Credible learning of hydroxychloroquine and dexamethasone effects on covid-19 mortality outside of randomized trials

Chad Hazlett, Ami Wulf, Bogdan Pasanuic, Onyebuchi Arah, Kristine Erlandson, Brian Montague

What is already known on this topic: Considerable uncertainty remains regarding the effectiveness of various covid-19 therapies. Little evidence exists to suggest hydroxychloroquine is beneficial, with some evidence for potential harm. For dexamethasone and other glucocorticoids, evidence suggests they may have a mild benefit, with little evidence of harm. To date, observational studies have relied upon covariate-adjustment strategies.

What this study adds: Using an alternative to covariate-adjustment approaches called the stability-controlled quasi-experiment (SCQE), we examine electronic health records outside of randomized trials and conclude that (i) it is difficult to argue for a beneficial effect of hydroxychloroquine, and (ii) dexamethasone is useful under a wide range of plausible assumptions and it is nearly impossible for dexamethasone to have had a significantly harmful effect in this sample. More broadly, the SCQE approach offers a safe and rigorous way to characterize what can be learned from rapid changes in the use of experimental therapies outside of non-randomized trials, as complements to randomized trials or where those trials are not forthcoming.

ABSTRACT

Objectives: To investigate the effectiveness of hydroxychloroquine and dexamethasone on coronavirus disease (covid-19) mortality using patient data outside of randomized trials.

Design: Phenotypes derived from electronic health records were analyzed using the stability-controlled quasi-experiment (SCQE) to provide a range of possible causal effects of hydroxychloroquine and dexamethasone on covid-19 mortality.

Setting and participants: Data from 2,007 covid-19 positive patients hospitalized at a large university hospital system over the course of 200 days and not enrolled in randomized trials were analyzed using SCQE. For hyrdoxychloroquine, we examine a high-use cohort (n=766, days 1 to 43) and a later, low-use cohort (n=548, days 44 to 82). For dexamethasone, we examine a low-use cohort (n=614, days 44 to 101) and high-use cohort (n=622, days 102 to 200).

Outcome measure: 14-day mortality, with a secondary outcome of 28-day mortality.

Results: Hydroxycholoroquine could only have been significantly (< 0.05) beneficial if baseline mortality was at least 6.4 percentage points (55%) lower among patients in the later low-use than the earlier high-use cohort. Hydroxychloroquine instead proves significantly harmful if baseline mortality rose from one cohort to the next by just 0.3 percentage points. Dexamethasone significantly reduced mortality risk if baseline mortality in the later (high-use) cohort (days 101-200) was higher than, the same as, or up to 1.5 percentage points lower than that in the earlier (low-use) cohort (days 44-100). It could only prove significantly harmful if mortality improved from one cohort to the next by 6.8 percentage points due to other causes – an assumption implying an unlikely 94% reduction in mortality due to other causes, leaving an in-hospital mortality rate of just 0.4%.

Conclusions: The assumptions required for a beneficial effect of hydroxychloroquine on 14 day mortality are difficult to sustain, while the assumptions required for hydroxychloroquine to be harmful are difficult to reject with confidence. Dexamethasone, by contrast, was beneficial under a wide range of plausible assumptions, and was only harmful if a nearly impossible assumption is met. More broadly, the SCQE provides a useful tool for making reasoned, limited and credible inferences from non-randomized uses of experimental therapies, when randomized trials are still ongoing and will take long, or to provide corroborative evidence from different populations.

CH thanks the California Center for Population Research at UCLA (CCPR) for support. CCPR receives population research infrastructure funding (P2C-HD041022) from the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD). KE was partially supported by National Institute on Aging (NIA) grant R01AG054366-05. UCLA IRB approval #20-000981.

INTRODUCTION

Although randomized controlled trials (RCTs) are the gold standard for learning causal effects of treatments on outcomes, running and awaiting the results of RCTs remains challenging and sometimes infeasible. This is particularly evident in the case of the SARS-CoV-2 infection that led to the coronavirus disease (covid-19) pandemic, where a multitude of treatments were adopted in the clinic on an urgent basis. It is particularly in these cases where the ability to draw credible inferences regarding the effects of treatment used outside of RCTs is of enormous interest to patients, healthcare workers and researchers. Yet, conventional approaches to such observational studies have well-known limitations, particularly their vulnerability to uncontrolled confounding, which can bias results such that harmful treatments could appear beneficial or vice versa without warning. Physicians and other expert consumers of medical research are often (rightly) wary of drawing conclusions about treatment effects – be they null, beneficial, or harmful – from non-randomized comparisons. Nevertheless, particularly in emergencies such as the covid-19 pandemic, healthcare providers often need to make decisions before RCTs have been completed or for individuals not well represented in those trials. Further, the global response to covid-19 has seen numerous treatments provided off-label or through emergency access provisions in parallel with ongoing RCTs, raising the question of what can credibly be learned from the experiences of patients receiving these treatments outside of RCTs.

