}[Y(0)|\text{group}]$) is the same in both groups—also implies that δ (the overall trend in $\mathbb{E}[Y(0)]$, pooling across groups) equals the over-time change in $\mathbb{E}[Y(0)]$ for just the control group. Thus, to rely on DID is to assert that δ is precisely equal to the observed change in outcomes seen in the control group. As with the IV comparison, the DID estimate could be shown as just one "row" on the SCQE plot at the value of δ set equal to the observed change found in the untreated group. Also like the IV comparison, this fact highlights the fragility of the parallel trends assumption, placing that result in the context of the estimates obtained under all the values of δ (each entailing violation of parallel trends) that the user is not able to refute. Here again, existing sensitivity analysis approaches have posited ways of reasoning about the parallel trends assumptions, for example by arguing it may be some multiple of the empirical difference in trends observed in prior time periods (Rambachan and Roth, 2023). The relative value of the SCQE as an approach to DID or IV sensitivity depends on whether reasoning about the baseline trend is natural in a given setting, and whether it illuminates which conclusions can or cannot be reached.

Variations in temporal cohort design There are several adjustments or extensions of the simple SCQE design described above, as discussed in Wulf (2021). The two cohorts may reflect a decrease in treatment usage rather than an increase (as in the second case below). Or, we may see more than two cohorts as treatment use oscillates (also illustrated below). More generally, any definition for the cohorts is allowable, so long as the user can reason about differences between the cohorts in baseline mortality risk, with all due conservativeness.

A prospective approach to SCQE is also possible: if we suspect that a change in non-random usage of some treatment is imminent, we could encourage a care team to refrain from making other practice changes over that time, enabling a narrower defensible range of δ values due to the absence of such secondary changes. Investigators can even conduct a form of SCQE analysis with only one cohort: given some treatment and outcome rates in the single cohort, we can determine the counterfactual outcome rates (had there been

no such treatment use) that we would have to defend in order to claim the treatment was beneficial or harmful.

SCQE for contemporaneous groups. While not employed here, we note that SCQE can also be constructed cross-sectionally, comparing a treatment-eligible (high-use) sample to a separate, possibly contemporaneous treatment-ineligible (no-use) sample. This design may resemble the "external control arm" (ECA) approach (e.g., Ventz et al., 2019; Seeger et al., 2020; Schmidli et al., 2020), but it differs in two key respects. First, ECA compares an observed, fully treated group to an untreated external group, making its validity depend on the comparability of treated and untreated patients—a setup that may invite more problematic confounding, since the treated actively self-selected for partly unobserved reasons. Second, SCQE is designed to produce cautious inference by reporting results across a range of assumed values for δ, whereas ECA analyses typically present a single estimate as if the external controls had exactly the same mean potential outcomes as the treated group.

3. Methods: Studying three COVID-19 therapies

3.1 Data construction and analysis template

Turning to the three applied studies, we briefly outline the common data construction and analysis procedures that will be utilized across all three cases in terms of cohort construction, plausible baseline trends, and the SCQE analysis itself.

Cohort construction. First, with IRB approval, data were extracted from the electronic medical records of two large tertiary care referral hospital systems. We use data from one of these systems to study dexamethasone and hydroxychloroquine (cases 1 and 2 below), and the other system to study remdesivir. In each case, the study populations are composed of individuals with recorded diagnoses for COVID-19 or positive PCR tests for SARS-CoV-2 infection. Clinical data extracted included pre-admission factors such as demographics and comorbidities, early hospitalization laboratory values and vital signs, COVID-19 treatments received, and hospital outcomes. Our primary outcome was cumulative mortality incidence within 28 days (henceforth, "mortality rate"), though study conclusions were unchanged for a 14-day mortality rate outcome.

For each treatment of interest, we start by identifying "high-use" and "low-use" cohorts. We are free to construct the start and end times of these cohorts in whatever manner we please, and choose to do so such that we maximize differences in the level of treatment seen in the two cohorts (as indicated by a large F-statistic in a regression of treatment use on cohort membership), while minimizing other changes between cohorts (e.g. the time-gap between them or the usage rates of other treatments). Details of the cohort construction are given in each case below.

Eliciting *ex ante* plausible baseline trends. Second, knowing the date ranges for the low-use and high-use cohorts for each treatment, we look to several sources to help inform us of a range of baseline trends (δ) we should declare to be plausible *ex ante*.

We start by assessing how remaining differences in baseline covariates and other treatments may be related to baseline mortality risk, and thus their potential contribution to baseline trends, basing these assessments in evidence available at the conclusion of the

study periods rather than months or years later. We also use statistical models trained on individual characteristics to fit mortality data, trained in the low-use cohort and then used for prediction both there and in the high-use cohorts. The average risk of mortality according to these models can be compared across cohorts. This provides a reasoned way to combine the covariates based on their prognostic value to inform arguments regarding the plausible range of baseline trends. Threats to the validity of these models are discussed in Appendices B and C, though we stress that that δ cannot be identified or directly estimated from the data, and as such our intention is only to inform values that we want to include in our plausible range.

Finally, we consider expert judgment of trend plausibility, proposed by treating physicians. When possible, we provided them with dates and full suite of baseline covariates above for each cohort in addition to the mortality rate in the low-use cohort to help them reason about possible δ ranges, i.e. their guesses for how different the mortality rate was in the high-use cohort. To minimize bias, we did not provide information about the shift in usage rates of the treatment under study (as recommended by Wulf, 2021). These conversations also allowed them to raise observations not in the data, for example information about other changes in treatment practices not recorded, about the impact of hospital crowding, about any change in the nature of the disease over this time, etc. These discussions occurred within months of the data collected here.

Aggregating the above suggestions of potential baseline trends, we declare our *ex ante* range of plausible δ values. We do so with a conservative approach, widening the range beyond the set of suggested values above until we can arguably defend the claim that the true but unknown δ is contained therein.

SCQE analysis. The final stage in each analysis is to conduct the SCQE analysis and examine the resulting plots. This can be done using the `SCQE` package for the `R` statistical software (Landsiedel et al., 2020), or through a web application available at `amiwulf.shinyapps.io/SCQE_demo`, which produces the same plots from only a series of summary statistics (rather than requiring unit-level data access). The SCQE plot shows the relationship between any postulated baseline trend and the consequent treatment effects. These plots show the results at δ values over as wide a range as we would like to see (for example, over a range sufficient to reach a research conclusion with either sign). They therefore reveal the ATT not only for the *ex ante* plausible values of δ but more broadly allow researchers, reviewers, and readers to clearly see "what would have to be believed" (about the baseline trend) to support a given conclusion.

