Agreement of treatment effects from observational studies and randomized controlled trials evaluating hydroxychloroquine, lopinavir-ritonavir, or dexamethasone for covid-19: meta-epidemiological study

Osman Moneer,[1] Garrison Daly,[2] Joshua J Skydel,[3] Kate Nyhan,[4,5] Peter Lurie,[2] Joseph S Ross,[6,7,8] Joshua D Wallach[5]

For numbered affiliations see end of the article

Correspondence to: J D Wallach
Joshua.wallach@yale.edu
(or @JoshuaDWallach on Twitter
ORCID 0000-0002-2816-6905)

Additional material is published online only. To view please visit the journal online.

Cite this as: *BMJ* 2022;377:e069400
http://dx.doi.org/10.1136/bmj-2021-069400

Accepted: 02 March 2022

ABSTRACT

OBJECTIVE

To systematically identify, match, and compare treatment effects and study demographics from individual or meta-analysed observational studies and randomized controlled trials (RCTs) evaluating the same covid-19 treatments, comparators, and outcomes.

DESIGN

Meta-epidemiological study.

DATA SOURCES

National Institutes of Health Covid-19 Treatment Guidelines, a living review and network meta-analysis published in *The BMJ*, a living systematic review with meta-analysis and trial sequential analysis in *PLOS Medicine* (The LIVING Project), and the Epistemonikos "Living OVerview of Evidence" (L·OVE) evidence database.

ELIGIBILITY CRITERIA FOR SELECTION OF STUDIES

RCTs in *The BMJ*'s living review that directly compared any of the three most frequently studied therapeutic interventions for covid-19 across all data sources (that is, hydroxychloroquine, lopinavir-ritonavir, or dexamethasone) for any safety and efficacy outcomes. Observational studies that evaluated the same interventions, comparisons, and outcomes that were reported in *The BMJ*'s living review.

DATA EXTRACTION AND SYNTHESIS

Safety and efficacy outcomes from observational studies were identified and treatment effects for dichotomous (odds ratios) or continuous (mean differences or ratios of means) outcomes were calculated and, when possible, meta-analyzed to match the treatment effects from individual RCTs or meta-analyses of RCTs reported in *The BMJ*'s living review with the same interventions, comparisons, and outcomes (that is, matched pairs). The analysis compared the distribution of study demographics and the agreement between treatment effects from matched pairs. Matched pairs were in agreement if both observational and RCT treatment effects were significantly increasing or decreasing ($P<0.05$) or if both treatment effects were not significant ($P\geq0.05$).

RESULTS

17 new, independent meta-analyses of observational studies were conducted that compared hydroxychloroquine, lopinavir-ritonavir, or dexamethasone with an active or placebo comparator for any safety or efficacy outcomes in covid-19 treatment. These studies were matched and compared with 17 meta-analyses of RCTs reported in *The BMJ*'s living review. 10 additional matched pairs with only one observational study and/or one RCT were identified. Across all 27 matched pairs, 22 had adequate reporting of demographical and clinical data for all individual studies. All 22 matched pairs had studies with overlapping distributions of sex, age, and disease severity. Overall, 21 (78%) of the 27 matched pairs had treatment effects that were in agreement. Among the 17 matched pairs consisting of meta-analyses of observational studies and meta-analyses of RCTs, 14 (82%) were in agreement; seven (70%) of the 10 matched pairs consisting of at least one observational study or one RCT were in agreement. The 18 matched pairs with treatment effects for dichotomous outcomes had a higher proportion of agreement (n=16, 89%) than did the nine matched pairs with treatment effects for continuous outcomes (n=5, 56%).

CONCLUSIONS

Meta-analyses of observational studies and RCTs evaluating treatments for covid-19 have summary

WHAT IS ALREADY KNOWN ON THIS TOPIC

Randomized controlled trials are generally considered to be the gold standard for studying safety and efficacy of treatments, but can have limited generalizability and require a substantial amount of time to plan and complete

The covid-19 pandemic has highlighted the potential role of observational studies to provide insight into the clinical value of candidate treatments, although concerns have been raised about the rapid dissemination of potentially low quality evidence

Little is known about the agreement between individual or meta-analyzed observational studies and randomized controlled trials evaluating the same covid-19 treatments, comparators, and outcomes

WHAT THIS STUDY ADDS

Of the matched observational study and RCT pairs comparing hydroxychloroquine, lopinavir-ritonavir, or dexamethasone to an active or placebo comparator for any safety or efficacy outcomes of covid-19, more than three quarters had treatment effects that were in agreement

Overall, agreement was higher in matched pairs of meta-analyses of observational studies and meta-analyses of RCTs (82%) and in those evaluating treatment effects for dichotomous outcomes (89%), than in those of only one observational study and/or one RCT (70%) and in those evaluating treatment effects for continuous outcomes (56%), respectively

Despite concerns about evidence from individual observational studies evaluating covid-19 treatments, meta-analyzed evidence from observational studies can complement, but should not replace, evidence collected from randomized controlled trials

treatment effects that are generally in agreement. Although our evaluation is limited to three covid-19 treatments, these findings suggest that meta-analyzed evidence from observational studies might complement, but should not replace, evidence collected from RCTs.

Introduction

The covid-19 pandemic has necessitated the rapid generation of evidence to better characterize the benefits and harms of therapies available for its treatment. Randomized controlled trials (RCTs) are generally considered to be the gold standard for determining therapeutic safety and efficacy.[1] However, despite numerous strengths, including a trial's role in minimizing the influence of bias and confounding factors, RCTs have important limitations that can undermine their generalizability to real world clinical practice.[2] In particular, trials tend to have strict inclusion and exclusion criteria and are subject to recruitment difficulties, which can be associated with small sample sizes and short follow-up durations.[3] In addition, they are logistically challenging and expensive, sometimes taking years to plan and complete.[2] These limitations and the urgent nature of the pandemic have highlighted the potential role of observational studies, including those that use real world data to provide insight into the clinical value of candidate treatments.[4] Although some rigorously designed studies that used real world data for other medical conditions have replicated the results obtained from RCTs across various conditions studied,[5-8] other studies have suggested poor agreement between the treatment effects from randomized and non-randomized studies.[9-11]

Since March 2020, the clinical and public health communities have searched for effective and safe interventions by conducting thousands of observational studies on potential covid-19 treatments while awaiting the results from ongoing and planned RCTs. However, concerns have been raised about the rapid dissemination of potentially low quality studies,[12] with prominent retractions in high impact journals undermining the confidence in observational evidence.[13 14] To explore the role of observational studies for future pandemic decision making, our first objective was to systematically identify, match, and compare the agreement between treatment effects and study demographics from observational studies and RCTs evaluating the same interventions, comparators, and outcomes in studies evaluating therapeutics for covid-19. We also aimed to evaluate the consistency of results across the number of studies available for each comparison (that is, individual *v* meta-analyzed treatment effects) and outcome types (dichotomous *v* continuous).

Methods

Identification of covid-19 interventions

Given the large number of covid-19 interventions (eg, drugs, biologics, and procedures) that have been evaluated across thousands of studies, our a priori established approach was to limit our evaluation to three prominent interventions for which multiple RCTs and observational studies were likely. This approach also ensured the feasibility of conducting new meta-analyses of observational studies across all potential clinical outcomes.

To assemble a list of potentially eligible therapeutic interventions for covid-19 that were evaluated in both observational studies and RCTs, we used four relevant sources: the National Institutes of Health Covid-19 Treatment Guidelines,[15] a living review and network meta-analysis published in *The BMJ*,[16] a living systematic review with meta-analysis and trial sequential analysis published in *PLOS Medicine* (The LIVING Project),[17] and the Epistemonikos "Living OVerview of Evidence" (L·OVE) evidence database (https://app.iloveevidence.com/covid19). From each source, we recorded all therapeutic interventions with at least two RCTs by 9 February 2021. To ensure that the most prominent interventions were selected, we then narrowed down this list to the three interventions with the most unduplicated trials pooled across all four sources: hydroxychloroquine, lopinavir-ritonavir, and dexamethasone.

Although the protocol for this study was not published before the study commenced, the objectives and methods were prespecified before we analysed any data.

Identification of observational studies

To identify observational studies of the three interventions, four authors (OM, GD, KN, and JDW) reviewed the 36 covid-19 evidence databases included in the Center for Science in the Public Interest covid-19 Evidence hub, a resource that seeks to aggregate all international databases related to evidence on covid-19.[18] From among these databases, we selected the L·OVE evidence database because of its comprehensive search strategy. Briefly, the database searches through 41 bibliographical and grey literature sources (eg, Medline, Embase, bioRxiv, ClinicalTrials.gov, medRxiv), and had identified 94 893 records related to covid-19 as of 9 February 2021.

On 23 February 2021, we worked with a medical librarian (KN) to narrow the sample of all covid-19 records in the L·OVE database by further filtering for records tagged as "prevention or treatment," as opposed to those that were tagged as only "diagnostic," "epidemiology," "etiology," "epidemiology," or "prognosis," under the "Select type of question" heading. Within the "Select intervention" heading, we identified records that were tagged for any of our three interventions. This approach of identifying observational studies via screening classified records from an evidence hub, rather than relying on keyword searches in traditional bibliographical databases, was endorsed by a medical librarian (KN) after empirical testing.

The resulting sample included 4774 records, which were imported into Covidence software to remove duplications and be screened by four investigators (OM, GD, JS, and JDW) at the title and abstract level. Two investigators (OM and GD) then evaluated potentially eligible records at the full text level to identify prospective or retrospective observational studies and case-control studies that evaluated the comparative effectiveness of interventions. We excluded studies that were not in English; were case study reports, case series, or cohort studies with a sample size of <15; were interventional studies or studies that did not include a comparator group; or were cross sectional studies or case-control studies that did not evaluate the comparative effectiveness of an intervention. Any uncertainties were resolved by consensus and discussion between two investigators (OM and JDW).

Identification of RCTs

To identify RCTs for the three interventions, we selected one source among the four sources used to locate the most prominent covid-19 interventions: a living systematic review and network meta-analysis on drug treatments for covid-19 published in *The BMJ*.[16] We selected *The BMJ*'s living review because of its comprehensive search strategy and frequent updates.[16] In particular, *The BMJ*'s living review conducts daily searches of the World Health Organization covid-19 database, monthly searches of six Chinese databases, and regularly monitors the L·OVE database among other living evidence retrieval services. On 2 July 2021, we identified all individual RCTs included in the fourth version (published on 31 March 2021) of *The BMJ*'s living review directly comparing hydroxychloroquine, lopinavir-ritonavir, or dexamethasone to an active or placebo comparator. *The BMJ*'s living review searches were completed on 1 March 2021 in the fourth version, which closely matched our observational study search date. To minimize the potential of selecting specific outcomes based on the direction and strength of the treatment effects, we recorded all safety or efficacy outcomes considered by *The BMJ*'s living review.