This study employs the stability controlled quasi-experiment (SCQE)[1, 2] approach to investigate treatment effects on covid-19 patients, which differs from conventional approaches for observational studies in two key ways. First, unlike standard covariate-adjustment strategies (regression, matching, weighting, and stratification), SCQE does not rely on the assumption that there are no unobserved confounders, i.e. that the treated and untreated groups are comparable after accounting for observed covariates. Instead, SCQE produces estimates that depend only on what the user is willing to assume about the *baseline trend*, here meaning changes in the covid-19 mortality rates from one cohort to another that are not caused by the treatment in question. Second, whereas conventional approaches present a single estimate and confidence interval that is correct only under the assumption of no unobserved confounding and no other sources of bias, SCQE is designed to show the user the entire range of estimates obtained over a plausible range of assumptions about this baseline trend. These results can be restated to reveal *what assumptions about the baseline trend in mortality would have to be defended* in order to argue that the treatment was beneficial, null, or harmful. Such an exercise avoids reliance on narrow assumptions. Yet, as illustrated here, it can be informative about the range of plausible effects of a treatment.

Using electronic health records from a university hospital cohort of over 2007 patients admitted over 200 days, we apply SCQE to investigate what can be concluded regarding the impact of hydroxychloroquine and dexamethasone on mortality in patients with covid-19.

METHODS

Approach and assumptions

To build intuition for the SCQE approach, let us consider a “natural experiment” that leverages changes in treatment prevalence over time. Suppose there are two cohorts of patients. In the first, no patients have access to a given treatment, and mortality is 20%. In another cohort (e.g., taken from a later period of time at the same facilities), 50% of patients are administered a new treatment. They do so not at random, but based on patient and physician judgement and choice. Suppose the overall mortality rate in the second cohort is 15%. With an assumption that the two cohorts of patients are comparable (i.e. they would have the same average outcomes, absent treatment differences) we can estimate that being in the second ("high-use") cohort reduced mortality by 5 percentage points. Further, since all this benefit comes from the half of patients who opted to take treatment, the benefit per treated patient must be twice that (i.e. a 10 percentage point benefit per treated patient). Note that the required assumption here regards comparability of the cohorts, and not comparability of the treated to the untreated within either the first or the second cohort. This is beneficial as we acknowledge that treatment decisions can be made in part due to unobservable factors, making the treated and control groups incomparable regardless of efforts to adjust for all measured or observed variables.

Such an approach, however, is limited by the assumption that the two cohorts would have the same average mortality rate, absent changes in the treatment. The SCQE takes the more flexible position of allowing the cohorts to differ in this regard by variable, postulated degrees. That is, we allow for some “baseline trend” that describes how differently the cohorts would have fared on their average outcomes, if not for changes in the treatment in question. Equivalently, this baseline trend can be defined as the difference in average outcomes between cohorts that we would have seen if no patients in either cohort had used the treatment. For instance, if treatment changes (other than the one in question) and/or changes in the composition of the cohorts would have generated a mortality rate that was 2 percentage points lower in the later cohort than the earlier one, the baseline trend for that analysis would be -2 percentage points.

The key mathematical fact is that in this case, *for any assumed baseline trend, we can estimate the treatment effect experienced by the treated patients*, without additional assumptions or covariates[1, 2]. For intuition behind this result, we return to the natural experiment considered above, where no individual in the first cohort takes the treatment, and hence the average outcome we observe is the “average non-treatment outcome”.[1] If we add to this the assumed baseline trend, we obtain the average non-treatment outcome in the second cohort, i.e. the outcome we would expect if we could see how all individuals in this cohort fared absent the treatment (regardless of whether they actually took the treatment).

Algebraically, this average non-treatment outcome over *everybody* in the second cohort is the sum of two terms: (i) the average non-treatment outcome we observe from the untreated patients in this cohort, times the proportion that were untreated, and (ii) the (unobservable) average non-treatment outcome that the treated would have had, times the proportion that were treated. Since the average non-treatment outcome for the treated is the only unknown in this equation, we can solve for this quantity (see [1] for a complete description employing mathematical notation). Next, the (observed) average treatment outcome for the treated minus this average non-treatment outcome for the treated is the average treatment effect among the treated (ATT). Figure 1 illustrates this reasoning graphically, using values similar to those from the dexamethasone study below.

[Figure 1 about here.]

Finally, rather than place our confidence in a single assumption, we “invert” the analysis to reveal *the needed assumptions about the baseline trend in mortality to declare that a given treatment had a beneficial, null, or harmful effect.* Note that (95%) confidence intervals can be constructed for the effect estimate at any given choice of the baseline trend assumption, using the approach described in [2]. We describe an effect as a significantly or detectably “beneficial” or “harmful” estimated effect when its 95% confidence interval excludes zero, which is equivalent to a two-sided p-value at or below 0.05.

While no analysis can determine the true value of the baseline trend, beliefs about this quantity can be defended or challenged through auxiliary analyses, such as examining the change in the composition and risk factors of the patients in the two cohorts and changes in any other documented treatment practices. We consider what baseline trends can be deemed plausible or implausible in the Discussion below.

While we have described the approach in its simplest form, several extensions are important, some employed here. First, we need not have a cohort with zero use of the treatment, just two cohorts with sufficiently different levels of treatment.[2] Second, the two cohorts do not actually need to be cohorts separated by time; they could be cohorts from separate hospitals, for example. We need only be able to consider how widely the two cohorts may have differed in their average outcomes, had treatment levels not differed between the cohorts. Third, while we employ individual observations for the analysis and a range of auxiliary variables that are an aid to validating the approach, the SCQE can be used to estimate effects where we are only given average outcomes and the proportion treated in two cohorts.[3]

Data Collection

Data were extracted from the electronic medical records for a multicenter hospital system including an academic tertiary referral hospital. Hospital courses were identified based on a documented covid-19 infection indicated by either recorded diagnosis or identification of a positive PCR test for the SARS-CoV-2 virus. Data were extracted for all persons with hospitalizations that began between 3/8/2020 and 10/7/2020. Where patients had multiple hospitalizations within 14 (or 28) days of the index hospitalization following diagnosis, analysis included these extra hospitalizations.