3.2 Case 1: Dexamethasone

Cohort construction. Our studies of dexamethasone and hydroxychloroquine begin with data from 186 days of admissions at the first hospital system between 03/08/2020 to 09/24/2020. In this system, dexamethasone use increased over time. Cohorts obtain maximal treatment rate differences when ending the earlier ("low-use") cohort on day 101 and letting the later ("high-use") cohort include patients admitted between days 102 and 186. Further, we start the low-use cohort on day 44 in order to exclude a period of high hydroxychloroquine-usage before then, thereby avoiding the need to accommodate a possible hydroxychloroquine-caused shift in baseline mortality. The resultant low-use cohort com-

prised 614 patients with a 5.7% dexamethasone treatment rate, compared to the 534-patient high-use cohort's treatment rate of 46% (F=327, $p< 1 \times 10^{-15}$). The 28-day mortality rate was 10.9% in the low-use cohort and 5.4% in the high-use cohort, for a substantial raw risk difference (RD) of -5.5 (t=-3.36, $p=8 \times 10^{-4}$) percentage points (pp).

Table 1: Average baseline characteristics, treatments, and modeled risks, by dexamethasone cohort, first hospital system. Low-Use cohort: N=614 and dexamethasone treatment rate=5.7%. High-Use cohort: N=534 and dexamethasone treatment rate=46%.

	Covariate	Low-Use (Early)	High-Use (Late)
(A)	Age, years	55.0	54.5
	Male	53%	49%
	White	50%	53%
	Hispanic	50%	46%
	BMI, kg/m^2 (% missing)[a]	32 (45)	32 (40)
	Weight, lb (% missing)	187 (3)	195 (4)
	Admit from skilled care facility[b]	10%	2%
	Admit from non-healthcare location	76%	81%
	Within 24 hours of admission:		
	C-reactive protein, mg/L (% missing)	107 (18)	102 (24)
	White blood cells, 10^9/L (% missing)	8.88 (40)	8.03 (48)
	Ferritin, μg/L (% missing)	584 (37)	578 (44)
	Procalcitonin, μg/L (% missing)	0.88 (44)	0.81 (52)
	Ventilator use	10%	5%
	Intensive care unit admission	14%	8%
(B)	Treatment with:		
	Remdesivir	12%	27%
	Convalescent plasma	22%	20%
	Proning	3%	1%
	Hydroxychloroquine	3%	1%
	Tocilizumab	3%	2%
	Corticosteroid (not dexamethasone)[c]	10%	5%
(C)	Modeled 28-day mortality risk:		
	Linear (all covariates)	10.9%	8.8%
	KRLS (all covariates)	10.3%	8.6%
	Linear (only pre-admit or day-of)	11.1%	9.6%
	KRLS (only pre-admit or day-of)	10.4%	9.1%

[a]Means for this and other covariates calculated after excluding missing values. [b]Includes skilled nursing and long term acute care facilities. [c]Methylprednisolone, hydrocortisone, or prednisone.

Plausible baseline trends. Table 1 compares the cohorts on numerous baseline characteristics (A), other treatments that may have been in flux (B), and modeled baseline outcomes (C). Considering baseline characteristics (A), compared to the low-use (earlier)

cohort, the high-use cohort saw fewer ICU admissions and ventilator use within 24 hours of admission, and had fewer patients admitted from a skilled care facility. Each of these would suggest an anticipated drop in baseline mortality. Coming to changes in other therapies (B), we see there was higher usage of remdesivir in the high-use cohort (12% vs 27%). Remdesivir's effectiveness in reducing mortality was uncertain at the time, with RCTs showing either no benefit or benefit only in some post-hoc subset analyses (Wang et al., 2020b; Beigel et al., 2020).

Section (C) of Table 1 shows the modeled risk of mortality given by four model types and specifications.[2] This is simply a data-driven approach to combine and summarize the information above as it predicts any shift in the baseline mortality. Across these models, the predicted baseline mortality risk fell by 1.3-2.1pp.

Outside of these data, the treating physicians on our research team anticipated a decrease in mortality risk due to generally improving treatment practices through this period of time. All lines of evidence thus suggest a possible drop in baseline mortality in the range of a couple percentage points. Taking these considerations collectively, and maintaining a conservative attitude for this *ex ante* approach, we argue that the baseline trends plausibly lies between no cohort-to-cohort change and a reduction of up to 5pp (i.e. –46% from the 10.9% mortality in the low-use cohort).

SCQE analysis. Figure 2 shows the effect of dexamethasone on mortality (horizontal axis) as a function of possible baseline trend assumptions (vertical axis). Over the range of baseline trends previously deemed plausible (0 to -5pp), the point estimates lie in the negative (beneficial) direction over the entire range, with 95% confidence intervals excluding zero for nearly half this range (from 0 down to -2.3pp).

Turning to *ex post* argumentation, the plot shows that we could only sustain the claim that dexamethasone (statistically significantly) benefited those taking it if baseline mortality increased, stayed level, or fell by as much as 2.3pp (-21%). Alternatively, the point estimate remains negative (beneficial) so long as baseline mortality does not drop by more than 5.5pp – a 50% drop in mortality. Finally, to claim dexamethasone was (statistically significantly) harmful given our data, we would have to defend a decrease in baseline mortality of at least 8.7pp, an 80% drop to a mortality rate of just 2.2%. In this approach, while we cannot rule out δ values lower than -2.3pp, we do reject the possibility of an 80% drop, particularly without a known cause that escaped attention of the physicians on the team and does not appear in the observable differences in cohorts. Thus, while we cannot confidently defend the assumptions necessary to claim statistically significant benefits from dexamethasone, we can confidently reject those values that produce estimates of significant harm.

2. Two are very simple linear models (OLS), while two use a more powerful non-linear machine learning technique, kernel-regularized least squares (Hainmueller and Hazlett, 2014). For each model type, one specification uses the variables age; sex; race; Hispanic ethnicity; indicators for the two admission sources in Table 1; and indicators for ventilator use and ICU admission within 24 hours. The other specification additionally includes indicators for treatment with remdesivir, convalescent plasma, and any corticosteroid other than dexamethasone. Appendices B and C discuss relevant modeling choices and risks.