Matching of observational studies and RCTs

To identify observational studies evaluating the same clinical questions as the RCTs included in *The BMJ*'s living review (that is, matched pairs), we first developed and undertook a prespecified hierarchical matching process. At least two individual authors (OM, JJS, GD, and JDW) independently screened and matched individual observational studies to individual RCTs if the observational studies and RCTs considered the same therapeutic intervention, comparator, and outcome measures. If eligible studies evaluated multiple therapeutic interventions (eg, hydroxychloroquine, dexamethasone, and placebo), they were included in multiple matches (eg, hydroxychloroquine *v* dexamethasone, dexamethasone *v* placebo, and hydroxychloroquine *v* placebo). To match *The BMJ*'s living review method,[16] we did not differentiate between interventions based on the dosage or duration of treatment, we considered placebo and standard of care comparators clinically equivalent, and we allowed flexibility in phrasing of outcomes (eg, time to symptom resolution may have been matched with time to clinical improvement). Similar to *The BMJ*'s living review, we did not require matching based on severity of illness or other study demographic characteristics (eg, sex distribution, age).

Data extraction

For each eligible observational study identified through the L·OVE database and RCT included in *The BMJ*'s living review, we recorded the study title, date of publication, design, intervention, comparator, sample size (intention-to-treat sample size for RCTs), center status (multicentre or single center), disease severity (that is, mild to moderate, severe, critical), proportion of female and male participants, and age distribution (mean (standard deviation) or median (interquartile range)). For RCTs, we also determined whether the study was masked (none, open, double or higher, unknown). Given that *The BMJ*'s living review calculated mean differences or ratios of means for meta-analyses of continuous outcomes, two authors (OM and JDW) extracted means or medians and their corresponding confidence intervals, standard deviations, standard errors, or any other available data used to calculate treatment effect and measures of precision for the individual observational studies and RCTs. *The BMJ*'s living review reported odds ratios (ORs) for meta-analyses of dichotomous outcomes so we abstracted counts for all relevant intervention and comparator groups from the observational studies and RCTs. For observational studies, we prioritized the counts from propensity score matched populations, whenever reported. When values were not reported in the text of eligible studies, an online digitizer (https://apps.automeris.io/wpd/) was used to reverse engineer Kaplan-Meier curves to extract the underlying numerical data.

Data analysis

Meta-analyses of observational studies and randomized controlled trials

To generate matched pairs, we first verified all our abstractions for the eligible RCTs using data shared by the authors of *The BMJ*'s living review.[16] Next, we calculated ORs (95% confidence intervals) for all dichotomous outcomes and converted all medians and corresponding interquartile ranges, minimum and maximum values, or 95% confidence intervals to means and standard deviations, ensuring that the observational treatment effect estimates matched those reported in *The BMJ*'s living review.

When at least two observational studies or two RCTs were identified evaluating the same therapeutic, comparator, and outcome measure, we used the DerSimonian and Laird procedure for random effects to conduct separate meta-analyses of all observational studies and RCTs. This process resulted in meta-analyzed matched pairs with the same summary treatment effects. We did not rely on the summary treatment effects reported in *The BMJ*'s living review because those were estimated using a Bayesian framework. Instead, we combined observational studies and re-evaluated the summary treatment effects for the RCTs using the DerSimonian and Laird procedure for random effects, which assumes that the identified studies are estimating different effects, is widely implemented, and does not require any assumptions about priors for the variance and effect parameters. For studies with outcomes with zero cell frequencies, we used a continuity correction of 0.5. We assessed the proportion of total variability due to heterogeneity between studies using the I^2 statistic.

Comparison of summary demographics from the matched pairs

We used descriptive statistics to compare the demographic characteristics (that is, sex, age) and disease severity among the patient populations included in the matched observational study and RCT pairs. Sex distribution was considered to be concordant if pairs included studies with only one sex or if pairs included studies with both sexes. Age distribution was concordant if pairs included studies in which the mean age ranges fell within the same age range (that is, pediatric (0-17 years), adult (18-64 years), or elderly (≥65 years)). A pair would still be considered concordant if the RCTs or observational studies included only one of the three age ranges and the matched observational studies or RCTs included that age range and additional age ranges. For disease severity, we determined whether the studies in a matched pair had any patients from each of the three disease severity categories. We assigned concordance for disease severity if pairs included studies that had any overlap in the severity of included patients (eg, a pair would be considered concordant if the RCTs included patients with only mild to moderate disease and the observational studies included patients with mild to moderate and severe disease).

Comparison of treatment effect estimates from matched pairs

Individual or summary treatment effects from matched pairs were separately characterized on the basis of their significance (that is, $P<0.05$ *v* $P\geq0.05$) and direction (that is, increased for odds ratios and ratios of means greater than 1 or mean differences greater than 0, and decreased for odds ratios and ratios of means less than 1 or mean differences less than 0). Treatment effect estimates from matched pairs were concordant if the direction of the observational and RCT treatment effect estimates was concordant and both the treatment effect estimates were significant, or if the observational study and RCT treatment effect estimates were both not significant. Treatment effect estimates from matched pairs that did not fulfil either of these criteria were classified as discordant. Although P values are imperfect measures, our binary classification system, based on the traditional alpha cut-off value of 0.05, is useful for showing how significance is most often defined in the literature.

As secondary measures of concordance, we determined how often the observational study and RCT treatment effects from matched pairs had overlapping 95% confidence intervals; whether the observational study treatment effects were included in RCT treatment effects' 95% confidence intervals[19]; and ratios of ORs, ratios of ratios of means, or differences between standardized mean differences between observational and RCT treatment effects. For each matched pair with treatment effects of dichotomous outcomes, we determined the ratio of ORs by exponentiating the differences between the natural log-scale ORs from the observational study and the RCT ORs. For each matched pair with treatment effects of continuous outcomes, we determined the ratio of ratios of means by exponentiating the differences between the natural log-scale ratios of means for the observational study and that for the RCT, or determined the difference between standardized mean differences by subtracting the standardized mean difference of the RCT from that of the observational study. Ratios of ORs and ratios of ratios of means greater than 1.0 and differences between standardized mean differences greater than 0.0 implied greater summary treatment effects in the observational studies, while values less than 1.0 or 0.0, respectively, implied the opposite. We calculated 95% confidence intervals for these values by taking the square root of the sum of the variance for the two original outcome measures from the summary RCT and observational treatment effect estimates. These variance calculations assumed independence between observational study and RCT outcomes. We considered $P<0.05$ to be significant for all two sided tests. All analyses were done using the meta package in R (version 4.1.2).

Publication timing

We also compared when the individual studies in the matched pairs were published. Using ClinicalTrials.gov, we determined the date of trial registration and date of publication for individual RCTs to classify whether all, any, or none of the observational studies were published before the RCTs were registered or published.

Risk-of-bias assessment

For individual RCTs, we abstracted the risk of bias evaluations reported in *The BMJ*'s living review, which were based on a revision of the Cochrane tool for assessing risk of bias in randomized trials (RoB 2.0).[20] RCTs were only rated at a low risk of bias if all domains received a classification of probably low risk or low risk of bias. For the individual observational studies, two authors (OM and JDW) conducted formal assessments of risk of material bias using the ROBINS-I tool for non-randomized studies.[21] Studies were rated serious or critical risk of bias overall if any of the domains received a classification of serious or critical risk of bias.

Sensitivity analyses

We repeated our analyses using the summary treatment effects from the network meta-analysis reported in *The BMJ*'s living review. Traditional meta-analyses only include RCTs that conduct direct or head-to-head comparisons between two interventions, whereas network meta-analyses can incorporate evidence from pairs of interventions that are not formally compared in individual RCTs but are included in a network of potential interventions.[22] In particular, if two RCTs of different interventions have a shared comparator, an indirect comparison can be made between the two interventions from different studies. As a post-hoc analysis, we repeated our analyses using the Hartung-Knapp-Sidik-Jonkman method for random effects meta-analyses and planned to repeat our concordance evaluation limited to observational studies and RCTs that were low risk of bias.[23]

Patient and public involvement

We did not involve members of the public or patients when we designed our study, interpreted the results, or wrote the manuscript. However, we asked members of the public to read our manuscript after submission to ensure it was understandable.

Results

Search results and characteristics of studies

Of 4774 records identified through the literature search on 23 February 2021, 1575 were excluded as duplicates and 2981 were excluded during the initial screening based on title and abstract (fig 1, supplementary tables 1 and 2). We reviewed 216 studies at the full text level to identify all individual observational studies comparing hydroxychloroquine, lopinavir-ritonavir, or dexamethasone to an active or placebo comparator for any safety or efficacy outcomes. After matching individual observational studies to individual RCTs or meta-analyses of RCTs, we identified 46 observational studies evaluating the same interventions, comparisons, and outcomes as six individual RCTs and 21 meta-analyses of RCTs (fig 1).