Clinical data extracted included demographic factors (age, sex, race/ethnicity),and body mass index (BMI) at or prior to the period of hospitalization, baseline laboratory assessments (white blood cell count, C-reactive protein, ferritin, procalcitonin), medication use (remdesivir, convalescent plasma, hydroxychloroquine, dexamethasone, prednisone, methylprednisolone, hydrocortisone, and use of proning for assistance with ventilatory support. We additionally extracted whether the patient had been admitted by transfer from a skilled nursing facility, and disposition at discharge.

Use of dexamethasone and hydroxycholoroquine, our treatments of interest, were defined as any use during the hospital stay(s). Hydroxychloroquine was administered as a standard 5-day course. For Dexamethasone, the prescribed course was variable in the few cases given in the low-use cohort (days

[1] Borrowing from the potential outcomes framework, we can conceptualize a patient’s outcome both had they taken the treatment (their treatment outcome) and had they not (their non-treatment outcome), regardless of their actual treatment status. An "average non-treatment outcome" for a cohort, then, is the average outcome we would observe had no patients received treatment.

[2] This does change the interpretation of effect in terms of the population of patients to whom it applies; see [2].

[3] Analyses using only aggregate data of this kind can be conducted using our web-based software available at https://amiwulf.shinyapps.io/SCQE_demo/.

44-101). Its prescription in the high-use cohort became more standardized as a result of its potential effectiveness [3], typically administered at 6mg for 10 days.

Cohort Construction

Hydroxychloroquine. Initially, hydroxychloroquine was widely used, given to 62% of patients admitted in the first two weeks. Usage then began to fall steadily, with fewer than 2% of patients admitted in week 7 or later receiving it. Cohorts were constructed based on patients' day of admission. Data from all days (1 to 200) were first split into two cohorts based on the split-point that would maximize the strength of relationship between cohort and probability of receiving hydroxychloroquine, as judged by the F-statistic, which occurred at day 44. Next, the second cohort was trimmed, to avoid covering a period in which dexamethasone use rose. Ending the second cohort on day 82 minimized the difference in proportion of patients receiving dexamethasone in the first and second cohorts.When used in either cohort, hydroxychloroquine was administered as a standard 5-day course.

Dexamethasone. Use of dexamethasone began low and remained at 5% or lower for the first 15 weeks, after which it steadily rose and peaked near 50% in week 21. Cohort construction proceeded by first choosing the split date that maximized the difference between dexamethasone use in the two cohorts, as judged by the F-statistic, which occurred at day 102. We then trimmed the first cohort to begin on day 44, ensuring little change in hydroxychloroquine usage between the cohorts. Dexamethasone use in the first, low-use cohort was largely unstandardized. The transition to higher use in the later cohort was driven partly by promising preliminary trial results [3], after which dosing become more standardized with most patients receiving 6mg once daily for up to 10 days.

RESULTS

Little evidence supporting beneficial role for Hydroxychloroquine

The "high-use" (first) cohort included 766 patients admitted from day 1 to 43, of which 36% used hydroxychloroquine, and a "low-use" (second) cohort of 548 patients admitted between days 44 and 82 of which only 2.9% used hydroxychloroquine. The F-statistic for difference in hydroxychloroquine use between cohorts was 242.3, p<1e-15). Mortality at 14-days was 11.6% in the high-use cohort (89/766) and 8.6% (47/548) in the low-use cohort, for a raw risk difference (RD) of 3 percentage points (t=1.79, p=0.07).

[Figure 2 about here.]

Figure 2 shows the results for hydroxychloroquine. The vertical axis shows different assumptions regarding the baseline trend, i.e. mortality shifts absent changes in hydroxychloroquine use. These are shown in terms of the mortality change going from the low-use to high-use cohorts, and because the high-use cohort came first and the low-use came second, a value of 0.02, for example, reflects an improvement (decrease) over time in mortality by 2 percentage points.

We find that hydroxychloroquine can only be claimed to have had a benefit if baseline mortality decreased between the first and second cohort by 6.4 percentage points. In other words, one must argue there was a 55% reduction in covid-19 mortality among inpatients in this short time, due to factors other than changes in hydroxychloroquine use. Second, hydroxychloroquine is harmful at the $p < 0.05$ level if baseline mortality instead worsened over-time by just 0.3 percentage points (2.6% of the original 11.6% mortality) or more. At this boundary the point estimate for hydroxychloroquine is roughly a 10 percentage point increase in mortality. For all baseline trend assumptions in between these, we would not reject the null hypothesis of zero effect.

We also consider 28-day mortality for comparability with existing studies. For hydroxychloroquine to have been significantly beneficial would require that baseline mortality improved by 6.8 percentage points, a 47% drop from the first cohort's 28-day mortality of 11.5%. Hydroxychloroquine would prove harmful at the $p < 0.05$ level if baseline mortality rose by 0.7 percentage points.