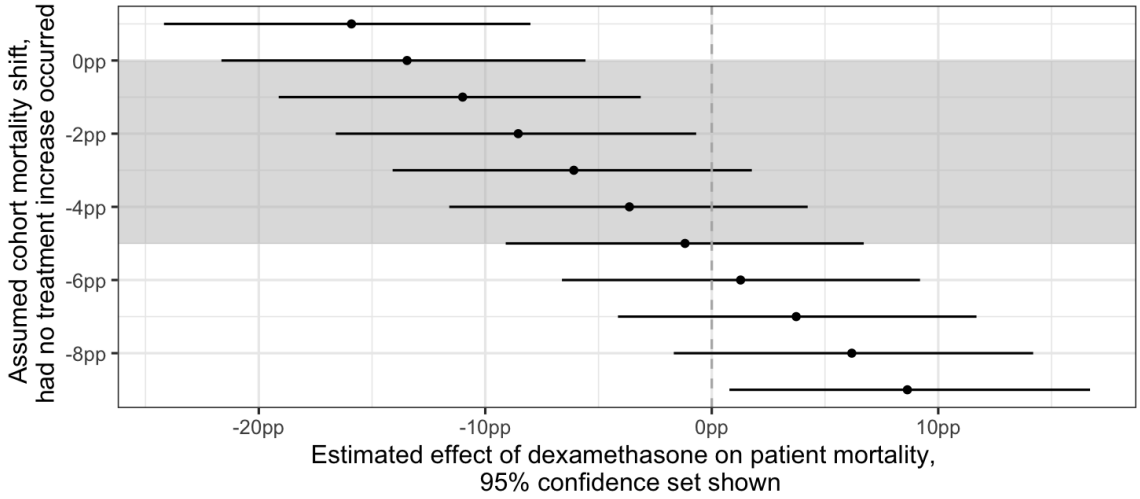


Figure 2: Estimated average treatment effect of dexamethasone on the 28-day mortality of those patients receiving it in the later cohort due to its increased usage. Estimates and confidence sets are displayed across values of the baseline trend δ, the counterfactual outcome shift between cohorts had dexamethasone use not increased. Plausible δ values are those ranging from 0 to -5pp, which support claims of either beneficial or null effects of dexamethasone.

3.3 Case 2: Hydroxychloroquine

Cohort construction. Within the same hospital system, hydroxychloroquine use, administered in a standard 5-day course, dropped over time, meaning that in this application the high-use cohort will precede the low-use cohort. In looking to maximize treatment rate differences between cohorts, the high-use (earlier) cohort is set to days 1-43. To avoid other major treatment differences between cohorts, we limit the low-use cohort to days 44-82, excluding a period of high dexamethasone use that follows. The resulting high-use cohort comprised 766 patients with a 36% hydroxychloroquine treatment rate, while the treatment rate for the 548 patients admitted in the low-use cohort was 2.9% (F=242, $p< 1 \times 10^{-15}$). The 28-day mortality rate was 14.5% in the high-use cohort and 11.5% in the low-use cohort, for a raw risk difference of -3.0pp (t=-1.58, p=0.11).

Plausible baseline trends. Cohort comparisons are shown in Table 2. Baseline characteristics (A) were largely similar, though the proportion identifying as Hispanic rose over time (from 38% to 49%), as did the fraction of patients coming from skilled care facilities (from 4% to 10%). Taken alone, these compositional shifts might suggest a small upward shift in baseline mortality risk over time. On the other hand, fewer patients in the later cohort required mechanical ventilation within a day of admission, suggesting a downward shift in risk is likely. Furthermore, two treatment practices (B) that could have potentially impacted risk increased over time between these cohorts: remdesivir (from 3% to 11%) and convalescent plasma (from 5% to 22%), though treatment guidance at the time suggested lit-

Table 2: Average baseline characteristics, treatments, and modeled risks, by hydroxychloroquine cohort, first hospital system. High-Use cohort: N=766 and hydroxychloroquine treatment rate=36%. Low-Use cohort: N=548 and hydroxychloroquine treatment rate=2.9%.

	Covariate	High-Use (Early)	Low-Use (Late)
(A)	Age, years	57.9	55.8
	Male	58%	54%
	White	51%	50%
	Hispanic	38%	49%
	BMI kg/m^2 (% missing)[a]	31 (41)	32 (55)
	Weight, lb (% missing)	194 (2)	187 (3)
	Admit from skilled care facility[b]	4%	10%
	Admit from non-healthcare location	77%	74%
	Within 24 hours of admission:		
	C-reactive protein, mg/L (% missing)	110 (22)	107 (17)
	White blood cells, 10^9/L (% missing)	7.73 (49)	8.75 (38)
	Ferritin, μg/L (% missing)	716 (32)	586 (36)
	Procalcitonin, μg/L (% missing)	0.65 (27)	0.92 (43)
	Ventilator use	16%	10%
	Intensive care unit admission	18%	15%
(B)	Treatment with:		
	Remdesivir	3%	11%
	Convalescent plasma	5%	22%
	Proning	3%	4%
	Tocilizumab	7%	3%
	Dexamethasone	4%	5%
	Other corticosteroid[c]	12%	11%
(C)	Modeled 28-day mortality risk:		
	Linear (all covariates)	11.8%	11.5%
	KRLS (all covariates)	11.6%	10.7%
	Linear (only pre-admit or day-of)	13.2%	11.5%
	KRLS (only pre-admit or day-of)	12.0%	10.8%

[a]Means for this and other covariates calculated after excluding missing values. [b]Includes skilled nursing and long term acute care facilities. [c]Methylprednisolone, hydrocortisone, or prednisone.

tle mortality benefit from either treatment (COVID-19 Treatment Guidelines Panel, 2020). As a conservative approach, suppose that all the 14.5% of patients who would have died in the low-use cohort (based on the rate in the earlier, high-use cohort) received treatment with remdesivir and/or convalescent plasma, and that these therapies, in any combination, reduced mortality by as much as 30%. These exceedingly generous assumptions would suggest a baseline mortality trend as low as -4.3pp. Looking to modeled outcomes (C), we

