PEER REVIEW HISTORY

BMJ Medicine publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

This paper was submitted to a another journal from BMJ but declined for publication following peer review. The authors addressed the reviewers' comments and submitted the revised paper to BMJ Medicine. The paper was subsequently accepted for publication at BMJ Medicine.

ARTICLE DETAILS

TITLE (PROVISIONAL)	COVID-19 Trial Preprints: Consistency with later publications and impact for decision-making
AUTHORS	Zeraatkar, Dena; Pitre, Tyler; Leung, Gareth; Cusano, Ellen; Agarwal, Arnav; Khalid, Faran; Escamilla, Zaira; Cooper, Matthew; Ghadimi, Maryam; Wang, Ying; Verdugo-Paiva, Francisca; Rada, Gabriel; Kum, Elena; Qasim, Anila; Bartoszko, Jessica; Siemieniuk, Reed; Patel, Chirag; Guyatt, Gordon; Brignardello-Petersen, Romina

VERSION 1 - REVIEW

REVIEWER 1	Camaradou, Jennifer**;** EUPATI Patient Expert. Competing Interest: I am a costed PPI lead on a study at UCL University College London, NIHR132914: "CICADA: Coronavirus Inter-sectionalities: Chronic Conditions and Disabilities AndMigrants and other Ethnic minorities" I have received commercial fees from GSK Glaxco Smith Klein, MEDABLE Inc and Hoffman-La Roche Limited I am a patient partner at the European Academy of Neurology WG ANS autonomic nervous system disorders I am a patient editor on the Advances in Therapy Journal I am a OneNeurology Global Partnership Ambassador I am a citizen partner at the Global Evidence Commission/COVID END I am a lay member on the NICE Covid expert Panel I am a lay member on the scientific committee at QResearch Oxford university I am a lay member on the NIHR AI AWARD panel and a member of NHS AI Lab/AAC EAG group
REVIEW RETURNED	26-Jun-2022

GENERAL COMMENTS	This is an important paper that adds to the information around Covid-19 and is a matter of public interest, the review is written from a patient and public perspective. This research benefits patients and the public, as it is important to understand how evidence around the pandemic COVID-19 has been translated into clinical management guidelines and policy instruments and adds value by boosting public confidence in transparency and integrity of trusted research and clinical trials that can attract >/ 10% of people and less so from under represented groups in research. Will it help our readers to make better decisions and, if so, how? This study is helpful for the public as it provides concrete data around the relative effectiveness, reliability and credibility of pre-prints compared to published papers around results emerging from clinical trials for medications for Covid-19 looking at all stages of

	disease including severe and hospitalised admissions, during the time period of the analysis of July 2020 and3 July 2021 looking at results from 356 trials globally, 101 pre-prints and 181 publications. It builds on looking at living systematic reviews 114 and network meta-analyses (SRNMAs) of drug treatments, antiviral antibodies and cellular therapies, and 115 prophylaxis for COVID-19, that provides real-time summaries 116 addressing the comparative effectiveness of treatments and prophylaxis for COVID-19. The article will help readers make better decisions by providing up to date accurate analysis on data from trials that do not solely focus on severity of illness but is wide encompassing and the article is balanced in suggesting that there is a need for evidence users and evidence intermediaries and the broader implementation science community to scrutinize preprints for falsified data and take into account the poor quality of evidence arising from e.g retrospective studies, so a need for sensitivity analysis and thinking subjectively is of paramount importance, the article provides reference to additional resources to help mitigate these concerns. Since many challenges exist in identifying the appropriate evidence, disseminating it to different stakeholders, implementing it and collaborating at break speed across different settings, the need to bridge the gap between what science knows and what is the know-do gap as determined by evidence makers is critically important and there is a need to male better decisions in faster time frames so that future pandemics can be better handled. Will the article add enough to existing knowledge? The study sheds light on the rates of precision and estimates in pre-prints compared to published articles on trial results and the effect of pre-prints on meta analytic studies in relation to evidence quality using GRADE system, and in particular in relation to falsified data and errors addressing issues around publication bias and therefore adds to existing knowledge. It is interesting to see that trials with industry sponsors and government funding and those reporting on severe Covid-19 were published faster although there were discrepancies in the reporting of methods and results between pre-print and published reports although overall effects do not appear to be affected a lot which is re-assuring. I would have liked to have seen the findings discussed in light of recent Cochrane Convenes report and some of the other work with regards to the Global Evidence Commission to address future societal challenges as that may have provided an additional layer of contextual analysis of interests to the target audience and in particular perhaps helped create better incentives to improve clinical research infrastructure in different countries and equitably distributed capacities to help intermediaries, evidence bodies and government bodies better collaborate to produce, share and use evidence. In addition to perhaps more emphasis on how to mitigate the bias seen itself in some of the results, which is primarily attributed to the fact that more than 2/3 of the trials created had an open label design within the framework of ensuring future sustainability and agility of working practices when another global emergency occurs. The Global Evidence Commission found that “Global commissions are also silent on the need to have the protocols for randomized-controlled trials and other study designs, as well as national evidence-support systems and a broader global evidence architecture, ‘ready to go’ or already in use” section 7.2 (20) recommendations Global Evidence Commission. The methods used seem appropriate though as a lay am not fully qualified to comment on this, but it would appear that they do