Probing possible trends. In assessing the plausibility of different baseline trends it is useful to examine possible changes in the composition of the cohorts and in the treatments provided. Table 1 describes these cohorts in terms of characteristics determined prior to or very shortly after admission (A), the treatments received (B), and the predicted risk of mortality according to a range of models (C). As the purpose of such comparisons is to inform our beliefs about the plausible range of baseline mortality differences between the cohorts absent hydroxychloroquine, statistical inferences regarding the comparisons are irrelevant.

Looking first at patient characteristics prior to or shortly after admission (A), we see the two cohorts are similar overall, particularly on known risk factors such as age, gender, weight, and BMI. The proportion who identify as Hispanic rises somewhat, from 38% to 49%. Taken alone, and given documented differences in outcomes in Hispanic patients, this would contribute towards an upward shift in baseline mortality risk over time. Similarly, the fraction of patients coming from skilled nursing facilities rises somewhat (from 4% to 9%), which could also increase baseline mortality in the second cohort.

Table 1 Comparison of hydroxychloroquine cohorts

	Cohort means:	
A. Characteristics	High-use	Low- use
age (years)	58	56
over 65 y.o.	36%	33%
female	42%	46%
ethnicity: Hispanic	38%	49%
weight (Lb)	194	187
BMI (kg/m^2)	31	32
from skilled nursing facility	4%	9%
ICU in first 24h	18%	15%
CRP (mg/L)	115	108
WBC (per mcL)	7.72	8.63
ferritin (μg/L)	747	601
procalcitonin (ng/mL)	0.89	1.11

B. Other treatments	High Use	Low Use
remdesivir	3%	11%
tocilizumab	7%	3%
convalescent plasma	5%	22%
proning	3%	4%
dexamethasone	4%	5%
methylprednisolone	10%	7%
prednisone	1%	2%
hydrocortisone	3%	4%
nitazoxanide	1%	0%

C. Modeled risk of 14-day mortality	High Use	Low Use
linear model (pre-tx)	10.5%	9.8%
linear model (all)	10.9%	9.2%
KRLS model (pre-tx)	10.2%	9.5%
KRLS model (all)	10.4%	9.1%

Note: Comparison of cohorts with high (left) or low (right) use of hydroxychloroquine, considering (A) various patient characteristics, (B) other treatments received, and (C) model-estimated risk of 14-day mortality. Lab measures (CRP, WBC, ferritin, procalcitonin) refer to the first measurement taken.

Recall that, because the high use cohort precedes the low use cohort, potential increases in baseline mortality over time (as might be caused by these changes) represent negative baseline trends (i.e. moving downwards on Figure 2), and lead to more harmful estimated effects of hydroxychloroquine.

On the other hand, two treatment practices (B) that could have potentially improved mortality increased over time between these cohorts as well: remdesivir (from 3% to 11%) and convalescent plasma (from 5% to 22%). Were these treatments to improve mortality, they would encourage us to consider possible improvements in mortality over time, moving upwards on Figure 2. However, the change in baseline mortality due to these alone would not likely be large given the low usage rates. Suppose that nearly all of the 11.6% 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. Suppose these drugs, in any combination, reduce mortality by 30%. This would reduce mortality by 3.5 percentage points overall. Assuming a baseline trend of 3.5 would then be generous, given these assumptions and that other factors such as ethnicity suggest mortality change in the opposite direction. Yet, even at an assumed baseline trend of 3.5, hydroxychloroquine would not prove significantly beneficial.

Finally, these differences in the cohorts are important only insofar as they suggest different baseline mortality rates. Using simple linear probability models (C), we can predict 14-day mortality using only patient characteristics prior to treatment ("Linear model (pre)"), or using those characteristics plus information on treatments ("Linear model (all)") (see Supplement for details of all models). The same predictions can instead be made using a more flexible and powerful machine learning model, kernel-regularized least squares (KRLS [4]). These models are reasonably predictive: the linear model with all variables explains 17% of the variation in mortality; the KRLS model with all variables explains 52%. Yet, the overall risk levels in the two cohorts appears similar, as shown in Table 1. The low-use (second) cohort has slightly lower risk by 0.7 to 1.7 percentage points. Such model estimates only inform the range of plausible baseline trends considered. If we consider, for example, a 1 percentage point drop in baseline mortality (a baseline trend represented by .01 on Figure 2), this corresponds to a non-significantly harmful increased risk of 6 percentage points (95% CI=[-0.04, 0.16]).

Dexamethasone: plausibly beneficial with very low risk of harm

In the low-use (first) cohort, 5.7% (35/614) were given dexamethasone, and the 14-day mortality rate was 8.1% (50/614). In the high-use (second) cohort, 46% (287/622) of patients were given dexamethasone, and the 14-day mortality rate was 4.0% (25/622).

In terms of *ex ante* plausible baseline trends, it would be difficult to support claims that baseline mortality dropped or increased by more than perhaps 50% (4.1 percentage points, either direction). Given the possibility that treatment practices are otherwise improving over time (e.g. preferences for non-invasive oxygen supplementation, improved ventilator management, or other treatments being attempted), we might expect some decrease in mortality (a negative trend).

[Figure 3 about here.]

SCQE formalizes the simple logic that while mortality fell in the higher-use cohort, this only implies a benefit of dexamethasone if mortality would not have improved too greatly "on its own." In Figure 3, the vertical axis again represents different postulated baseline trends, ie how much higher baseline mortality is assumed to be in the high-use than in the low-use cohort. Because the transition from low-use to high-use for dexamethasone is now the transition from the earlier to later cohort, positive trends indicate higher (worse) baseline mortality over time.