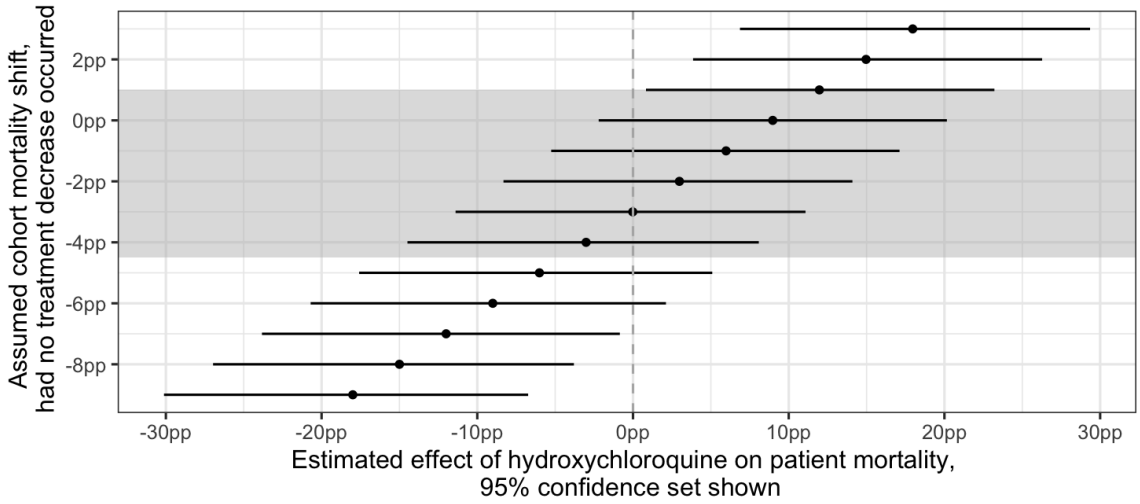


Figure 3: Estimated average treatment effect of hydroxychloroquine on the 28-day mortality of those patients who were treated in the early cohort but would not have been treated had they been in the later, low-use cohort (or equivalently, the effect on those late cohort patients left untreated because of its decreased usage). Estimates and confidence sets are displayed across values of the baseline trend δ, the counterfactual outcome shift between cohorts had hydroxychloroquine use not decreased. Plausible δ values are those ranging from +1pp to -4.5pp, which support claims of either harmful or null effects of hydroxychloroquine.

see that the predicted risk in the low-use cohort ranges from 0.3pp to 1.7pp lower than the predicted risk in the high-use cohort.[3]

As in the first case, considering the observed cohort differences as well as possible improvements in care over time suggested by practicing clinicians, a wide range of baseline trends are possible. To leave a conservative margin, we consider as plausible δ values ranging from a 1pp (7%) increase to a 4.5pp drop (-31%) in baseline mortality.

SCQE analysis. In a simple comparison of mortality-rates between cohorts, mortality drops when hydroxychloroquine use drops. The naive interpretation of this observation—that hydroxychloroquine was harmful—arises from a presumption that baseline mortality was not otherwise changing (i.e. $\delta = 0$). SCQE formalizes such reasoning, recognizing possible trends in baseline mortality. Figure 3 displays the effect of hydroxychloroquine on mortality implied by various baseline trends. Matching intuition, if there was no change in baseline mortality between these cohorts ($\delta = 0$), hydroxychloroquine appears to be substantially harmful with a point estimate of 9pp, i.e. a 33% increase in mortality due to use of hydroxychloroquine, though given the small sample size, the confidence interval still

3. The four models closely resemble those used in the dexamethasone case above. The only change in model type or specification is that the two models that include treatment variables (labeled "all covariates") use an indicator for *any* corticosteroid use rather than excluding dexamethasone use itself.

includes 0. If instead baseline mortality was worsening (in time) by 0.7pp (+2.6%) or more, hydroxychloroquine would prove statistically significantly harmful with a point estimate of over 11pp higher mortality.

In assessing the possibility of treatment benefit, we can only believe hydroxychloroquine to have been significantly beneficial if we can defend a baseline mortality risk decrease from the earlier to the later cohort by almost half (-46%, or -6.7pp). Even if this is possible, it would be very difficult to defend. It also falls outside the window we deemed *ex ante* plausible given expert knowledge and information about the characteristics of both cohorts and other treatment changes over this period. More generally, over much of our range of *ex ante* plausible baseline trend values, the point estimate lies in the positive (harmful) direction. Altogether, the results indicate that in this group, hydroxychloroquine either had no statistically significant effect, or had a significantly harmful effect.

3.4 Case 3: Remdesivir

Cohort construction. In the second hospital system, where we study 169 days of admission data from 03/11/2020 to 08/26/2020, the use of remdesivir started high (for patients admitted on days 1 through 30), dropped to zero (days 31 to 74), and then returned to a high level (day 75 onward), producing a "high-use, no-use, high-use" arrangement of cohorts. We exclude patients after day 109 to avoid the baseline trend risks introduced by a set patients with high dexamethasone usage who were admitted soon afterwards. However, we can combine the two high-use cohorts, disentangling time from cohort membership, while mitigating the impact of risk factors that shift monotonically over time. That is, while there may be notable differences between our cohorts, treatment practices and prognostic factors that decreased or increased monotonically across the study period would partially cancel out between high-use and no-use cohorts once we combine the early and late high-use periods into one cohort. The resulting no-use cohort is made up of 53 untreated patients, while the 83 patients in the pooled high-use cohort had a 31.3% remdesivir treatment rate (F=23.8, p=3×10^{-6}). The 28-day mortality rate was 24.5% in the no-use cohort and 13.3% in the pooled high-use cohort, for a raw risk difference of -11.2pp (t=-1.69, p=0.09).

Cohort comparison and plausible baseline trends. Table 3 compares the cohorts using the set of covariates available from the second hospital system. We see (A) that patients in the high-use cohort were younger, less likely to arrive from a nursing home, by ambulance, or have a history of hyperlipidemia (though comorbidity prevalences are based on particularly small samples), showed fewer abnormal lab results, and received supplemental oxygen less often and at lower levels, implying the possibility of lower mortality risk (CDC COVID-19 Response Team et al., 2020; Chen et al., 2020; Huang et al., 2020; Wang et al., 2020a). Members of the high-use cohort were more frequently male, however, which had been linked to higher risk (Peckham et al., 2020). Notable treatment differences in the first 7 days (B) for this cohort included higher use of dexamethasone and hydroxychloroquine and lower use of ceftriaxone, leronlimab, and tocilizumab. Our modeling procedure (C) predicts baseline risk differences of -7.6pp and -10.2pp.[4]

4. Given the smaller sample size for this hospital system, models with too many covariates risked overfitting on the 53-patient no-use cohort. Models with pre-admission markers, admission-day status, and other treatments given were infeasible. We thus limited the covariates used by both the linear and KRLS

Table 3: Average baseline characteristics, treatments, and modeled risks, by remdesivir cohort, second hospital system. No-Use cohort: N=53. High-Use cohort: N=83 and remdesivir treatment rate=31.3%.