The 46 individual observational studies conducted 66 direct comparisons across multiple efficacy or safety outcomes for hydroxychloroquine versus standard of care (32 studies, 70%), hydroxychloroquine-azithromycin versus standard of care (15, 33%), lopinavir-ritonavir versus standard of care (6, 13%), hydroxychloroquine versus lopinavir-ritonavir (5, 11%), hydroxychloroquine

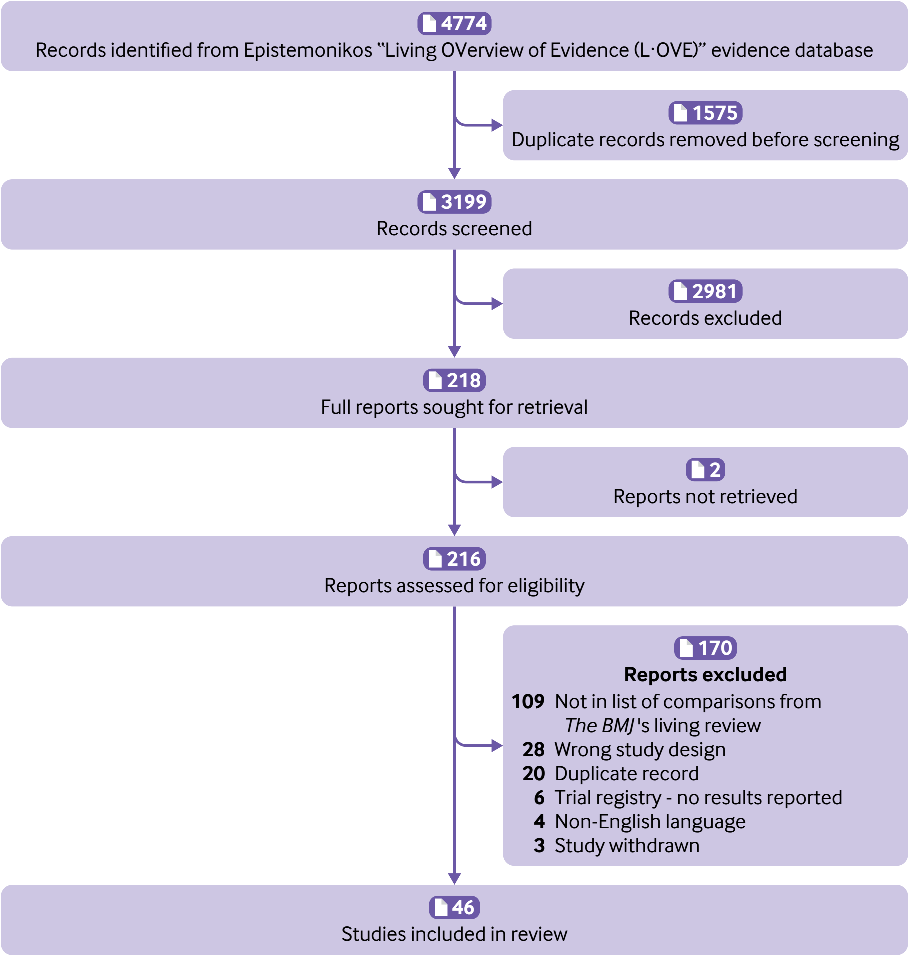

4774
Records identified from Epistemonikos "Living OVerview of Evidence (L·OVE)" evidence database
1575
Duplicate records removed before screening
3199
Records screened
2981
Records excluded
218
Full reports sought for retrieval
2
Reports not retrieved
216
Reports assessed for eligibility
170
Reports excluded
109 Not in list of comparisons from The BMJ's living review
28 Wrong study design
20 Duplicate record
6 Trial registry - no results reported
4 Non-English language
3 Study withdrawn
46
Studies included in review

Fig 1 | Study flow chart

Table 1 | Characteristics of individual observational studies and randomised controlled trials included in the matched pairs. Data are number (%)

Characteristic	Observational studies (n=46)	RCTs (n=37)
Study design		
Retrospective cohort	44 (96)	NA
Prospective cohort	2 (4)	NA
Masking		
None	NA	18 (49)
Open	NA	2 (5)
Unknown	NA	1 (3)
Double or higher	NA	16 (43)
Center status		
Single center	25 (54)	14 (38)
Multicenter	21 (46)	22 (60)
Unknown	0 (0)	1 (3)
Study location		
China	3 (7)	6 (16)
France	8 (17)	1 (3)
Italy	3 (7)	0 (0)
Saudi Arabia	3 (7)	0 (0)
South Korea	4 (8)	0 (0)
Spain	3 (7)	1 (3)
United States	13 (28)	8 (22)
Other	9 (20)	14 (38)
Multiple countries	0 (0)	7 (19)
Individual studies for comparisons included in matched observational study and RCT pairs*		
Hydroxychloroquine *v* standard of care/placebo	32 (70)	29 (78)
Dexamethasone *v* standard of care/placebo	3 (7)	3 (8)
Lopinavir-ritonavir *v* standard of care/placebo	6 (13)	6 (16)
Hydroxychloroquine-azithromycin *v* standard of care/placebo	15 (33)	4 (11)
Hydroxychloroquine *v* lopinavir-ritonavir	5 (11)	2 (5)
Hydroxychloroquine *v* azithromycin	4 (9)	2 (5)
Hydroxychloroquine *v* ivermectin	1 (2)	2 (5)
Study population demographics		
No of patients per individual study, median (range)	464 (16-8075)	146 (2-6425)
Percentage of female individuals per study, median (IQR)	40.8 (33.5-46.7)	40.2 (36.4-45.0)
Study population disease severity† (across n=27 matched pairs)		
Mild or moderate disease	23 (85)	22 (82)
Severe disease	22 (82)	18 (67)
Critical disease	19 (70)	2 (7)

NA=not applicable; RCT=randomised controlled trial; IQR=interquartile range.

*Percentages within this category might not add to 100% because one observational study or RCT could have included multiple treatment comparisons.

†Percentages within this category might not add to 100% because one observational study or RCT could have included patients of multiple disease severity types (eg, one observational study might have included patients with severe disease and critical disease).

versus azithromycin (4, 9%), dexamethasone versus standard of care (3, 7%), hydroxychloroquine versus ivermectin (1, 2%; table 1). Over half (25, 54%) of the observational studies were done at one center. The most common geographical areas were United States (13, 28%), followed by France (8, 17%) and South Korea (4, 8%). Almost all observational studies were retrospective cohort studies (44, 96%). About a quarter (12, 26%) used propensity score matching.

The 37 individual RCTs conducted direct comparisons for hydroxychloroquine versus standard of care (29 trials, 78%), hydroxychloroquine-azithromycin versus standard of care (4, 11%), lopinavir-ritonavir versus standard of care (6, 16%), dexamethasone versus standard of care (3, 8%), hydroxychloroquine versus lopinavir-ritonavir (2, 5%), hydroxychloroquine versus azithromycin (2, 5%), and hydroxychloroquine versus ivermectin (2, 5%; table 1). More than half (22, 60%) of the RCTs were conducted at multiple centers; seven (19%) were in multiple countries, eight (22%) in the United States, and six (16%) in China. Just fewer than half (16, 43%) of the trials were designed to have double or higher masking.

Comparison of matched pairs

We conducted 17 new, independent meta-analyses of observational studies evaluating the same interventions, comparisons, and outcomes as 17 meta-analyses of RCTs included in *The BMJ*'s living review (table 1). Ten additional matched pairs had only one eligible observational study and/or one eligible RCT, making 27 matched pairs in total.

Matched observational studies, compared with RCTs, had the same median number of included studies (2 (range 1-26) *v* 2 (1-28)) and a lower median number of total participants (1188 (40-42 859) *v* 1288 (85-11 655); table 1, supplementary tables 3 and 4). The median proportion of female participants in matched observational studies was 40.8% (interquartile range 33.5-46.7) and in RCTs was 40.2% (36.4-45.0). Across the 27 matched pairs of observational studies versus RCTs, patients with mild to moderate covid-19 disease severity were similar in number; however, more patients in the observational group had severe disease and critical disease (table 1). In 23 (85%) matched pairs, the observational studies included patients with mild to moderate covid-19 disease severity, 22 (82%) with severe covid-19 disease severity, and 19 (70%) with critical covid-19 disease severity. 22 (82%) matched pairs had RCTs that included patients with mild to moderate covid-19 disease severity, 18 (67%) with severe covid-19 disease severity, and two (7%) with critical covid-19 disease severity.

All matched pairs included studies with a similar proportion of female participants (supplementary table 5). After excluding one pair with missing information about the ages of participants in at least one individual study, participant age was concordant for all remaining 26 pairs. After excluding five pairs with missing information about the disease severity of participants in at least one individual study, all remaining 22 pairs had concordant disease severity.

Concordance between treatment effects from matched pairs

Seventeen matched pairs consisted of meta-analyses of observational studies and meta-analyses of RCTs and 10 additional matched pairs had only one observational study and/or only one RCT (table 2 and table 3). Overall, 21 (78%) of the 27 total matched pairs had treatment effects that were concordant in terms of direction of effect and statistical significance. For 11 (52%) of 21 concordant pairs, the treatment effects were larger (that is, further from the null value) for the observational studies than for the RCTs.

Of the 17 matched pairs consisting of meta-analyses of observational studies and meta-analyses of RCTs, 14 (82%) had treatment effects that were concordant for direction and statistical significance (table 2 and

Table 2 | Concordance between treatment effect estimates from 27 matched observational study and RCT pairs

	RCT treatment effect estimates			
Observational study treatment effect estimates	Increased, significantly*	Decreased, significantly*	Increased, but not significantly†	Decreased, but not significantly†
Matched pairs of meta-analyses of observational studies and meta-analyses of RCTs				
Increased, significantly*	0‡	0	2	0
Decreased, significantly*	0	0‡	0	0
Increased, but not significantly†	0	0	4‡	2‡
Decreased, but not significantly†	0	1	5‡	3‡
Additional matched pairs consisting of one observational study and/or one RCT				
Increased, significantly*	0‡	0	0	1
Decreased, significantly*	0	0‡	0	1
Increased, but not significantly†	1	0	3‡	1‡
Decreased, but not significantly†	0	0	1‡	2‡

RCT=randomized controlled trial.
*Statistically significant based on P<0.05.
†Not statistically significant based on P≥0.05.
‡Pairs classified as concordant.

table 3, fig 2, fig 3, fig 4, fig 5). Overall, the summary observational study treatment effects were larger than the summary RCT treatment effect estimates for six (43%) of the 14 concordant pairs. Of the 10 additional matched pairs consisting of one observational study and/or one RCT, seven (70%) were concordant. Overall, the summary observational study treatment effects were larger than the summary RCT treatment effects for five (71%) of the seven concordant pairs. Twelve (57%) of 21 meta-analyses of RCTs and three (15%) of 20 meta-analyses of observational studies had I^2 values of 0% (supplementary figures 1-27). Twelve (60%) of 20 meta-analyses of observational studies had I^2 values of more than 75%.

Treatment effects for dichotomous outcomes

Among the 27 total matched pairs, 18 had ORs as their treatment effect (fig 2, fig 3, table 3, supplementary figs 1-18). Of these, 16 (89%) were concordant in terms of direction and significance.

Of 12 matched pairs consisting of meta-analyses of observational studies and meta-analyses of RCTs with ORs as their treatment effect, 11 (92%) were concordant in terms of direction and statistical significance (fig 2, fig 3, table 3). All 11 concordant matched meta-analysis pairs had non-significant summary ORs for RCTs and observational studies. Of these, six (55%) had summary observational study ORs that were contained within the 95% confidence interval of the summary RCT ORs. Of the 11 concordant pairs, the summary observational study ORs were larger (further from the null value of 1.0) than the summary RCT ORs for four (36%) of the pairs (median ratio of ORs, 0.96 (range 0.33 to 1.53)). Although one meta-analysis of RCTs (dexamethasone *v* standard of care for mortality) had a statistically significant summary OR, the corresponding matched meta-analysis of observational studies was not concordant.