SCQE finds that we would conclude dexamethasone had a significant benefit ($p < 0.05$) if mortality was increasing, flat, or going down by as much as 1.5 percentage points for reasons other than dexamethasone use. Unlike the conditions required for hydroxychloroquine to be beneficial, this window includes a range of plausible baseline trends. Consequently, we conclude there is a reasonable possibility that dexamethasone has a benefit, though it remains far from defensible with certainty. This is particularly useful as weighed against the risks: for dexamethasone to be significantly harmful at the $p < 0.05$ level, baseline mortality would have to have improved by 6.8 percentage points, which would bring mortality to 1.3%. This can nearly be ruled out as there is no reason to believe any changes in the composition of patients or in other treatments made available could have reduced mortality this dramatically.

Regarding the secondary outcome of 28-day mortality, results are again similar. Dexamethasone proves statistically beneficial at the $p < 0.05$ level so long as mortality rose, stayed flat, or fell by as much as 2.3 percentage points. Further, to prove harmful (at the $p < 0.05$ level), baseline mortality would have to drop by 8.7 percentage points. Given the first cohort 28-day mortality rate of 10.9%, this would mean arguing that mortality was reduced to just 2.2% in the second cohort for reasons other than dexamethasone use.

Probing possible trends. Table 2 aids reasoning about possible baseline trends by comparing the high- and low-use cohorts on numerous characteristics.

Table 2 Comparison of dexamethasone cohorts

A. Characteristics	High Use	Low Use
age (years)	55	55
over 65 y.o.	33%	32%
female	53%	47%
ethnicity: Hispanic	46%	50%
weight (Lb)	195	187
BMI (kg/m^2)	32%	32%
from skilled nursing facility	1%	8%
ICU in first 24h	8%	14%
CRP (mg/L)	101	109
WBC (per mcL)	8.24	8.79
ferritin (μg/L)	522	596
procalcitonin (ng/mL)	0.73	1.08
B. Other treatments	**High Use**	**Low Use**
hydroxychloroquine	1%	3%
remdesivir	28%	12%
tocilizumab	2%	3%
convalescent plasma	21%	22%
proning	1%	3%
methylprednisolone	3%	7%
prednisone	0%	2%
hydrocortisone	2%	4%
nitazoxanide	0%	0%
C. Modeled risk of 14-day mortality	**High Use**	**Low Use**
linear model (pre-tx)	5.0%	7.0%
linear model (all)	4.6%	7.3%
KRLS model (pre-tx)	5.0%	6.4%
KRLS model (all)	4.4%	6.8%

Note: Comparison of cohorts with high and low use of dexamethasone. Lab measures (CRP, WBC, ferritin, procalcitonin) refer to the first measurement taken.

Most differences between the cohorts are small and do not revise the range of baseline trends we can consider plausible. One worrying exception is remdesivir, with increased usage (from 12% to 28%) alongside dexamethasone. Remdesivir's effectiveness in reducing mortality remains uncertain, with the ACTT-1 trial[5] showing a benefit on time to recovery, while preliminary reports from the WHO Solidarity trial[?] show no significant mortality benefit. Nevertheless we must consider how this might affect the appropriate range of baseline trends. Even if remdesivir reduced mortality by 20 percentage points, then the increase in usage from 12% to 28% would suggest a drop in the baseline mortality by 3.2 percentage points. If we took this to be the baseline trend (-0.032), it would suggest a benefit of dexamethasone that no longer reaches significance (RD = -0.023, 95% CI=[-.088, .043]).

Looking to models of mortality risk, in every case the predicted risk of mortality instead fell going into the second (high-use) cohort, by 1.4-2.7 percentage points. The reduced risk forecasted by these models is due in part to the reduced proportion of patients transferred from skilled nursing facilities. If the baseline trend was believed to be approximately a two percentage points drop as suggested by these models, the corresponding effect estimate for dexamethasone would be a risk difference of -0.05 (95% CI=[-0.12, 0.01]). In summary, dexamethasone could very plausibly have had either a beneficial or a null effect, while we can nearly rule out that it had a harmful one.

DISCUSSION

Our study shows what can be inferred about the effects of hydroxychloroquine and dexamethasone use on covid-19 mortality using only electronic health records from outside of randomized trials.

Several considerations aid in gauging what baseline trends are (im)plausible or (im)probable, and hence the conclusions that can be supported. Studies of overall mortality in other populations show substantial decreases over time. For example [6] show large decreases in mortality in mid-April to May as compared to March among critical care patients in England which they argue are not due to changes in patient demographics. Such results do not generalize easily to our analysis given differences in the population, time period, outcomes, and most importantly that these reflect overall mortality inclusive of changes in treatments like dexamethasone, not the baseline mortality trends we require. In fact, [6] argue changes in mortality may be partly due to treatments employed in the RECOVERY trial, such as dexamethasone. Nevertheless there are numerous reasons to expect improvements in baseline mortality in our sample as well due to changes in other various treatment practices over time. Though the cohorts we compared had similar exposure to most therapies (Tables 1 and 2), changes in treatment practice that remain unobserved to us could have led to improvements, notably improvements in the timing and management of ventilatory support. Given such possibilities, the reduced mortality seen in other settings, and the otherwise similar demographics and estimated mortality risk in these cohorts, we would judge small increases in baseline mortality to be unexpected but still possible, while we judge large increases in baseline mortality—say by 20% or more—to be extremely unlikely.