	Covariate	No-Use (Middle)	High-Use (Early + Late)
(A)	Age, years	69.0	63.5
	Male	51%	59%
	White	34%	43%
	Latinx	32%	30%
	Diabetes (% missing)[a]	46% (75)	33% (75)
	Asthma (% missing)	15% (75)	10% (75)
	Hyperlipidemia (% missing)	38% (75)	0% (75)
	BMI, kg/m^2 (% missing)	26.4 (6)	27.9 (6)
	Arrived by ambulance	17%	8%
	Admit from nursing home	30%	11%
	Within 24 hours of admission:		
	C-reactive protein, mg/L (% missing)	103 (19)	104 (30)
	White blood cells, 10^9/L (% missing)	8.65 (0)	8.76 (0)
	Neutrophils, 10^9/L (% missing)	6.93 (2)	6.41 (7)
	Lymphocytes, 10^9/L (% missing)	0.94 (2)	1.30 (7)
	Creatinine, mg/dL (% missing)	1.96 (2)	1.26 (7)
	Oxygen saturation (% missing)	92.7% (0)	93.8% (0)
	Oxygen flow, L/min (% room air)	6.16 (19)	3.57 (28)
	Ventilator use	13%	19%
	Intensive care unit admission	34%	29%
(B)	Within 7 days of admission:		
	Dexamethasone	0%	8%
	Other corticosteroid[b]	15%	19%
	Proning	8%	6%
	Convalescent plasma	2%	0%
	Hydroxychloroquine	13%	22%
	Heparin	47%	53%
	Enoxaprin	57%	54%
	Azithromycin	57%	54%
	Ceftriaxone	70%	55%
	Leronlimab trial[c]	25%	5%
	Tocilizumab	11%	6%
(C)	Modeled 28-day mortality risk:		
	Linear (pre-admit/day-of covariates)	24.5%	14.3%
	KRLS (pre-admit/day-of covariates)	23.3%	15.8%

[a]Throughout, prevalences and means calculated after excluding missing values. [b]Prednisone, Methylprednisolone, or Hydrocortisone. [c]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.

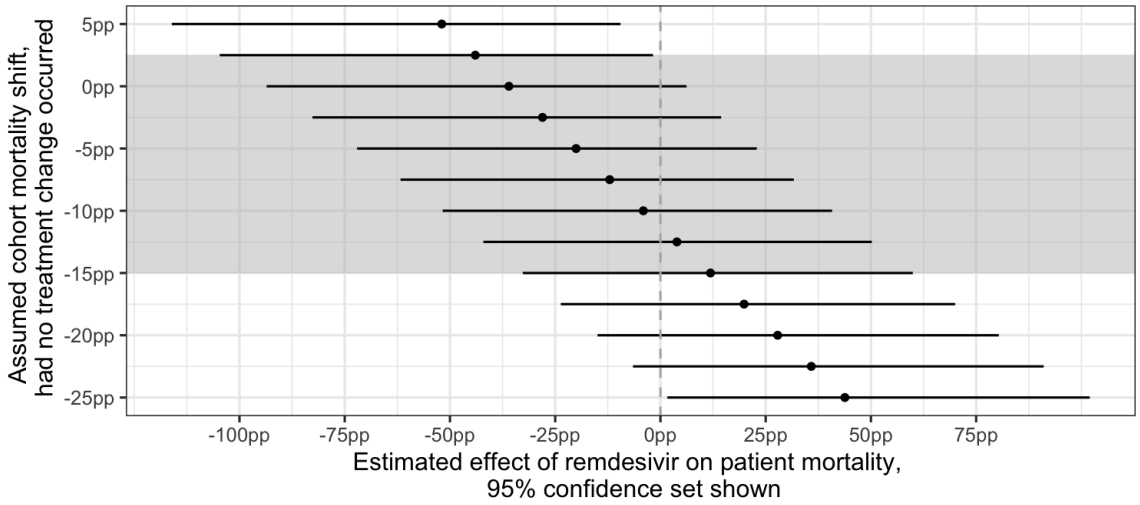


Figure 4: Estimated average treatment effect of remdesivir on the 28-day mortality of treated patients in the high-use cohort. Estimates and confidence sets are displayed across values of the baseline trend δ, the counterfactual outcome shift from the no-use cohorts to the high-use cohort had remdesivir been used in neither cohort. Plausible δ values are those ranging from +2.5pp to -15pp, which support claims of either beneficial or null effects of remdesivir.

One of the authors, an infectious disease specialist on the care team for many patients in this hospital system, judged that these differences may suggest a reduction in baseline mortality in the high-use cohort relative to the no-use cohort, leading to a conservative range of baseline trend values from -5pp to -15pp. Given the split high-use cohort, the kind of over-time improvements in disease management that factored into physician assessment of δ in the other two cases are less concerning here as noted above. Combining these considerations, and acknowledging added uncertainty due to the small sample, we suggest a conservative range of baseline trend values ranging from +2.5pp (+10%) to -15pp (-61%) cannot be definitively ruled out.

SCQE analysis. Figure 4 shows the effect of remdesivir we would obtain under each assumed baseline trend value. Over the range of baseline trends deemed plausible (2.5pp to -15pp), almost all values lead to non-significant estimates, mostly but not entirely falling in the beneficial direction, with one edge of the range implying statistically significant beneficial effects of remdesivir. If baseline mortality was 1.9pp higher in the high-use cohort(s) compared to no-use cohort (just within the range proposed as plausible), we would conclude that remdesivir significantly lowered mortality rates. Trends of -24.5pp or lower would imply statistically harmful effects. Note that the mortality rate in the no-use cohort was 24.5%, and thus to believe that remdesivir was significantly harmful would require believing that,

models to age; sex; admission from a nursing home; and oxygen saturation, white blood cell count, and transfer to the ICU within 24 hours.

absent changes in remdesivir use, mortality would have dropped to precisely 0 — clearly an impossible claim to make. Our conclusions in this case are especially hampered by the small sample size and resulting uncertainty, even at given values of δ. However, we can conclude that over the range of δ values deemed plausible, remdesivir would have either shown a marginally significant benefit or an effect indistinguishable from zero (at the $p < 0.05$ level), while we can confidently rule out that it was significantly harmful.

3.5 Comparison to covariate-adjustment with sensitivity analysis

How do these findings differ from those one would arrive at using conventional, covariate-adjustment approaches? Without sensitivity analysis, a covariate-adjustment approach would produce estimates as if an undefended assumption (no unobserved confounding) holds precisely. We are instead interested in comparing what research conclusions would have been credibly defensible by applying covariate-adjustment with sensitivity analysis (to unmeasured confounding), to those conclusions that are credibly defensible via the SCQE approach. These approaches can be differentially informative in different settings, depending on our ability to make bounding assumptions on unobserved confounding versus those on the baseline trend.

Dexamethasone. In a regression of 28-day mortality on dexamethasone treatment, we limit the covariates used for adjustment to: age; sex; race; Hispanic ethnicity; admission sources; within-24-hour transfer to the ICU and ventilation; and treatment with other corticosteroids, convelescent plasma, remdesivir, tocilizumab, and use of proning.[5] The regression coefficient for dexamethasone in the model just described is -3.5pp (t=1.9, p=0.06). The ATT estimate of -3.4pp is similar.[6] Using these results to claim a marginally beneficial effect (with questionable statistical significance) relies on the assumption of precisely zero unobserved confounding, while sensitivity analysis can reveal the fragility of that finding in the face of violations to the assumption. Using the approach of Cinelli and Hazlett (2020), we find first that confounding that implies a larger true benefit than observed would only have to explain 0.15% of residual variance in treatment (mortality) and outcome (dexamethasone use) to make the estimate statistically significant at the 0.05 level. Confounding that works against the apparent benefit would only have to explain 11% of residual variance in treatment and outcome to imply that the true estimate (if we could adjust for such confounding) would have been statistically significant in the harmful direction.