	address the nature of both confounding and publication bias sufficiently. It may be useful to refer to another recommendation of the Global Evidence commission here that has called for journal publishers to improve the ways in which they support the use of best evidence. “Journals can mandate the use of reporting guidance and critical-appraisal checklists by reviewers, the placement of single studies in the context of evidence syntheses, and the sharing of anonymized study data. They can also commit to publishing non-positive research reports and replication studies, avoiding ‘spin,’ and acting quickly when apprised of scientific misconduct. Journals need to find a timely way to publish updates to living evidence products. Journals also need to ensure that publication delays never hinder the public sharing of evidence that is urgently needed for decision-making (and reciprocally that public sharing does not preclude later publication in a journal)” section 7.2 (23) recommendations Global Evidence Commission. Scientific reliability The design of the study includes many confounding variables and the authors address some of the limitations of the study well such as the fact that not all servers of published reports are included in the WHO Covid-19 database and that there may still be errors contains in some that are in the database, but does not give that much detail on approaches that could be taken to eliminate bias so e.g. assigning participants to alternative interventions using randomly generated interventions. I would have liked to see more refence to some of the current rationale behind increasing transparency of reporting requirements as a compulsory requirement for registration with regulatory authorities so the UK MHRA for example has made this a pre-requisite as well as patient inclusion and it would be perhaps interesting to other ways that traditional reporting can better integrate outcomes that matter to patients though this may be outside the scope of the current study. This work is able to meaningfully contribute to the research field around data integrity and transparency of trial results gathered during the Covid-19 pandemic and likely to lead to incremental change in perception of and usage of pre-prints in other fields in the medium term. Given the speed with which some of the early clinical research studies and RCTS took place, it has been previously reported that PPI patient public involvement practices took a bit of a back seat, so any future analysis of other variants in the current time frame of 2022/23 should look at how patient reported outcomes and patient experience data could be better embedded in the design, identification and setting of end points in particular in relation patients that experienced mild disease and patients that now meet criteria for PACS known as LongCOVID. Importance of the work to general readers The inclusion of patients is notable, in that they have helped select the outcomes which are listed as being mortality, mechanical ventilation, duration of hospitalization, time to resolution, clinical improvement and virological outcomes and we are told that patients were involved in “generalization of recommendations part of SRNMA” but we are not told more details, which I feel is an omission. Patients and member of the public want to know what they can present to their doctor to get better advice and medication, but they may not understand the finer nuances in either the publication authorship process of the fact that the totality of the body of evidence is taken into account when forming national clinical guidelines, or how confirmation and information bias contribute to