It is more difficult to say how large a drop in baseline mortality would be too large to be believed. For many other, longer-running disease, it might be reasonable to suggest baseline mortality drops by no more than perhaps 10% in a matter of months. For covid-19 however, given rapid changes, a much more generous bound is required. However, given information about about treatment practices in this health system (two of the authors are physicians there), we do not expect any otherwise undocumented highly effective treatment was initiated and widely used in this period. While would argue that a 50% drop in baseline mortality can neither be ruled out nor defended with certainty, while we regard a drop of 80% or more to be highly improbable.

In the case of hydroxychloroquine, mortality rate decreased as hydroxychloroquine use decreased. It would be a mistake to conclude from this alone that hydroxychloroquine was harmful, as such an inference depends on how mortality would have changed anyway, i.e. the baseline trend. For example, if mortality would have fallen even faster absent the drop in hydroxychloroquine usage, then the observed data would be consistent with a benefit of hydroxychloroquine. Specifically, hydroxychloroquine is demonstrably beneficial only if baseline mortality would have improved from the earlier to later cohort by 6.4 percentage points (55%). This is possible, but we must accept that it is far from confidently defensible, and certainly not supported by the modeled changes in mortality risk in these cohorts. Further, against the difficulty of defending a beneficial effect, one must consider the risk that hydroxychloroquine was harmful. If mortality worsened from one cohort to the next by even 0.3 percentage points, then hydroxychloroquine must have had a statistically significant harmful effect. Notably, these results are consistent with evidence from randomized trials testing hydroxychloroquine for early treatment of mild covid-19 in adults [7], for reduced mortality among hospitalized patients (RECOVERY trial[8]), or prophylactic protection against infection among exposed participants [9], all of which concluded hydroxychloroquine had null or potentially harmful effects on their varied outcomes (see Supplement for a review of observational studies).

SCQE shows that dexamethasone was significantly beneficial, if baseline mortality was increasing over time between cohorts, stayed flat, or fell by up to 22% (1.5 percentage points). Such baseline trends are far from certain, but are certainly plausible and would not be ruled out. This potential benefit is weighed against the risk of harm. Here the results are rather definitive: statistically significant evidence of harm requires that baseline mortality improved between cohorts by at least 6.8 percentage points, which amounts to an 84% or a baseline mortality of just 1.3% in the later cohort. We regard as highly unlikely given the small differences between the cohorts and our belief that no undocumented but highly effective treatment had been discovered and widely used in the second cohort. Our results are consistent with, though more reserved than, conclusions drawn from the CoDEX trial[10] showing increased days alive without mechanical ventilation and the RECOVERY trial [3] showing lower mortality, specifically for those under mechanical ventilation or with oxygen supplementation at randomization.

Limitations

The central limitation of this study and approach is also its strength: that it avoids providing a narrow claim, because it avoids reliance upon a point assumption that is unlikely to be defensible. This may remain unsatisfying for many readers accustomed to more specific claims, though it serves to transparently communicate what can be claimed subject to what assumption, leaving the reader to argue positively for the assumptions that would be required to reach a conclusion and illustrating the limits of our knowledge.

Another limitation, specific to this study, regards sample size. The sample is larger than those in some randomized trials, and more than sufficient for SCQE. The downside of the modest sample size, however, is that the estimate effect has to be relatively large (roughly 10 percentage points or more) for the 95% CI to exclude zero. This in turn means that our conclusions will be less decisive over a given plausible range of baseline trends than they may have been with similar estimates but a larger sample. We note that a variant of the SCQE approach used here can be used to re-analyze observational studies, many of which are much larger, as demonstrated for existing studies of hydroxychloroquine in the Supplement.

Finally, in both of the studies, the cohorts we defined were largely similar in their composition and use of other treatments, which is not necessary but makes it far easier to reason about plausible bounds on the baseline mortality difference between cohorts. That said, differences in the use of remdesivir remain non-trivial in both cases, with convalescent plasma use also changing in the hydroxychloroquine study. We have discussed the degree to which these could influence the baseline mortality difference, and what this means for our estimates. Yet, the existence of these changes is a nuisance that widens the range of plausible estimates and thus reduces the chances of a more decisive conclusion being reached. A promising option suitable in some contexts for future research would be a "design-based" version of the SCQE in which hospitals plan to make a new treatment available, again by choice rather than as part of an RCT, while intentionally limiting other changes in practice or patient composition over a period of time around this transition. To the degree this is feasible it can buoy arguments for baseline trends or smaller magnitude, resulting in a narrower range of plausible effect estimates preserving the ability to offer patients and doctors choice in treatment.

Conclusions

Our results are largely consistent with those of existing trials on hydroxychloroquine and dexamethasone, despite examining outcomes for patients outside of randomized trials using only electronic health records. This study provides not only corroborative evidence from other populations regarding these treatments, but also a useful and accessible application of the SCQE approach to aid adoption.

More broadly, as observational studies are likely to remain part of the research landscape, the use of SCQE can offer a valuable, rigorous way to understand what can be safely learned from patient experiences with non-randomized treatments. SCQE estimates may be particularly useful prior to the availability of data from randomized trials, or in domains such as quality-improvement studies in which randomized trials are not always performed. This approach can also complement randomized trials, as evident here, by offering corroborating evidence and assessing efficacy in a population that will often differ from those enrolled in trials.