Among the six matched pairs consisting of one observational study and/or one RCT with ORs, five (83%) were concordant in terms of direction and statistical significance (fig 2, fig 3, table 3). All five concordant pairs had non-significant RCT and observational study ORs. Of the five concordant pairs, four (80%) had summary observational study ORs that were contained within the 95% confidence interval of the summary RCT ORs. Of the five concordant pairs, the summary observational study ORs were larger (further from the null value of 1.0) than the summary RCT ORs for three (60%) of the pairs (median ratio of ORs, 1.61 (range 0.45 to 3.47)). Although one observational study (hydroxychloroquine *v* standard of care for hospital admission) had a statistically significant OR, the corresponding matched RCT OR was not concordant.

Treatment effects for continuous outcomes

Among the 27 total matched pairs, nine (33%) had mean differences or ratios of means as their treatment effects (fig 4, fig 5, table 3, supplementary figures 19-27). Of these, five (56%) were concordant in terms of direction and statistical significance.

Five matched pairs consisted of meta-analyses of observational studies and meta-analyses of RCTs with mean differences or ratios of means, of which three (60%) were concordant (fig 4 and fig 5, table 3). All three concordant matched meta-analysis pairs had non-significant summary treatment effects. For all three concordant pairs, the summary observational study treatment effects were contained within the 95% confidence interval of the summary RCT treatment effects. The summary treatment effects for the observational studies were larger (that is, further from the null value of 1.0) than those for RCTs, for two (67%) of the pairs. Although two meta-analyses of observational studies had significant summary treatment effects (hydroxychloroquine *v* standard of care or placebo for duration of hospital stay, and hydroxychloroquine-azithromycin *v* standard of care or placebo for duration of hospital stay), the matched RCT meta-analyses pairs were not considered to be concordant.

Among the four additional matched pairs consisting of only one observational study and/or only one RCT with continuous treatment effect estimates, two (50%) were concordant (fig 4 and fig 5, table 3). Both concordant pairs had non-significant treatment effect estimates for RCTs and observational studies. Of the two concordant pairs, one (50%) had observational treatment effect estimates that were contained within the 95% confidence interval of the summary RCT treatment effect estimates. For both concordant pairs, the summary treatment effect estimates for observational studies were smaller (that is, further from the null value of 1.0) than those for RCTs. Although one observational study had a significant summary treatment effect estimate (hydroxychloroquine *v* lopinavir-ritonavir for duration of hospital stay), the matched RCT pair was not concordant. For another pair, one RCT had a significant summary treatment effect estimate (dexamethasone *v* standard of care or placebo for ventilator-free days); however, the matched observational study was not concordant.

Table 3 | Comparison of treatment effect estimates for observational studies and RCTs among 27 matched pairs

	Primary measures of concordance			Secondary measures of concordance	
Comparison and outcome	Concordant direction of effect estimate	Significant summary treatment effect estimate (P<0.05)	Overall concordance*	Summary treatment effect estimate for observational studies within 95% CI of the summary treatment effect estimate for RCTs	Overlapping 95% CI
Treatment effects for dichotomous outcomes (n=18)					
Hydroxychloroquine *v* standard of care/placebo:					
Mortality	No	Neither	Yes	No	Yes
Mechanical ventilation	Yes	Neither	Yes	No	Yes
Admission to hospital†	Yes	Observational only	No	Yes	Yes
Viral clearance	Yes	Neither	Yes	Yes	Yes
Lopinavir-ritonavir *v* standard of care/placebo:					
Mortality	No	Neither	Yes	Yes	Yes
Mechanical ventilation	No	Neither	Yes	No	Yes
Viral clearance	No	Neither	Yes	No	Yes
Dexamethasone *v* standard of care/placebo:					
Mortality	Yes	RCT only	No	No	Yes
Hydroxychloroquine-azithromycin *v* standard of care/placebo:					
Mortality	No	Neither	Yes	Yes	Yes
Mechanical ventilation†	Yes	Neither	Yes	Yes	Yes
Viral clearance	No	Neither	Yes	No	Yes
Hydroxychloroquine *v* azithromycin:					
Mortality	Yes	Neither	Yes	Yes	Yes
Mechanical ventilation†	Yes	Neither	Yes	Yes	Yes
Hydroxychloroquine *v* lopinavir-ritonavir:					
Mortality	Yes	Neither	Yes	Yes	Yes
Mechanical ventilation†	No	Neither	Yes	No	Yes
Adverse events†	Yes	Neither	Yes	Yes	Yes
Viral clearance	Yes	Neither	Yes	Yes	Yes
Hydroxychloroquine *v* ivermectin:					
Mortality†	Yes	Neither	Yes	Yes	Yes
Treatment effects for continuous outcomes (n=9)					
Hydroxychloroquine *v* standard of care/placebo:					
Duration of hospital stay	Yes	Observational only	No	No	Yes
Time to symptom resolution†	Yes	Neither	Yes	No	Yes
Ventilator free days†	No	Neither	Yes	Yes	Yes
Time to viral clearance	No	Neither	Yes	Yes	Yes
Lopinavir-ritonavir *v* standard of care/placebo:					
Duration of hospital stay	Yes	Neither	Yes	Yes	Yes
Time to symptom resolution	Yes	Neither	Yes	Yes	Yes
Dexamethasone *v* standard of care/placebo:					
Ventilator free days†	Yes	RCT Only	No	No	Yes
Hydroxychloroquine-azithromycin *v* standard of care/placebo:					
Duration of hospital stay	Yes	Observational only	No	No	Yes
Hydroxychloroquine *v* lopinavir-ritonavir:					
Duration of hospital stay†	No	Observational Only	No	No	No

CI=confidence interval; RCT=randomized controlled trial.

*Concordant in terms of direction and statistical significance (that is, both significantly increasing or decreasing or both not significant).

†Pairs with only one observational study or only one RCT.

Publication timing

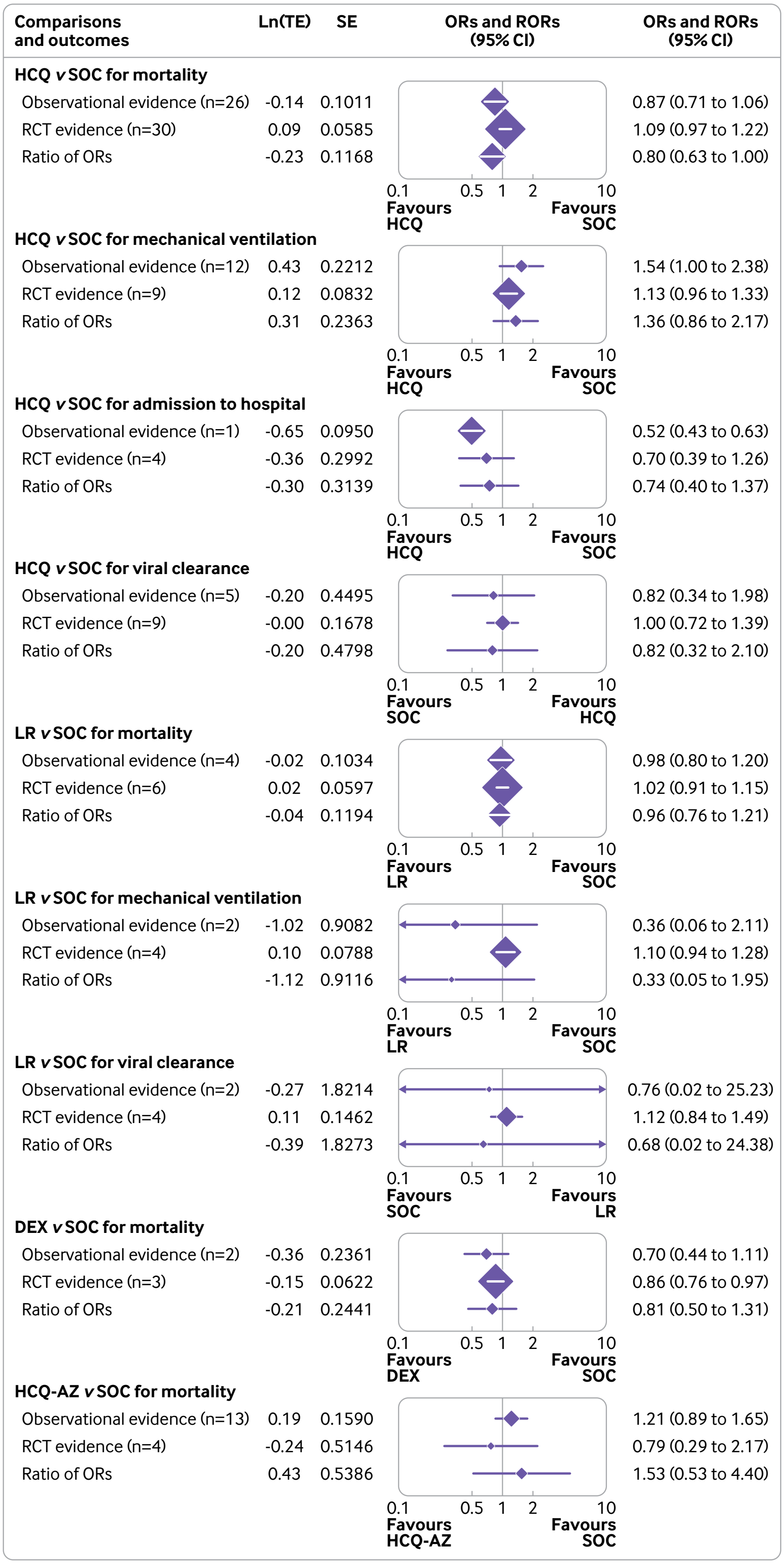
Twenty six (96%) of the 27 matched pairs had at least one observational study published before any RCT was published, and four (15%) had at least one observational study published before any RCTs were registered. Although 15 pairs (56%) had all observational studies published before all the RCTs were published, none had all observational studies published before all RCTs were registered (supplementary table 6).

Risk of bias

Among the 37 eligible RCTs, none was rated at low risk of bias across all domains (supplementary tables 7a and 7b). Among the 46 individual observational studies, none was rated at low risk of bias in all domains. Given that no meta-analyses pairs had at least two studies at low risk of bias, we did not repeat our analyses stratified by high and low risk-of-bias evidence.