	information mismanagement. It may have been useful to think about data visualization in the form of an infographic and other ways of communicating the content of the article as the trial results tables are fairly technical and dense for the average lay reader. Does the article read well and make sense? Does it have a clear message? As the topic and methods of evidence and the broader networks around it such as multilateral organizations, development banks, the Organisation for Economic Co-operation and Development, the G20,national and sub-national government policymakers, evidence intermediaries, including those who do not currently function as evidence intermediaries (such as journalists for the most part), evidence producers (such as units engaged in producing and supporting the use of data analytics, modelling, evaluation, behavioural / implementation research, qualitative insights, evidence syntheses, technology assessment / cost-effectiveness analysis, and guidelines makers) are largely not that well understood by citizens at large or members of the public and sometimes patient advocacy groups, it may have been useful for the patients involved in the outcome selection to have written a short lay summary for the article even though the target audience is primarily healthcare professionals as this may represent an opportunity for broader engagement work. I also feel that given some movements by national and European patient organisations and aggregator of interests such as EURORDIS, representing the voice of rare disease patients in Europe, EFNA representing neurological patient associations and other EU government funded groups in structures programs by the EU IHI and ERAnets it might have been good to try and compare and feed some of their concerns and positions in relation to interruption of clinical trials and to look at further analysis of differences in particular regions and in closer detail around some of the outcomes chosen in relation to clinical improvement and virological outcomes as people with existing conditions may have not met full inclusion criteria for some of the work and/or the faster results in relation to potential therapeutic solutions that meet rigorous scientific standards from e.g RCTs can impact actual health service improvement and delivery of care, for example many patients in clinically vulnerable categories e.g rare disease patients may have been awaiting direction of guidance from appropriate medical bodies e.g. if for example not being able to use a particular type of medication due to it being contra-indicated, so it is important to highlight the extra re-assurance that data can bring to patients and their families to keep up with sometimes conflicting information messages that may not have been tailored to their condition, due to for example not having access to specialist care as a result of the pandemic or their own clinical trial being halted for allocation of resources in the hospital elsewhere. This demographic of patients may want to see specific results tailored with additional data pertaining to their condition so an additional level of filtering and also reference to what that means in terms of service improvement within the UK picture please see ARDENT by the Cambridge Rare Disease network detailing the effect the pandemic has had on rare disease patients. Given the international emergency of the pandemic and the imperative for swift evidence-based medicine, it is imperative to ensure quicker results on what works in terms of clinical management. This article has a clear message that data from preprints is reliable and can further enhance trust and transparency

	in clinical trials reporting, a subject that is becoming increasingly important with reference to patient advocacy groups if one looks at work such as the Good Clinical Trials Initiatives guidelines and advocacy campaigns Transparimed. This study can be used to guide national policies on the subject matter at hand and adds to the knowledge base around the importance of efficient but flexible and agile methods for evidence synthesis that are grounded in methodological rigour but can adapt to ever changing circumstances as those dictated in international emergencies like pandemics. The research team are very strong References: https://www.mcmasterforum.org/networks/evidence-commission https://www.camraredisease.org/ardent/ https://www.clinicaltrialsarena.com/news/transparimed-trial-results/

REVIEWER 2	Godolphin, Peter; Institute for Clinical Trials and Methodology, University College London. Competing Interest: None
REVIEW RETURNED	26-Jun-2022

GENERAL COMMENTS	This article by Zeraatkar and colleagues describes a methodological study that aimed to assess the trustworthiness and impact of COVID-19 RCT preprints. The paper is original, clearly laid out, well written and interesting. The study appears of high standard, with a protocol included as supplemental material and data underpinning the study available on OSF – it is good to see a methodological study conducted in this fashion. I enjoyed reading the paper but have a concern regarding whether the authors have answered their primary research question and have a few other additional minor comments. These are detailed below: Major comment 1. Is “trustworthiness” adequately addressed? This paper hinges around assessing the trustworthiness of preprints. It is first defined at the bottom of page 5, as “complete and consistent reporting of key aspects of the methods and results” and is measured in this study by comparing the preprint methods/results against the published paper methods/results. I am not convinced that this is really trustworthiness, and instead my opinion is that this is looking to see if there is consistency between the preprint and published trial. Is the preprint trustworthy just because the peer-reviewed paper is similar? Surely to be able to answer the key question of “are preprints trustworthy” you would need to assess the data underpinning the preprint. I worry that referring to this as trustworthiness exceeds what the authors have done and readers who see only the headline message/abstract may misinterpret the findings. Minor comments 2. Prominence given to point estimates When comparing meta-analysis results from MAs that include/exclude preprints, the authors focus on point estimates (and direction of point estimates) as their approach to determine if conclusions change. I strongly recommend that the authors consider the variability around the point estimate before they claim