We endorse the argument that even—or especially—in moments of urgency such as a pandemic, every effort should be made to launch and complete coordinated, well-designed randomized trials [11]. Nevertheless, there remains an important role for credible observational studies that avoid risks of producing misleadingly confident results built on fragile assumptions.

Numerous opportunities remain to apply this approach to covid-19 therapeutics in development, including convalescent plasma, monoclonal antibodies, and additional antivirals and anti-inflammatory agents currently being used experimentally. At a time when ongoing randomized trials often coexist with parallel access to experimental therapies under expanded access provisions, the reduced uptake of randomized trials may additionally increase the importance of methodological frameworks such as SCQE to evaluate observational data.

REFERENCES

[1] Hazlett C. Estimating causal effects of new treatments despite self-selection: The case of experimental medical treatments. Journal of causal inference. 2019;7(1).

[2] Hazlett C, Maokola W, Wulf DA. Inference without randomization or ignorability: A stability controlled quasi-experiment on the prevention of tuberculosis. Statistics in Medicine. 2020;First online, September.

[3] Group RC. Dexamethasone in hospitalized patients with Covid-19—preliminary report. New England Journal of Medicine. 2020;.

[4] Hainmueller J, Hazlett C. Kernel regularized least squares: Reducing misspecification bias with a flexible

and interpretable machine learning approach. Political Analysis. 2014;p. 143–168.

[5] Beigel JH, Tomashek KM, Dodd LE, Mehta AK, Zingman BS, Kalil AC, et al. Remdesivir for the treatment of Covid-19. New England Journal of Medicine. 2020;.

[6] Dennis J, McGovern A, Vollmer S, Mateen BA. Improving COVID-19 critical care mortality over time in England: A national cohort study, March to June 2020. medRxiv. 2020;.

[7] Mitjà O, Corbacho-Monné M, Ubals M, Tebe C, Peñafiel J, Tobias A, et al. Hydroxychloroquine for early treatment of adults with mild Covid-19: a randomized-controlled trial. Clinical Infectious Diseases. 2020;.

[8] Horby P, Mafham M, Linsell L, Bell JL, Staplin N, Emberson JR, et al. Effect of Hydroxychloroquine in Hospitalized Patients with COVID-19: Preliminary results from a multi-centre, randomized, controlled trial. MedRxiv. 2020;.

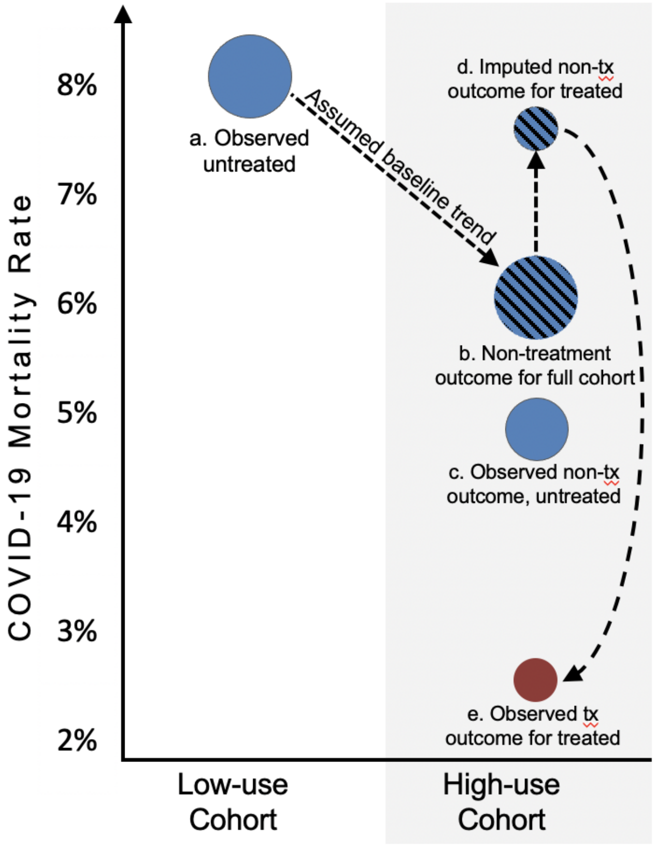
[9] Boulware DR, Pullen MF, Bangdiwala AS, Pastick KA, Lofgren SM, Okafor EC, et al. A randomized trial of hydroxychloroquine as postexposure prophylaxis for Covid-19. New England Journal of Medicine. 2020;.

[10] Tomazini BM, Maia IS, Cavalcanti AB, Berwanger O, Rosa RG, Veiga VC, et al. Effect of dexamethasone on days alive and ventilator-free in patients with moderate or severe acute respiratory distress syndrome and COVID-19: the CoDEX randomized clinical trial. Jama. 2020;.

[11] Lane HC, Fauci AS. Research in the Context of a Pandemic. Mass Medical Soc; 2020.