Sensitivity analyses

Repeating our analyses using the treatment effects from *The BMJ*'s network meta-analysis (that is, considering both indirect and direct comparisons of interventions) produced results consistent with our primary findings (supplementary tables 8-10). Our findings remained unchanged after we applied the Hartung-Knapp-Sidik-Jonkman method for random effects meta-analyses to the summary treatment effects that were significant



Comparisons and outcomes	Ln(TE)	SE	ORs and RORs (95% CI)
HCQ *v* SOC for mortality			
Observational evidence (n=26)	-0.14	0.1011	0.87 (0.71 to 1.06)
RCT evidence (n=30)	0.09	0.0585	1.09 (0.97 to 1.22)
Ratio of ORs	-0.23	0.1168	0.80 (0.63 to 1.00)
HCQ *v* SOC for mechanical ventilation			
Observational evidence (n=12)	0.43	0.2212	1.54 (1.00 to 2.38)
RCT evidence (n=9)	0.12	0.0832	1.13 (0.96 to 1.33)
Ratio of ORs	0.31	0.2363	1.36 (0.86 to 2.17)
HCQ *v* SOC for admission to hospital			
Observational evidence (n=1)	-0.65	0.0950	0.52 (0.43 to 0.63)
RCT evidence (n=4)	-0.36	0.2992	0.70 (0.39 to 1.26)
Ratio of ORs	-0.30	0.3139	0.74 (0.40 to 1.37)
HCQ *v* SOC for viral clearance			
Observational evidence (n=5)	-0.20	0.4495	0.82 (0.34 to 1.98)
RCT evidence (n=9)	-0.00	0.1678	1.00 (0.72 to 1.39)
Ratio of ORs	-0.20	0.4798	0.82 (0.32 to 2.10)
LR *v* SOC for mortality			
Observational evidence (n=4)	-0.02	0.1034	0.98 (0.80 to 1.20)
RCT evidence (n=6)	0.02	0.0597	1.02 (0.91 to 1.15)
Ratio of ORs	-0.04	0.1194	0.96 (0.76 to 1.21)
LR *v* SOC for mechanical ventilation			
Observational evidence (n=2)	-1.02	0.9082	0.36 (0.06 to 2.11)
RCT evidence (n=4)	0.10	0.0788	1.10 (0.94 to 1.28)
Ratio of ORs	-1.12	0.9116	0.33 (0.05 to 1.95)
LR *v* SOC for viral clearance			
Observational evidence (n=2)	-0.27	1.8214	0.76 (0.02 to 25.23)
RCT evidence (n=4)	0.11	0.1462	1.12 (0.84 to 1.49)
Ratio of ORs	-0.39	1.8273	0.68 (0.02 to 24.38)
DEX *v* SOC for mortality			
Observational evidence (n=2)	-0.36	0.2361	0.70 (0.44 to 1.11)
RCT evidence (n=3)	-0.15	0.0622	0.86 (0.76 to 0.97)
Ratio of ORs	-0.21	0.2441	0.81 (0.50 to 1.31)
HCQ-AZ *v* SOC for mortality			
Observational evidence (n=13)	0.19	0.1590	1.21 (0.89 to 1.65)
RCT evidence (n=4)	-0.24	0.5146	0.79 (0.29 to 2.17)
Ratio of ORs	0.43	0.5386	1.53 (0.53 to 4.40)

Fig 2 | Forest plot of treatment effects from matched observational studies and randomized controlled trials, shown with Odds ratios (ORs) and ratios of ORs (RORs). Numbers next to comparisons indicate the number of individual studies included in the meta-analyses. Az=azithromycin; CI=confidence interval; DEX=dexamethasone; HCQ=hydroxychloroquine; IVER=ivermectin; LR=lopinavir-ritonavir; Ln=natural logarithmic scale; SE=standard error; SOC=standard of care or placebo; TE=treatment effect

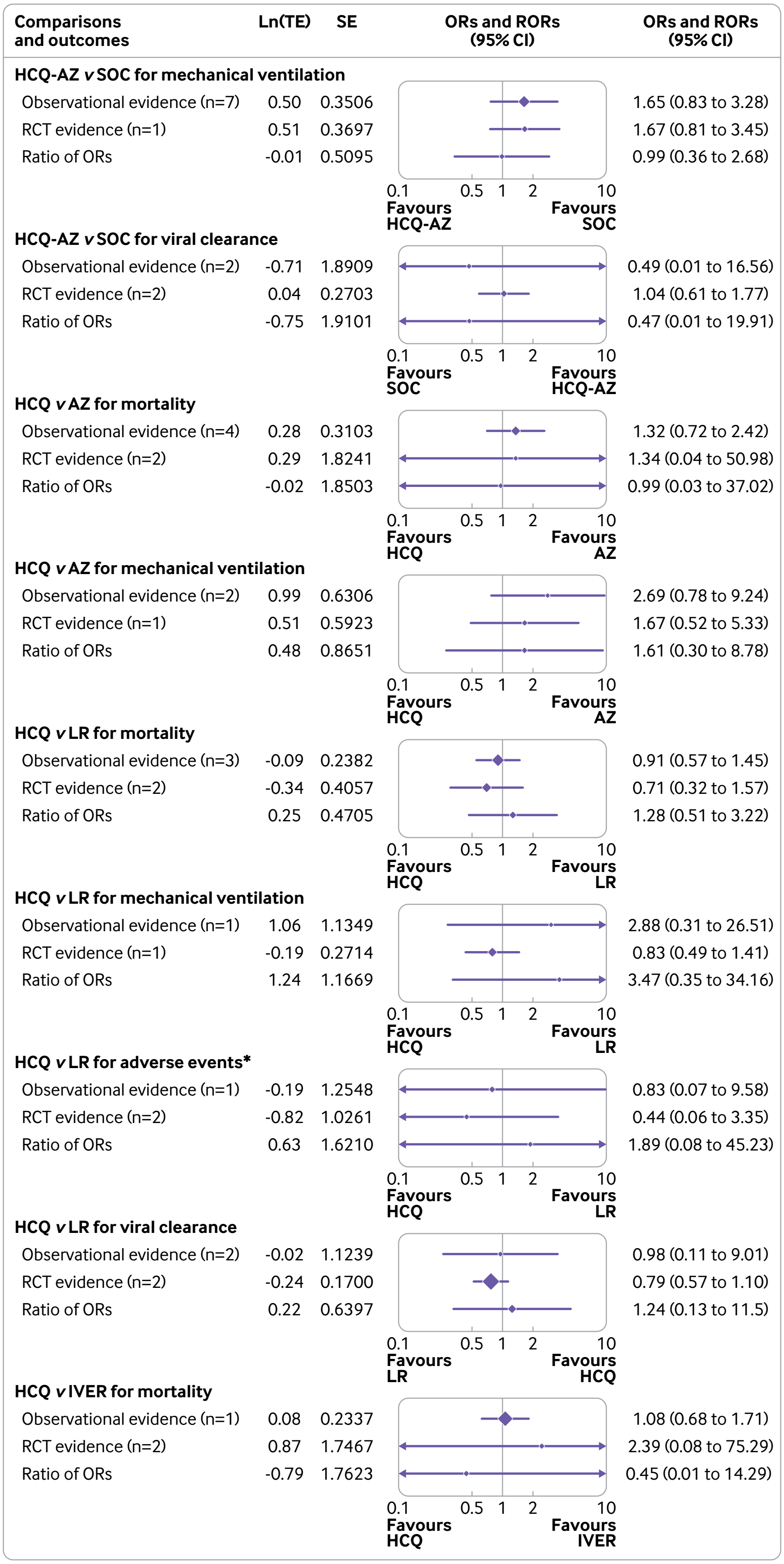
according to the DerSimonian and Laird method for random effects meta-analyses (supplementary text 1).

Discussion

In this meta-epidemiological study, we found that more than three quarters of the matched observational study and RCT pairs had treatment effects that were in agreement in terms of direction of effect and significance. However, agreement was higher when matched pairs were limited to meta-analyses of observational studies and meta-analyses of RCTs (82%), compared with matched pairs with only one observational study and/or one RCT (70%), or both. We noted higher agreement in matched pairs' treatment effects for dichotomous outcomes (89%) than in those of continuous outcomes (56%). Although our evaluation is limited to three covid-19 treatments, and therefore might not be generalizable to all interventions and outcomes, these findings suggest that meta-analyzed evidence from observational studies can complement, but should not replace, evidence collected from RCTs.

We observed relatively high proportions of agreement between matched observational study and RCT meta-analysis pairs, especially among matched pairs consisting of meta-analyses of observational studies and meta-analyses of RCTs. However, a recent cross-sectional study suggested that only 12% of individual non-randomized studies reporting significant survival benefits of potential anti-covid-19 drugs are replicated by large RCTs.[24] Several reasons explain why our findings are not directly comparable to that study and might lead to slightly different conclusions. First, our evaluation considered meta-analyzed evidence for three treatments with the greatest number of RCTs across four sources, and we did not include studies evaluating groups of interventions (eg, corticosteroids). In fact, the only overlapping analyses across both evaluations were for hydroxychloroquine and hydroxychloroquine-azithromycin versus placebo or standard of care. Second, we did not limit our comparisons to statistically significant findings. We classified evidence from observational studies and RCTs as in agreement if both observational and RCT treatment effects were significantly increasing or decreasing (P<0.05) or both treatment effects were not significant. Lastly, our evaluation did not include evidence from cohort studies with sample sizes of less than 15 and from cross sectional studies. However, our findings of high levels of agreement are consistent with previous evaluations comparing the relative treatment effects from observational studies and RCTs. For instance, in a Cochrane review, which summarized evidence from 14 methodological reviews across 228 different medical conditions, 11 reviews found that observational studies and RCTs were highly concordant in terms of effect direction and magnitude.[25] In one of these reviews, very high correlation was observed among treatment effects of observational studies and RCTs (r=0.75) across 45 different clinical topics.[11]

In contrast with the higher proportion of agreement between relative treatment effects in matched