We then ask whether unobserved confounders could plausibly explain 11% or more of these residual variances. If we cannot rule out confounding of that degree, we cannot in turn rule in or out any conclusion about statistically significant benefit or harm.[7] We

5. We construct linear probability models, i.e., we use OLS despite the binary outcome. This is now standard in many fields, because it is still a well-defined conditional expectation function, particularly in cases where all fitted values can be expected to fall well within $[0, 1]$. See e.g. Angrist and Pischke (2008).
6. Using regression imputation/g-computation we used a model trained on the untreated to predict the treated patients' non-treatment outcomes, and compared those to their predicted treatment outcomes outcomes from a model trained on the treated, to produce the average treatment effect among the treated (ATT).
7. As pointed out in Cinelli and Hazlett (2020) and many other pieces using these techniques, users may be positioned to defend such arguments in some contexts, for example by leveraging "benchmarks" that

judge confounding at these levels to be very difficult to rule out, owing to the surfeit of unobserved factors that may influence both treatment decisions and mortality risk. This assessment is supported by the judgment of treating physicians involved in making these decisions—including authors on this paper—who attest that many unmeasured factors affect clinical judgments in these treatment decisions and could also influence mortality risk. These include, but are not limited to, indicators of health, resilience, and disease severity not recorded in the data, family support, and patient preference. Thus while sensitivity analysis for covariate adjustment is very useful for demonstrating the fragility of estimates, and sometimes reveals robustness to even strong degrees of confounding that can arguably be ruled out (see e.g. Hazlett and Parente, 2023), in this case it does little to credibly defend or rule out any conclusion about the impact of the treatment.

By comparison, leveraging arguable bounds on the baseline trend, the SCQE suggests that the point estimate was very likely in the beneficial direction, and possibly statistically significant. SCQE further shows that a point estimate in the harmful direction requires a claim that the baseline mortality rate dropped by at least half, and supporting a statistically significant harmful effect requires defending the possibility of an 80% drop in baseline mortality. Even though both approaches map uncertainty about an assumed value to uncertainty about the treatment impact, in this setting reasoning with the baseline trend assumption provides enough traction to rule out one conclusion (a statistically significant harm) while reasoning with the degree of unobserved confounding does not rule out any conclusions by our judgments.

Hydroxychloroquine. We use similar covariates for adjustment as above, though here we include dexamethasone with other corticosteroids as a control covariate. The regression coefficient for hydroxychloroquine indicates a 5.2pp higher risk of mortality (t=2.4, p=0.01). The ATT estimate is similar (6.0pp). While the coefficient estimate and its implication of statistically significant harm falls well within the range of SCQE estimates, it is not itself *defensible*, as it depends on the claim of no unobserved confounding. Sensitivity analysis again helps to show the fragility of this result. If unobserved confounding explains even 1% of residual variance in the treatment and the outcome, then it would imply the (so adjusted) effect estimate would not be statistically significant at the 0.05 level. Confounding that explains 7% of residual variance in treatment and outcome would change the sign of the estimate. Confounding that explains 12% of residual variance in the treatment and the outcome would imply that the adjusted estimate would have been a statistically significant *beneficial* effect at the 0.05 level. As above, we see no reason to claim that unobserved confounding explains less than 12% of residual variance in the use of hydroxychloroquine and in the outcome. Thus, the covariate-adjusted comparison should not be understood to make or rule out any claim at all; once paired with an appropriate sensitivity analysis, the point estimate from this approach could coincide with a true beneficial, null, or harmful effect. By contrast, the SCQE analysis suggests the baseline trend assumption needed to claim a statistically significant benefit is implausible, while baseline trends implying a lack of benefit or even significant harm are quite possible.

compare the strength of unobserved confounding to what is explained by factors we know are essential to explaining the treatment uptake or the outcome. In this case, because of the great potential for unobserved confounding/limited information about how treatment was assigned, we did not find such approaches to convincingly limit confounding to degrees that would bear on the conclusion.

Remdesivir. In this case, with a smaller and different dataset, the regression controls for age; sex; admission from a nursing home; arrival by ambulance; within-24-hour oxygen saturation, white blood cell count, and admission to the ICU; and treatment with dexamethasone. That model suggests those given remdesivir had a 1.6pp lower risk of mortality, not statistically different from 0 (t=0.34, p=0.73). The ATT estimate is again similar (-1.2pp). Statistical uncertainty alone is enough to preclude any strong conclusion from being made. The question, however, is how different this might have been under various degrees of confounding we cannot rule out. If confounding was working against finding a benefit, then confounding that explains 15% or more of residual variance in remdesivir use and mortality would imply the actual effect to be significantly beneficial. Similarly, if confounding was inflating the apparent benefit, confounding that explains 18% of residual variance in remdesivir and the outcome would imply that remdesivir would have actually had a significantly harmful effect. While these potential degrees of confounding are larger than those in the cases above, they do not seem to us to be implausible, particularly given the smaller sample size and fewer covariates. By comparison, the plausibility arguments we were able to make around an assumption on the baseline trend enabled us to effectively rule out a harmful effect.

4. Discussion

The SCQE method offers one approach for drawing safe yet potentially informative inferences from treatments attempted outside of randomized trials, applied here to study three COVID-19 therapies. Two features of this approach may be unfamiliar to many readers and investigators. First, SCQE does not make any assumptions about the absence or strength of unobserved treatment-outcome confounding. Given the extensive list of treatment considerations, clinical judgment, and unrecorded factors influencing treatment decisions for COVID-19 patients in the early pandemic, assumptions of this type would be difficult to persuasively defend. Such assumptions are replaced in SCQE with an interval assumption of what baseline mortality trend ranges can be ruled plausible or implausible. Second, as an example of a partial identification approach, this strategy avoids producing a focal estimate that is claimed to be the result, or a "best guess" for purposes of future decision-making. Just as we have accommodated to the need for confidence intervals to reflect what we don't know due to statistical uncertainty, the wider range of estimates here reflect uncertainty resulting from the plausible range of causal assumptions, here regarding δ. While potentially frustrating, this feature helps to protect us against reaching over-confident conclusions rooted in indefensible assumptions, and serves as a built-in analogue of sensitivity analyses as applied to other identification strategies.