	benefit/harm/no effect. 3. Greater discussion around meta-analysis Here are some things you may consider commenting on in the discussion that the paper touches on, but that I think could come out clearly in the discussion and (in my opinion) would be of interest to readers. • Unpublished data in a meta-analysis. A number of meta-analyses of COVID-19 trials included data from unpublished trials (i.e., not even published as preprints). Whilst this is a step even further, it may be an interesting discussion point. See for example the IL-6 prospective meta-analysis (Shankar-Hari et al. JAMA 2021). • Timing of when to conduct a MA. The authors focus around 4 crude timepoints for demonstration and it is interesting to see how the evidence accumulates over time. Do the authors have any comments on when you should conduct a meta-analysis? There are various approaches that tackle timing such as living systematic reviews, framework for prospective adaptive meta-analysis (FAME, Tierney et al. PLoS Med 2021) and ALL-IN meta-analysis (ter Schure and Grunwald 2021). 4. Justification for search date The last search was carried out on August 3rd 2021. Given this is linked to the living systematic review and network meta-analysis, I would have expected that the search would be far more up to date. The authors do comment that these searches are carried out daily, so it is not clear to me why they have chosen this date. Updating this or providing justification for this date would be useful. 5. Risk of bias Risk of bias was specified as high risk for many of these trials for an outcome of mortality due to the open-label nature. Personally, I disagree with this judgement, and do not consider that the open-label nature could introduce bias into an outcome as objective as mortality. I do recognise that the authors are consistent with their previous approach but still suggest they reconsider. Specific comments 1. Retractions Is there a typo in the results section? think there is a typo and it should say six instead of four 2. Supplement 3 – Is some text missing here? The RHS column is empty for some fields – e.g., p45 Probably low risk of bias 3. Supplement 4 – There is some missing text here, “43,849 records excluded for not being…” 4. Supplement 6 – fixed effect models have also been carried out. Please specify what model this is based on in the methods. 5. Supplement 6 – Instead of “current” it may be clearer to say “On X Date”

REVIEWER 3	Bero, Lisa; University of Colorado Anschutz Medical Campus. Competing Interest: None
REVIEW RETURNED	26-Jun-2022

GENERAL COMMENTS	The main new contribution of this paper is the evaluation of the effect of preprints on meta-analytic estimates. This paper did not assess "trustworthiness" of the trials. This term usually refers to the underlying data, not the reports (see, for a summary, https://www.cochranelibrary.com/cdsr/doi/10.1002/14651858.ED000152/full). The current paper assessed only reports (preprints and manuscripts). The objective of this study is more accurately described as assessing the reporting of preprints and publications, not their trustworthiness. The word trustworthiness should not be used in the title or conclusion of the manuscript. One could also argue that the word "impact" in the title is misleading as this paper studies a narrow impact of preprints – on meta-analytic estimates and GRADE ratings, not on social media indicators or guidelines or decisions, as other studies have done. INTRODUCTION The introduction to the paper needs to be updated as it does not summarize prior research evaluating COVID-19 preprints or comparing COVID-19 preprints to their final publications. See below, and there are likely more recent studies. (The preprint of one of these studies is cited in the discussion section, but, oddly, the final publication is not). Bero L, Lawrence R, Leslie L, et al. Cross-sectional study of preprints and final journal publications from COVID-19 studies: discrepancies in results reporting and spin in interpretation. BMJ Open 2021;11:e051821. doi:10.1136/ bmjopen-2021-051821 Nicolalde B, Anazco D, Mushtaq M, et al. Citations and publication rate of preprints on pharmacological interventions for COVID-19: the good, the bad and, the ugly. Res Sq 2020;version 2. Kataoka Y, Oide S, Ariie T, et al. COVID-19 randomized controlled trials in medRxiv and PubMed. Eur J Intern Med 2020;81:97–9. METHODS The protocol is submitted as an appendix but was not published. Why not publish in OSF or on some other open access platform? These platforms also allow comments and are publicly accessible. Line 135-136: How were "concerns regarding research integrity" monitored via Epistemonikos and the WHO database? The referenced supplement 2 does not provide any information on searching for concerns regarding research integrity. Furthermore, no data are reported on these concerns, so this phrase should be deleted from the methods section. As noted in https://www.cochranelibrary.com/cdsr/doi/10.1002/14651858.ED000152/full, retractions are only the tip of the iceberg for identifying problematic studies. The search for preprints does not include a comprehensive list of preprint servers (Appendix 2). This is mentioned as a limitation in the discussion section. The method for linking preprints with subsequent publications needs to be clarified. Does the living systematic review use a study-based register, rather than a records based register? The value of a study based registers is that all versions of all records linked to a publication would be identified. What about multiple versions of preprints and articles? What were the selection criteria for forming the pairs? Line 173 (in the data collection section) states "For preprints with more than one version, we extracted data from the first version of the preprint, which is the least likely to have been modified in response to peer review" How were multiple versions of manuscripts handled? A major gap is that information on harm outcomes reported was not