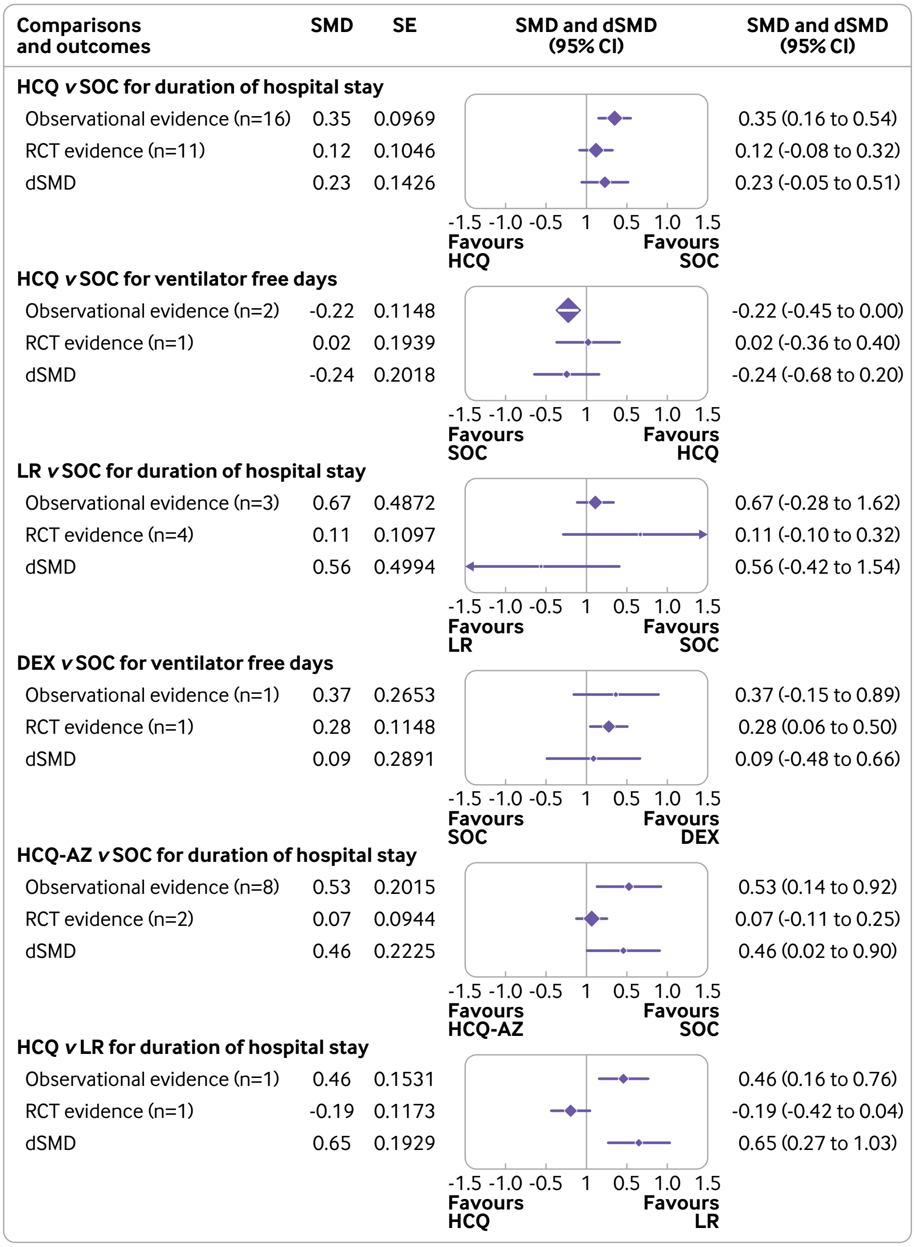
Comparisons and outcomes	Ln(TE)	SE	ORs and RORs (95% CI)
HCQ-AZ *v* SOC for mechanical ventilation			
Observational evidence (n=7)	0.50	0.3506	1.65 (0.83 to 3.28)
RCT evidence (n=1)	0.51	0.3697	1.67 (0.81 to 3.45)
Ratio of ORs	-0.01	0.5095	0.99 (0.36 to 2.68)
HCQ-AZ *v* SOC for viral clearance			
Observational evidence (n=2)	-0.71	1.8909	0.49 (0.01 to 16.56)
RCT evidence (n=2)	0.04	0.2703	1.04 (0.61 to 1.77)
Ratio of ORs	-0.75	1.9101	0.47 (0.01 to 19.91)
HCQ *v* AZ for mortality			
Observational evidence (n=4)	0.28	0.3103	1.32 (0.72 to 2.42)
RCT evidence (n=2)	0.29	1.8241	1.34 (0.04 to 50.98)
Ratio of ORs	-0.02	1.8503	0.99 (0.03 to 37.02)
HCQ *v* AZ for mechanical ventilation			
Observational evidence (n=2)	0.99	0.6306	2.69 (0.78 to 9.24)
RCT evidence (n=1)	0.51	0.5923	1.67 (0.52 to 5.33)
Ratio of ORs	0.48	0.8651	1.61 (0.30 to 8.78)
HCQ *v* LR for mortality			
Observational evidence (n=3)	-0.09	0.2382	0.91 (0.57 to 1.45)
RCT evidence (n=2)	-0.34	0.4057	0.71 (0.32 to 1.57)
Ratio of ORs	0.25	0.4705	1.28 (0.51 to 3.22)
HCQ *v* LR for mechanical ventilation			
Observational evidence (n=1)	1.06	1.1349	2.88 (0.31 to 26.51)
RCT evidence (n=1)	-0.19	0.2714	0.83 (0.49 to 1.41)
Ratio of ORs	1.24	1.1669	3.47 (0.35 to 34.16)
HCQ *v* LR for adverse events*			
Observational evidence (n=1)	-0.19	1.2548	0.83 (0.07 to 9.58)
RCT evidence (n=2)	-0.82	1.0261	0.44 (0.06 to 3.35)
Ratio of ORs	0.63	1.6210	1.89 (0.08 to 45.23)
HCQ *v* LR for viral clearance			
Observational evidence (n=2)	-0.02	1.1239	0.98 (0.11 to 9.01)
RCT evidence (n=2)	-0.24	0.1700	0.79 (0.57 to 1.10)
Ratio of ORs	0.22	0.6397	1.24 (0.13 to 11.5)
HCQ *v* IVER for mortality			
Observational evidence (n=1)	0.08	0.2337	1.08 (0.68 to 1.71)
RCT evidence (n=2)	0.87	1.7467	2.39 (0.08 to 75.29)
Ratio of ORs	-0.79	1.7623	0.45 (0.01 to 14.29)

Fig 3 | Forest plot of treatment effects from matched observational studies and randomized controlled trials, shown with odds ratios (ORs) and ratios of ORs (RORs). Numbers next to comparisons indicate the number of individual studies included in the meta-analyses. Az=azithromycin; CI=confidence interval; DEX=dexamethasone; HCQ=hydroxychloroquine; IVER=ivermectin; LR=lopinavir-ritonavir; Ln=natural logarithmic scale; SE=standard error; SOC=standard of care or placebo

observational study and RCT meta-analysis pairs, 56% of the matched pairs with continuous treatment effects were in agreement. Few studies have examined the agreement between continuous treatment effects from observational studies and RCTs, as compared with relative treatment effects. However, according to a previous evaluation comparing standardized treatment response (a continuous treatment effect estimate) in antidepressants between observational studies and RCTs, treatment effects in RCTs and observational studies differed significantly, with a greater magnitude of effect observed among RCTs than among observational studies.[26] Overall, why concordance was lower among matched pairs with treatment effects for continuous outcomes than was for dichotomous outcomes in this study is unclear. Higher agreement among the treatment effects for dichotomous outcomes might partially be explained by the fact that most dichotomous outcomes are hard outcomes, such as mortality, which can be less prone to detection biases. The higher agreement might also be explained by the fact that almost half of the matched pairs with continuous outcomes contained one observational study and/or one RCT. Unlike accumulated evidence, individual studies, especially those with smaller sample sizes, might be more likely to have spurious findings, which could lead to low agreement.[27] Additional research might be needed to track the consistency of treatment effects for continuous outcomes as evidence is accumulated.

Of the three interventions we evaluated, only dexamethasone has been recommended by the National Institutes of Health and World Health Organization's under each institution's treatment guidelines for therapeutic management of covid-19 in patients admitted to hospital.[15 28] However, the treatment effects from our DerSimonian and Laird random effects meta-analyses of observational studies (two studies, 2544 participants) and RCTs (three RCTs, 6742 participants) were not in agreement. Although both meta-analyses had ORs suggesting benefit with dexamethasone versus standard of care or placebo for mortality, only the accumulated RCT evidence was significant.

Across dichotomous outcomes, we found one observational study reporting a significant reduction in hospital admission among patients receiving hydroxychloroquine versus standard of care. However, this observational study was not a meta-analysis of studies and the corresponding evidence from RCTs was not significant. Three clinical questions across continuous outcomes showed significant treatment effects based on evidence from observational studies: hydroxychloroquine versus standard of care (n=16 studies), hydroxychloroquine-azithromycin versus standard of care (n=8 studies), and hydroxychloroquine versus lopinavir-ritonavir (n=1 study) for duration of hospital stay. According to these analyses, use of hydroxychloroquine was associated with lengthier hospital stays for all analyses. However, the corresponding evidence from RCTs was not