The SCQE is an additional tool that may provide useful leverage for bounding credible causal claims outside of randomized trials. Whether it is feasible and advantageous in a given setting depends on several features. On feasibility, for the temporal version of SCQE considered here, it must be possible to construct time periods such that the difference in treatment probability between them is relatively large. Second, the window of eligibility for being given the treatment must be relatively short, so that it is possible to label each unit as

clearly falling in the "low-use" or "high-use" cohort.[8] If these are satisfied, one advantage of SCQE is that it can be computed with only a few summary statistics on cohort membership, treatment status, and outcome rates. However, covariate means, expert assessments about outcome trends, and individual-level data are useful for better considering plausible values for δ.

Where feasible, SCQE always provides a mapping between any postulated baseline trend (δ) and the logically implied ATT or CATE. The result will be less informative where we are less able to argue for bounds on δ. This can be useful for shedding light on what research conclusions we should *not* be comfortable supporting. On the other hand, the results will be increasingly informative where the ability to reason about δ provides identification leverage that is not exploited by other strategies. For example, covariate adjustment with sensitivity analysis may provide useful leverage where expert knowledge facilitates claims about bounds on the strength of unobserved confounding. However, in settings where unobserved confounding is difficult to bound due to limited knowledge of the treatment assignment process (and influences on the outcome), but we can reasonably argue for some bounds on baseline trends, SCQE may open the door to more informative results. As also noted above, SCQE provides a generalization or relaxation of DID and IV methods, reexpressing their identification assumptions in terms of the equivalent baseline trend, and showing the consequences of violating those assumptions.

In the applications studied here, despite the small size of the samples available to us, we find that somewhat informative conclusions can be defensible given baseline trend ranges deemed plausible *ex ante*. Results are perhaps clearest for dexamethasone: under roughly half the plausible range, dexamethasone would have had a statistically significant beneficial effect, while under the remaining range SCQE persists in suggesting non-significant point-estimates in the beneficial direction. To sustain a statistically harmful effect one would need to argue cohort-to-cohort baseline risk dropped over 80%, which we believe to be implausible. Nearly the opposite picture arises for hydroxychloroquine: we cannot rule out baseline trends implying it had statistically significant harmful effects, and though most of our plausible range lead to statistically insignificant effect estimates, we can declare as implausible the baseline mortality drops (-46% or lower) that would lead to claims of hydroxychloroquine's beneficial effect. Finally, for remdesivir, the point estimates fall in the beneficial direction across most of our plausible baseline trends range. There is less promise here for a statistically significant benefit than in the case of dexamethasone, however. More clearly, the baseline trends necessary to conclude remdesivir was statistically harmful are implausible, as they require believing that there would have been no mortality at all in the high-use cohort in the hypothetical absence of remdesivir treatment changes.

We also compare the conclusions that can be supported through SCQE to those that could be supported through a covariate-adjustment approach accompanied by a suitable sensitivity analysis. In this case, we find that the vulnerability to unobserved confounding under the covariate adjustment approach produces a wide plausibility range, preventing us from ruling out significantly harmful or significantly beneficial effects for any of the therapies. By comparison, SCQE trades judgments on the plausible strength of unobserved

8. For the construction of SCQE that instead compares two cohorts from different, possibly contemporaneous groups, these challenges are alleviated but the burden shifts to our ability to constraint δ in this setting.

confounding for judgments about the plausible range of baseline trends. In this particular case, the latter offers leverage not available through the former.

These results would have proven useful to physicians and patients at the time before many of the largest trials were completed (Beigel et al., 2020; RECOVERY Collaborative Group, 2021; WHO Solidarity Trial Consortium, 2021, 2022). One virtue of the SCQE is that it allows us to study populations different from those enrolled in randomized trials. Accordingly, our population differs from those studied in RCTs for these therapies. Moreover, our sample is much smaller. Nevertheless, the conclusions we reach here remain consistent with the results of those trials and with the most recent trials and guidance (e.g. Lamontagne et al., 2023; COVID-19 Treatment Guidelines Panel, 2023), which describe the lack of benefit of hydroxychloroquine and the likely benefit of dexamethasone in particular.

Declarations

Ethics approval and consent to participate

UCLA IRB approval #20-000981; University of Colorado Multi-Institutional Review Board approval #20-0690.

Availability of data and materials

The datasets generated and analyzed during the current study are not publicly available due to institutional restrictions on data sharing and privacy concerns. The code used to analyze that data for one of the Cases is provided at github.com/dawulf/remdesivir-scqe-code to guide interested readers through the process and inform their own use of SCQE.

Competing interests

KME has received research funding and has consulted for Gilead Sciences, paid to the University of Colorado. Remaining authors declare that they have no competing interests.

Funding

CH was partially supported by the California Center for Population Research at UCLA (CCPR), which receives population research infrastructure funding (P2C-HD041022) from the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD). DGM was supported by the U.S. National Institute on Drug Abuse (grant K08DA048163). OAA was partially supported by the National Institutes of Health (NIH) National Center for Advancing Translational Sciences (NCATS) UCLA CTSI grant number UL1TR001881 and the UCLA David Geffen School of Medicine (DGSOM) – Broad Stem Cell Research Center (BSCRC) COVID-19 Research Award OCRC #20-44. KME was partially supported by National Institute on Aging (NIA) grant R01AG054366-05. KME and BTM were partially supported by NIH/NCATS Colorado CTSA Grant Number UL1 TR002535. Contents are the authors' sole responsibility and do not represent official views of NIH or any other agency.

Authors' contributions

DAW: Methodology, software, formal analysis, writing (original draft & editing), visualization. CH: Conceptualization, methodology, software, writing (original draft & editing), administration. BLH: Software, data curation, writing (review & editing). JNC: Software, data curation, writing (review & editing). DGM: Domain expertise, investigation, validation, writing (review & editing). BP: Investigation, validation, writing (review & editing), administration. OAA: Methodology, validation, writing (review & editing), administration. KME: Domain expertise, data curation, investigation, administration, writing (review & editing). BTM: Domain expertise, data curation, investigation, administration, writing (review & editing).

References

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

Joshua D Angrist and Jörn-Steffen Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press, 2008.

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 — final report. *New England Journal of Medicine*, 383(19):1813–1826, 2020.

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.

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.

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.

Carlos Cinelli and Chad Hazlett. An omitted variable bias framework for sensitivity analysis of instrumental variables. Unpublished Manuscript, 2021. URL `https://carloscinelli.com/files/Cinelli%20and%20Hazlett%20(2020)%20-%20OVB%20for%20IV.pdf`.

COVID-19 Treatment Guidelines Panel. Coronavirus disease 2019 (COVID-19) treatment guidelines. Oct 9, 2020 edition, National Institutes of Health, 2020.

COVID-19 Treatment Guidelines Panel. Coronavirus disease 2019 (COVID-19) treatment guidelines. Apr 20, 2023 edition, National Institutes of Health, 2023.

Lucy D'Agostino McGowan. Sensitivity analyses for unmeasured confounders. *Current Epidemiology Reports*, 9(4):361–375, 2022.