LIST OF FIGURES

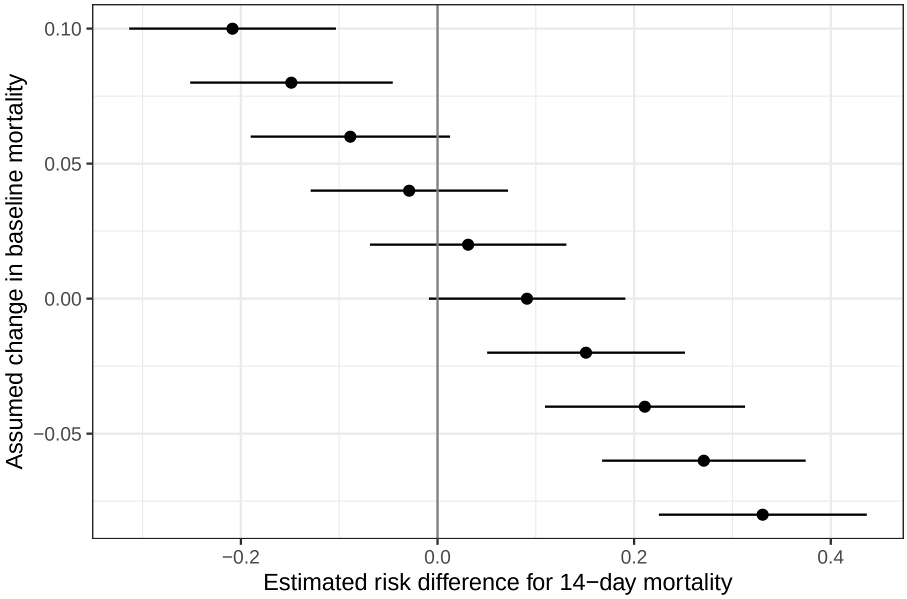
Figure 1 Understanding the SCQE



COVID-19 Mortality Rate
8%
7%
6%
5%
4%
3%
2%
a. Observed untreated
Assumed baseline trend
d. Imputed non-tx outcome for treated
b. Non-treatment outcome for full cohort
c. Observed non-tx outcome, untreated
e. Observed tx outcome for treated
Low-use Cohort
High-use Cohort

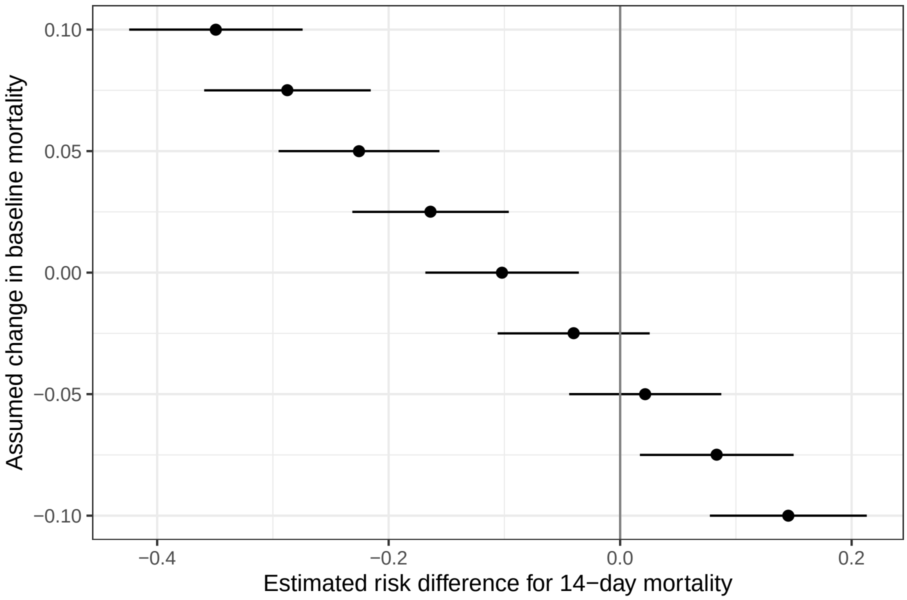
**Note:* Each ball represents a group, and the height represents that group's average outcome. Starting on the left, in the no-use cohort we observe the average mortality (8%) under non-treatment. We then impose an assumption regarding how the non-treatment outcome would have changed from one cohort to the next. Here this is a 2 percentage point drop, meaning the average non-treatment outcome over the entire high-use cohort, is assumed to be 6% (b). Because the value of (b) is weighted some of the non-treatment outcome for those who were not-treated and those who were treated, we can solve algebraically for the average non-treatment outcome that would have been experienced by the treated (d). Comparing the actual average (treatment) 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 the comparability of the treated and control (c and e) is made, only an assumption on the trend in the average non-treatment outcome.

Figure 2 SCQE Estimates of risk difference for hydroxychloroquine



**Note:* The vertical axis indicates an assumption about the baseline trend in mortality, i.e. how mortality is postulated to have changed going from the low-use to high-use cohorts, for reasons other than changes in hydroxychloroquine use. Because the high-use cohort is the earlier one here, positive values (towards the top of the figure) correspond to falling mortality in the direction of time. At each postulated mortality trend, we see the consequent effect estimate and its 95% confidence interval.

Figure 3 SCQE Estimates of risk difference for dexamethasone



**Note:* The vertical axis indicates an assumption about the baseline trend in mortality, i.e. how mortality is postulated to have changed going from the low-use to high-use cohorts, for reasons other than changes in dexamethasone use. Because the high-use cohort is now the earlier one, positive values (towards the top of the figure) correspond to increases in mortality over time. At each postulated mortality trend, we see the consequent effect estimate and its 95% confidence interval.