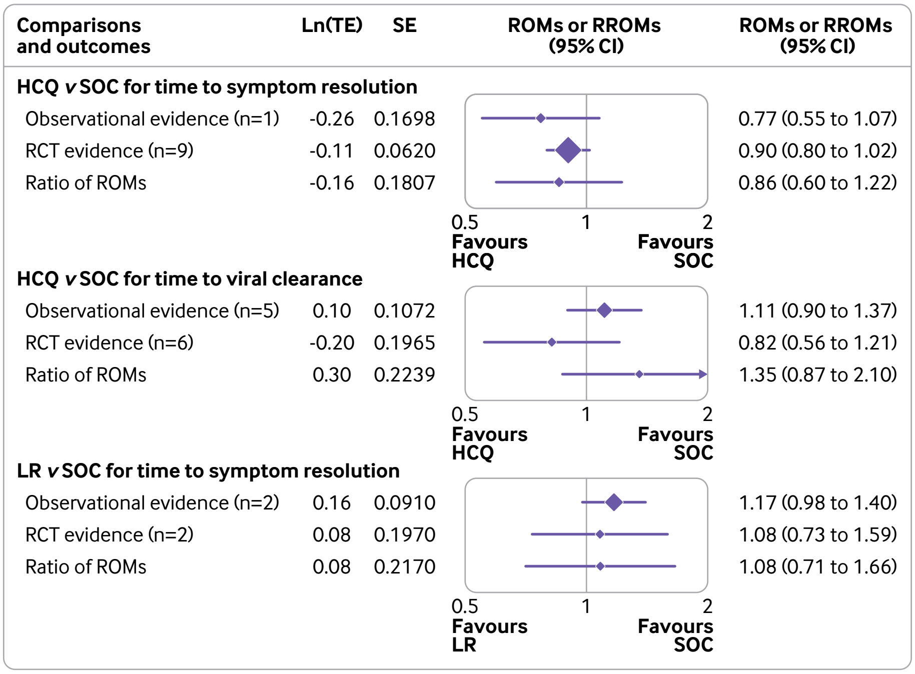


Comparisons and outcomes	SMD	SE	SMD and dSMD (95% CI)
HCQ *v* SOC for duration of hospital stay			
Observational evidence (n=16)	0.35	0.0969	0.35 (0.16 to 0.54)
RCT evidence (n=11)	0.12	0.1046	0.12 (-0.08 to 0.32)
dSMD	0.23	0.1426	0.23 (-0.05 to 0.51)
HCQ *v* SOC for ventilator free days			
Observational evidence (n=2)	-0.22	0.1148	-0.22 (-0.45 to 0.00)
RCT evidence (n=1)	0.02	0.1939	0.02 (-0.36 to 0.40)
dSMD	-0.24	0.2018	-0.24 (-0.68 to 0.20)
LR *v* SOC for duration of hospital stay			
Observational evidence (n=3)	0.67	0.4872	0.67 (-0.28 to 1.62)
RCT evidence (n=4)	0.11	0.1097	0.11 (-0.10 to 0.32)
dSMD	0.56	0.4994	0.56 (-0.42 to 1.54)
DEX *v* SOC for ventilator free days			
Observational evidence (n=1)	0.37	0.2653	0.37 (-0.15 to 0.89)
RCT evidence (n=1)	0.28	0.1148	0.28 (0.06 to 0.50)
dSMD	0.09	0.2891	0.09 (-0.48 to 0.66)
HCQ-AZ *v* SOC for duration of hospital stay			
Observational evidence (n=8)	0.53	0.2015	0.53 (0.14 to 0.92)
RCT evidence (n=2)	0.07	0.0944	0.07 (-0.11 to 0.25)
dSMD	0.46	0.2225	0.46 (0.02 to 0.90)
HCQ *v* LR for duration of hospital stay			
Observational evidence (n=1)	0.46	0.1531	0.46 (0.16 to 0.76)
RCT evidence (n=1)	-0.19	0.1173	-0.19 (-0.42 to 0.04)
dSMD	0.65	0.1929	0.65 (0.27 to 1.03)

Fig 4 | Forest plot of treatment effects from matched observational studies and randomized controlled trials, shown with standardized mean differences (SMD) and differences between SMDs (dSMD). Numbers next to comparisons indicate the number of individual studies included in the meta-analyses. Az=azithromycin; CI=confidence interval; DEX=dexamethasone; HCQ=hydroxychloroquine; LR=lopinavir-ritonavir; SE=standard error; SOC=standard of care or placebo

significant across any of these analyses, highlighting the complexity of relying only on observational evidence, even when studies are accumulated.

Even before the pandemic, interest was increasing for the use of observational methods and real world data to assess the comparative effectiveness of medical treatments.[29] In 2016, the US Congress passed the 21st Century Cures Act, which promoted the use of real-world evidence (that is, any longitudinal health data collected outside of RCTs) to support drug approvals.[30] The US Food and Drug Administration followed this legislative directive by releasing a guidance document outlining how drug sponsors can use real-world evidence to support regulatory decisions.[30-32] Similarly, the European Medicines Agency has encouraged the use of real world evidence by facilitating multiple real world databases to be used in both in-house and external studies.[33] Real world evidence was revisited during the covid-19 pandemic when clinicians and regulators urgently sought information about the effectiveness of pre-existing interventions to treat covid-19. However, concerns were raised about the rapid dissemination of potentially low quality observational studies and findings of treatment effectiveness (or ineffectiveness) from observational studies of covid-19 were in doubt until corresponding evidence was generated by RCTs. If the findings reported in observational studies are subsequently refuted by RCTs, the implications



Comparisons and outcomes	Ln(TE)	SE	ROMs or RROMs (95% CI)
HCQ *v* SOC for time to symptom resolution			
Observational evidence (n=1)	-0.26	0.1698	0.77 (0.55 to 1.07)
RCT evidence (n=9)	-0.11	0.0620	0.90 (0.80 to 1.02)
Ratio of ROMs	-0.16	0.1807	0.86 (0.60 to 1.22)
HCQ *v* SOC for time to viral clearance			
Observational evidence (n=5)	0.10	0.1072	1.11 (0.90 to 1.37)
RCT evidence (n=6)	-0.20	0.1965	0.82 (0.56 to 1.21)
Ratio of ROMs	0.30	0.2239	1.35 (0.87 to 2.10)
LR *v* SOC for time to symptom resolution			
Observational evidence (n=2)	0.16	0.0910	1.17 (0.98 to 1.40)
RCT evidence (n=2)	0.08	0.1970	1.08 (0.73 to 1.59)
Ratio of ROMs	0.08	0.2170	1.08 (0.71 to 1.66)

Fig 5 | Forest plot of treatment effects from matched observational studies and randomized controlled trials, shown with ratios of means (ROMs) and ratios of ROMs (RROMs). Numbers next to comparisons indicate the number of individual studies included in the meta-analyses. Az=azithromycin; CI=confidence interval; HCQ=hydroxychloroquine; LR=lopinavir-ritonavir; Ln=natural logarithmic scale; SE=standard error; SOC=standard of care or placebo

are important, especially because all but one of the matched pairs in our sample had at least one observational study published before any RCT was published. However, our study might suggest that neither RCTs nor observational studies should be automatically assumed to serve as a gold standard.[4] Beyond research integrity concerns (eg, falsification and fabrication), the methodological quality of covid-19 articles, across all study designs, has been found to be lower than in non-covid-19 articles.[34] As we observed in our study, very few individual RCTs and observational studies included in our evaluation were at low risk of bias.

In addition to biases, other methodological limitations can affect the direction and magnitude of treatment effect estimates.[35] For instance, RCTs typically have strict inclusion and exclusion criteria that might not necessarily reflect the populations that eventually receive the same treatments. Conversely, rigorous observational studies might have more representative, heterogeneous patient populations, containing participants that cannot be included in RCTs. Therefore, to expect identical results from observational studies and RCTs might not be reasonable, even when the studies evaluate the exact same interventions, comparators, and outcomes. Furthermore, RCTs have strict inclusion and exclusion criteria and are often subject to recruitment difficulties, which can lead to smaller sample sizes than in observational studies. In our sample, we found a higher median number of total participants across observational studies compared with RCTs. Although our findings might suggest that in future pandemics—or other similar situations requiring urgent evaluations of existing treatments for new indications—meta-analyses of observational studies could be used to complement the evidence generated by RCTs, caution is necessary before using any evidence to inform clinical and regulatory decision making. In situations where policy decisions to treat or not to treat individuals can affect millions of lives, prioritization of accumulated evidence is necessary, especially when evidence from rigorous, large RCTs is not available. For both observational and RCT evidence, the methods, populations, results, and overall quality of individual studies must be evaluated before use of the evidence to guide practice.

Limitations

Our study has several limitations. First, our study focused on the three covid interventions with the greatest number of RCTs. Although this approach ensured that the number of RCTs for our comparisons were adequate, our findings might not be generalizable to other interventions with few or no RCTs or to non-repurposed treatments for covid-19. Second, the observed agreement between RCTs and observational evidence in our study could be driven by the low number of identified studies for specific pairs, the low total power of the accumulated evidence, or the fact that most treatment effects were null. Third, we used a comprehensive and systematic process to identify observational studies but relied on the individual RCTs reported in a high quality living review.[16] Although the databases and search dates used to identify RCTs and observational studies were similar, use of two different approaches could have introduced bias or limited the number of eligible studies.

Fourth, more than half of the meta-analysis pairs had all observational studies published before all RCTs. However, whether one study design prompted the other in our sample is unclear. Given sample size restrictions, we were unable to formally compare the evidence from observational studies published before RCTs. Fifth, although we matched RCTs and observational studies based on population, intervention, comparator, and outcome, studies were unlikely to have had the same characteristics. For instance, the data for treatment effects from the observational studies are further limited by the potential heterogeneity between studies, including in the definition of standard of care in these treatment arms; patients within these groups could have been given a wide range of possible interventions. We did not formally evaluate the effect of study design characteristics on the agreement between matched pairs, and although we quantified statistical heterogeneity and conducted random effects meta-analyses, genuine heterogeneity is still possible for analyses with low I^2 values.

Finally, we did not request access to the raw data from all observational studies and RCTs, which would have been needed to directly compare the agreement between the study populations considered. However, we assessed the overlap between key demographic characteristics (that is, sex distribution, age, and disease severity) and found much agreement between matched RCT and observational study meta-analyses pairs. We did not include race as a demographic characteristic because we thought that the heterogeneity in study geographies and differences in reporting strategies meant that the information synthesis and derivation of useful claims would not have been possible without making overt generalizations. Of note, findings from RCTs have historically been incomplete because of challenges in recruiting diverse patient populations.[36-38]

Conclusion

In this meta-epidemiological study measuring the agreement between summary treatment effects from 27 matched pairs of observational studies and RCTs for three covid-19 treatments, we found that observational studies and RCTs generally have treatment effects that are in agreement in terms of direction and significance. Even greater agreement was noted among pairs consisting of meta-analyzed evidence from observational studies and RCTs and studies measuring relative effect estimates. Although these findings do not suggest that observational studies can replace RCTs in pandemic settings, meta-analyses of observational studies could complement evidence collected from RCTs.

AUTHOR AFFILIATIONS

[1]Yale School of Medicine, Yale University, New Haven, CT, USA

[2]Center for Science in the Public Interest, Washington, DC, USA

[3]Dartmouth-Hitchcock Medical Center, Lebanon, NH, USA

[4]Harvey Cushing/John Hay Whitney Medical Library, Yale University, New Haven, CT, USA

[5]Department of Environmental Health Sciences, Yale School of Public Health, New Haven, CT, USA

[6]Section of General Medicine and the National Clinician Scholars Program, Department of Internal Medicine, Yale School of Medicine, New Haven, CT, USA

[7]Center for Outcomes Research and Evaluation, Yale-New Haven Health System, New Haven, CT, USA

[8]Department of Health Policy and Management, Yale School of Public Health, New Haven, CT, USA

We thank the authors of *The BMJ*'s living systematic review and network meta-analysis on drug treatments for covid-19 for answering questions that allowed for the reanalysis of their data.