	collected. Lines 169-174 list the included outcomes. Although the “direction” of an outcome was assessed (e.g., a drug study designed to determine if a drug decreased mortality would measure an increase in mortality or a decrease in mortality), this is not the same as outcomes that specifically assess harms. Previous studies have found discrepancies in reporting harm outcomes between COVID 19 preprints and their final publications (eg, Bero 2021). The patient involvement description does not appear relevant to this study, but rather to the living SRNMA and guidelines. The analysis focuses on comparing the reporting of preprints and their corresponding publications, as well as the calculation of meta-analytic estimates that exclude or include data from preprints at different time intervals. “Key methods” appear to be the same as risk of bias criteria “Key methods include description of the randomization process and allocation concealment, blinding of patients and healthcare providers, extent of and handling of missing outcome data, blinding of outcome assessors and adjudicators, and prespecification of outcomes and analysis. The “key results” analyzed focus on the meta-analytic estimates but do not include other aspects of outcome reporting that could vary (see Bero 2021). RESULTS Line 253. How was the determination that a preprint reported on interim results made? Did this have to be stated in the preprint or were the Ns for the outcomes compared between preprints and final publications? If n’s were compared, were increases or decreases for n’s for reported outcomes both considered to indicate interim results? The differences in outcome reporting should be described. An example is given (box 2), but no information on the details of these discrepancies. For example, did the reported outcomes differ in the metrics used, measures, time point at which assessments were made, population (subgroup) analyzed, N, etc? Comparing just the statistics and numerical results reported could miss key discrepancies in outcome reporting. Line 292. 11 publication / preprint pairs showed reporting of “an outcome” in the publication that was not included in the preprint. Clarify if this was one, at least one, or multiple outcomes between each preprint-publication pair. Also, what was the nature of these outcomes? Additional time points, harm, different metrics or cutoffs used? This critical information to determine if preprints could be missing important information compared to subsequent publications is not reported. It is not surprising that inclusion of preprints in the MA would decrease imprecision as this criteria primarily relates to the quantity and heterogeneity of data included in the meta-analysis. If less data are included in the meta-analysis, one would expect wider confidence intervals and / or greater heterogeneity. Another relevant question would be to determine if final publications rather than preprints should be included in meta-analyses. This could be done by comparing the analytic estimates between meta-analyses that include preprints vs. meta-analyses that include the final publications of these preprints. DISCUSSION The findings and implications are overstated as this paper does not assess trustworthiness, and only narrow measures of impact. The recommendation that systematic reviewers and guideline developers

	consider evidence from preprints (line 386-387) is supported by the data presented in this paper. The discussion of false and fabricated data (lines 392-400) appears to be an add on that is not directly related to this paper. The cursory mention of methods for detecting fabrication and falsification in clinical trials (box 3), does not cover other detritmental research practices which may make a study actually untrustworthy. The section on “relation to previous work” is inadequate. It is incomplete (see mention above re other relevant studies) and inaccurately describe one of the studies as assessing spin, when it actually assessed characteristics of outcome reporting, as well as spin. It is accurate that a major contribution of this study (as other have assessed reporting and spin) is that assessed the effect of preprints on analytic estimates, but, as noted above, a limitation of this analysis is that it did not compare meta-analytic estimates when preprints vs final publications were included.