Jens Hainmueller and Chad Hazlett. Kernel regularized least squares: Reducing misspecification bias with a flexible and interpretable machine learning approach. *Political Analysis*, pages 143–168, 2014.

Chad Hazlett. Estimating causal effects of new treatments despite self-selection: The case of experimental medical treatments. *Journal of Causal Inference*, 7(1), 2019.

Chad Hazlett and Francesca Parente. From "is it unconfounded?" to "how much confounding would it take?": Applying the sensitivity-based approach to assess causes of support for peace in colombia. *The Journal of Politics*, 85(3):1145–1150, 2023.

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.

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.

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.

François Lamontagne, Arnav Agarwal, Bram Rochwerg, Reed AC Siemieniuk, Thomas Agoritsas, Lisa Askie, Lyubov Lytvyn, et al. A living who guideline on drugs for covid-19. *BMJ*, Update 12, published January 2023, 2023. doi: 10.1136/bmj.m3379. URL `https://www.bmj.com/content/370/bmj.m3379`.

Kirsten Landsiedel, Colleen Pinkelman, David Ami Wulf, and Chad Hazlett. scqe: An R package for the stability-controlled quasi-experiment, 2020. URL `https://github.com/chadhazlett/scqe`.

Danyu Y Lin, Bruce M Psaty, and Richard A Kronmal. Assessing the sensitivity of regression results to unmeasured confounders in observational studies. *Biometrics*, pages 948–963, 1998.

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.

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 identified by global COVID-19 meta-analysis as a risk factor for death and ITU admission. *Nature Communications*, 11(1):1–10, 2020.

Ashesh Rambachan and Jonathan Roth. A more credible approach to parallel trends. *Review of Economic Studies*, 90(5):2555–2591, 2023.

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

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

Heinz Schmidli, Dieter A Häring, Marius Thomas, Adrian Cassidy, Sebastian Weber, and Frank Bretz. Beyond randomized clinical trials: use of external controls. *Clinical Pharmacology & Therapeutics*, 107(4):806–816, 2020.

John D Seeger, Kourtney J Davis, Michelle R Iannacone, Wei Zhou, Nancy Dreyer, Almut G Winterstein, Nancy Santanello, Barry Gertz, and Jesse A Berlin. Methods for external control groups for single arm trials or long-term uncontrolled extensions to randomized clinical trials. *Pharmacoepidemiology and Drug Safety*, 29(11):1382–1392, 2020.

Steffen Ventz, Albert Lai, Timothy F Cloughesy, Patrick Y Wen, Lorenzo Trippa, and Brian M Alexander. Design and evaluation of an external control arm using prior clinical trials and real-world data. *Clinical Cancer Research*, 25(16):4993–5001, 2019.

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, 2020a.

Yeming Wang, Dingyu Zhang, Guanhua Du, Ronghui Du, Jianping Zhao, Yang Jin, Shouzhi Fu, Ling Gao, Zhenshun Cheng, Qiaofa Lu, et al. Remdesivir in adults with severe COVID-19: a randomised, double-blind, placebo-controlled, multicentre trial. *The Lancet*, 395(10236):1569–1578, 2020b.

WHO Solidarity Trial Consortium. Repurposed antiviral drugs for Covid-19—interim who solidarity trial results. *New England Journal of Medicine*, 384(6):497–511, 2021.

WHO Solidarity Trial Consortium. Remdesivir and three other drugs for hospitalised patients with COVID-19: final results of the who solidarity randomised trial and updated meta-analyses. *The Lancet*, 399(10339):1941–1953, 2022.

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.

Appendix A. Robust confidence sets for SCQE

In this paper, we quantify statistical uncertainty around a δ-conditional ATT estimate via the connection between SCQE and Instrumental Variables (IV). In Hazlett et al. (2020), 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 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.

Rather than using the traditional standard error estimator in IV, we employ the Anderson-Rubin confidence sets (Anderson and Rubin, 1949), which have correct coverage even in the face of weak instruments (i.e. if the change in treatment usage between cohorts were to be small, though in our three cases this is not a concern). 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` (Kang et al., 2021). We note that different values of δ give different confidence interval widths, although in our cases the range of δ is small enough and the sample size large enough that these differences are not visible in the plots. When SCQE can be implemented without unit-level data (as noted in the discussion section), Anderson-Rubin confidence sets can still be constructed (Wulf, 2021).

Appendix B. "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 Tables 1, 2, and 3 are strictly pre-treatment. Measures taken in the first 24 hours could in principle occur after a decision to give the treatment (the one under study in a given case), but we do not anticipate they could have had any effect within this time, and thus these measures are plausibly pre-treatment. Other treatments given, however, even those within 7 days, are more likely to be affected by the primary treatment's 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. Taking the remdesivir case as an example, if tocilizumab is affected by remdesivir use, then using observed cohort-to-cohort differences in tocilizumab's presence to inform δ will induce bias, as those differences may represent indirect effect pathways (tocilizumab use could be a mediator of remdesivir's effect on mortality) rather than violations of the exogeneity assumption we wish to include in δ. Additional related introductions and sources of bias are discussed in Wulf (2021).

Although sensitivity analyses may help identify how our covariate-driven δ-evaluation procedure is impacted by covariate choice and our definition of "baseline", we note that the

original procedure is not meant to be precise enough to trust a narrow δ range, and any range it produces should only be treated as one suggestion to consider. Still, the fact that models we trained to predict baseline risk with and without the post-admission treatments produced similar suggestions for δ is a welcome finding. Of course, this does not guarantee our model is correct, only that it is to some degree robust to that particular treatment inclusion choice.

Appendix C. 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.

In the remdesivir case, 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 Appendix B – 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 Appendix B is discussed in Wulf (2021). Note that this assumption is likely far more defensible than the standard conditional ignorability assumption, which can be written $Y(0) \perp\!\!\!\perp D \mid X$.

In the dexamethasone and hydroxychloroquine cases, we face a challenge in that neither cohort has 0 treatment use. We could choose to adjust the procedure by training our baseline risk model on only those patients that were not treated, or by training it on all patients in the low-use cohort, acting as though they had not been treated. Here, we opt for the latter. Training the baseline risk model on a subset of the population defined precisely by the selection process SCQE looks to avoid would be counterproductive. Instead, we acknowledge that some of what is being modeled as baseline mortality risk has in fact been influenced by the treatment. The resultant bias is minimized by the small amount of treatment in our two low-use cohorts, but can still mislead us to some degree when using the model predictions to obtain suggestions for δ values. We can respond by widening the range of δ values we consider plausible by an arbitrary amount, acknowledging that the results from these procedures are intended to be informative rather than prescriptive for our assessment of plausible baseline trends.