Contributors: OM, PL, JSR, and JDW conceived this study. OM, KN, PL, JSR, and JDW designed the study. OM, GD, JJS, and JDW acquired the data. OM and JDW conducted the statistical analysis. OM and JDW drafted the manuscript. All authors participated in the interpretation of the data and critically revised the manuscript for important intellectual content. JDW had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. JDW provided supervision. JDW is the guarantor. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Funding: JDW is supported by the National Institute on Alcohol Abuse and Alcoholism of the National Institutes of Health under award K01AA028258. The funders had no role in considering the study design or in the collection, analysis, interpretation of data, writing of the report, or decision to submit the article for publication.

Competing interests: All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf and declare: support from the National Institute on Alcohol Abuse and Alcoholism of the National Institutes of Health for the submitted work; JSR is the US outreach and associate research editor at *The BMJ* and currently receives research support through Yale University from Johnson & Johnson to develop methods of clinical trial data sharing, from the Medical Device Innovation Consortium as part of the National Evaluation System for Health Technology, from the US Food and Drug Administration for the Yale-Mayo Clinic Center for Excellence in Regulatory Science and Innovation program (U01FD005938), from the Agency for Healthcare Research and Quality (R01HS022882), from the National Heart, Lung and Blood Institute of the National Institutes of Health (R01HS025164, R01HL144644), and from the Laura and John Arnold Foundation to establish the Good Pharma Scorecard at Bioethics International; JSR is also an expert witness at the request of Relator's attorneys, the Greene law firm, in a lawsuit alleging violations of the False Claims Act and Anti-Kickback Statute against Biogen; JDW currently received research support from the US Food and Drug Administration and through Yale University from Johnson & Johnson to develop methods of clinical trial data sharing.

Ethical approval: Not required.

Data sharing: The dataset will be made available via a publicly accessible repository on publication.

The senior author (manuscripts guarantor, JDW) affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant registered) have been explained.

Dissemination to participants and related patient and public communities: Results from our study will be disseminated through a press release (Yale University, Center for Science in the Public Interest), websites (eg, Center for Science in the Public Interest), and social media (Twitter).

Provenance and peer review: Not commissioned; externally peer reviewed.

1 Akobeng AK. Understanding randomised controlled trials. *Arch Dis Child* 2005;90:840-4. doi:10.1136/adc.2004.058222.

2 NEJM. Real-world evidence—what is it and what can it tell us? Accessed 23 December, 2020. https://www.nejm.org/doi/pdf/10.1056/NEJMsb1609216

3 Kennedy-Martin T, Curtis S, Faries D, Robinson S, Johnston J. A literature review on the representativeness of randomized controlled trial samples and implications for the external validity of trial results. *Trials* 2015;16:495. doi:10.1186/s13063-015-1023-4.

4 Janiaud P, Agarwal A, Tzoulaki I, et al. Validity of observational evidence on putative risk and protective factors: appraisal of 3744 meta-analyses on 57 topics. *BMC Med* 2021;19:157. doi:10.1186/s12916-021-02020-6.

5 Fergusson DA, Hébert PC, Mazer CD, et al, BART Investigators. A comparison of aprotinin and lysine analogues in high-risk cardiac surgery. *N Engl J Med* 2008;358:2319-31. doi:10.1056/NEJMoa0802395.

6 Schneeweiss S, Seeger JD, Landon J, Walker AM. Aprotinin during coronary-artery bypass grafting and risk of death. *N Engl J Med* 2008;358:771-83. doi:10.1056/NEJMoa0707571.

7 Patorno E, Goldfine AB, Schneeweiss S, et al. Cardiovascular outcomes associated with canagliflozin versus other non-gliflozin antidiabetic drugs: population based cohort study. *BMJ* 2018;360:k119. doi:10.1136/bmj.k119.

8 Neal B, Perkovic V, Mahaffey KW, et al, CANVAS Program Collaborative Group. Canagliflozin and cardiovascular and renal events in type 2 diabetes. *N Engl J Med* 2017;377:644-57. doi:10.1056/NEJMoa1611925.

9 Hemkens LG, Contopoulos-Ioannidis DG, Ioannidis JPA. Agreement of treatment effects for mortality from routinely collected data and subsequent randomized trials: meta-epidemiological survey. *BMJ* 2016;352:i493. doi:10.1136/bmj.i493.

10 Ewald H, Ioannidis JPA, Ladanie A, Mc Cord K, Bucher HC, Hemkens LG. Nonrandomized studies using causal-modeling may give different answers than RCTs: a meta-epidemiological study. *J Clin Epidemiol* 2020;118:29-41. doi:10.1016/j.jclinepi.2019.10.012.

11 Ioannidis JPA, Haidich AB, Pappa M, et al. Comparison of evidence of treatment effects in randomized and nonrandomized studies. *JAMA* 2001;286:821-30. doi:10.1001/jama.286.7.821.

12 Pundi K, Perino AC, Harrington RA, Krumholz HM, Turakhia MP. Characteristics and strength of evidence of covid-19 studies registered on ClinicalTrials.gov. *JAMA Intern Med* 2020;180:1398-400. doi:10.1001/jamainternmed.2020.2904.

13 Mehra MR, Desai SS, Ruschitzka F, Patel AN. RETRACTED: Hydroxychloroquine or chloroquine with or without a macrolide for treatment of COVID-19: a multinational registry analysis. *Lancet* 2020;0:S0140-6736(20)31180-6. doi:10.1016/S0140-6736(20)31180-6.

14 Mehra MR, Desai SS, Kuy S, Henry TD, Patel AN. RETRACTED: Cardiovascular disease, drug therapy, and mortality in covid-19. *N Engl J Med* 2020;382:e102. doi:10.1056/NEJMoa2007621.

15 COVID-19 Treatment Guidelines Panel. COVID-19 treatment guidelines. Accessed 31 December, 2020. https://www.covid19treatmentguidelines.nih.gov/therapeutic-management/

16 Siemieniuk RA, Bartoszko JJ, Ge L, et al. Drug treatments for covid-19: living systematic review and network meta-analysis. *BMJ* 2020;370:m2980. doi:10.1136/bmj.m2980.

17 Juul S, Nielsen EE, Feinberg J, et al. Interventions for treatment of COVID-19: A living systematic review with meta-analyses and trial sequential analyses (The LIVING Project). *PLoS Med* 2020;17:e1003293. doi:10.1371/journal.pmed.1003293.

18 COVID-19 Evidence Hub. Center for Science in the Public Interest. 2020. https://cspinet.org/covid-19-evidence-hub

19 Franklin JM, Patorno E, Desai RJ, et al. Emulating randomized clinical trials with nonrandomized real-world evidence studies: first results from the RCT DUPLICATE initiative. *Circulation* 2021;143:1002-13. doi:10.1161/CIRCULATIONAHA.120.051718.

20 Sterne JAC, Savović J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. *BMJ* 2019;366:l4898. doi:10.1136/bmj.l4898.

21 Sterne JA, Hernán MA, Reeves BC, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ* 2016;355:i4919. doi:10.1136/bmj.i4919.

22 Chaimani A, Caldwell DM, Li T, Higgins JP, Salanti G. Undertaking network meta-analyses. In: *Cochrane Handbook for Systematic Reviews of Interventions.* John Wiley & Sons, Ltd, 2019: 285-320, doi:10.1002/9781119536604.ch11

23 IntHout J, Ioannidis JP, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. *BMC Med Res Methodol* 2014;14:25. doi:10.1186/1471-2288-14-25.

24 Shepshelovich D, Yahav D, Ben Ami R, Goldvaser H, Tau N. Concordance between the results of randomized and non-randomized interventional clinical trials assessing the efficacy of drugs for COVID-19: a cross-sectional study. *J Antimicrob Chemother* 2021;76:2415-8. doi:10.1093/jac/dkab163.

25 Anglemyer A, Horvath HT, Bero L. Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials. *Cochrane Database Syst Rev* 2014;4:MR000034. doi:10.1002/14651858.MR000034.pub2.

26 Naudet F, Maria AS, Falissard B. Antidepressant response in major depressive disorder: a meta-regression comparison of randomized controlled trials and observational studies. *PLoS One* 2011;6:e20811. doi:10.1371/journal.pone.0020811.

27 L'Abbé KA, Detsky AS, O'Rourke K. Meta-analysis in clinical research. *Ann Intern Med* 1987;107:224-33. doi:10.7326/0003-4819-107-2-224.

28 World Health Organization. Therapeutics and covid-19: living guideline. Accessed 5 November, 2021. https://www.who.int/publications-detail-redirect/WHO-2019-nCoV-therapeutics-2021.3

29 US Food and Drug Administration. Framework for FDA's real-world evidence program. Accessed 15 January, 2021. https://www.fda.gov/media/120060/download

30 Franklin JM, Glynn RJ, Martin D, Schneeweiss S. Evaluating the use of nonrandomized real-world data analyses for regulatory decision making. *Clin Pharmacol Ther* 2019;105:867-77. doi:10.1002/cpt.1351.

31 Gottlieb S. Breaking down barriers between clinical trials and clinical care: incorporating real world evidence into regulatory decision making, 28 January 2019. US Food and Drug Administration. 2019. https://www.fda.gov/news-events/speeches-fda-officials/breaking-down-barriers-between-clinical-trials-and-clinical-care-incorporating-real-world-evidence

32 Real-World Evidence Collaborative. Margolis Center for Health Policy. Accessed 9 February, 2021. https://healthpolicy.duke.edu/projects/real-world-evidence-collaborative

33 Cave A, Kurz X, Arlett P. Real-world data for regulatory decision making: challenges and possible solutions for Europe. *Clin Pharmacol Ther* 2019;106:36-9. doi:10.1002/cpt.1426.

34 Quinn TJ, Burton JK, Carter B, et al. Following the science? Comparison of methodological and reporting quality of covid-19 and other research from the first wave of the pandemic. *BMC Med* 2021;19:46. doi:10.1186/s12916-021-01920-x.

35 Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA* 1995;273:408-12. doi:10.1001/jama.1995.03520290060030.

36 Murthy VH, Krumholz HM, Gross CP. Participation in cancer clinical trials: race-, sex-, and age-based disparities. *JAMA* 2004;291:2720-6. doi:10.1001/jama.291.22.2720.

37 Flores LE, Frontera WR, Andrasik MP, et al. Assessment of the inclusion of racial/ethnic minority, female, and older individuals in vaccine clinical trials. *JAMA Netw Open* 2021;4:e2037640. doi:10.1001/jamanetworkopen.2020.37640.

38 Downing NS, Shah ND, Neiman JH, Aminawung JA, Krumholz HM, Ross JS. Participation of the elderly, women, and minorities in pivotal trials supporting 2011-2013 U.S. Food and Drug Administration approvals. *Trials* 2016;17:199. doi:10.1186/s13063-016-1322-4

Web appendix: Online appendix