REVIEWER 4	Reviewer 4
REVIEW RETURNED	26-Jun-2022

GENERAL COMMENTS	This is a very good study, that should be of interest to all evidence users, particularly stakeholders involved in data synthesis. It presents an analysis of key differences from preprints to peer-reviewed publications of COVID-19 clinical trials and changes in meta-analytic effect estimates and confidence assessments by including preprints or not. This is a substantial contribution to the decision-making process of planning and conducting meta-analyses. Below I list some major and minor points that, in my opinion, need revision. Major: 1. I do not see how the analyses presented inform on the trustworthiness of preprints. The study presents the changes in methods and results reported between preprint and peer-reviewed publication, but only some of the items assessed can be interpreted as improving trust by improving transparency. As this interpretation may be subjective and the authors do not wish to change the title, then I’d recommend adding clarification to the ‘Purpose’ section of the abstract. 2. Similarly, the abstract should include the clarification that the purpose of assessing impact is related to changes in meta-analysis as the impact on clinical practice or outcomes of the pandemic are not directly and systematically assessed. 3. An important limitation not mentioned is regarding the subjectivity of the GRADE framework for assessing certainty of evidence, which is one of the most relevant outcomes of the study. 4. On Table 1, single center and multicentre do not sum up to the total. Is there another option? There seems to be 10 studies missing in published without preprints, 14 missing in preprint only, and 4 missing in preprint first. 5. On Table 1, it is unclear from which statistical test the p values are obtained. This information is not anywhere of the results or methods sections. 6. Did you consider corrections for type 1 error rates, given the multiple comparisons performed? While these comparisons may not be the focus of the paper, I believe this is still an important consideration to be made even if corrections are not used.

	7. Why are there so many missing values (reported as undefined or NA) on Table 3? While upper bound confidence intervals may not be exact and some software yield such results, some of the missing values refer to medians (even though p values are reported). This warrants a revision of all results in this table. If there really is the case for inexact values, it should be stated in the legend. 8. Is it possible to present the median and interquartile range (or other summary) of the number of changes per preprint? The results on the most frequent types of changes are very interesting, but it led me to wonder whether there is an influence of a few poorly reported preprints as the overall number of changes are very low for most of the categories. 9. In lines 283-5, there's a statement that changes in outcome data are likely due to accumulating events over time. From these 20 trials with changes, how many of them had the preprint reporting that they are presenting interim analyses and that the outcome would continue to be evaluated? I think it could help you corroborate this claim. 10. In the text, when describing the two cases of changes in conclusions (results from Table 4), it would be helpful to also have at hand how many preprint trials are added to lead to such changes. In one case it was just one additional trial with proportionally very few patients added while in the other, there are 3 trials more with more substantial additions in patient. 11. You conclude that including preprints may affect the results of meta-analyses and encourage the use of preprints. However, the changes found can be positive or negative (improving or worsening certainty of evidence) and I believe it is important to highlight this also in the conclusion section. 12. Please consider openly sharing the raw data collected for the study. Minor: 13. Not all abbreviations are defined on first use, such as IQR and SD. While it might seem obvious to those used to them, it might not be so to others, particularly those for which English is a second/foreign language. 14. In the abstract, it may be useful to mention that certainty of evidence was assessed using the GRADE framework. 15. In many parts of the manuscript the authors state that preprints are relevant for health emergencies. I'd argue this is not the case and would like to ask the authors to consider if this is also not true for any disease, which will be an urgent issue for those affected by it. Of course, the generalizability of the results found is limited to COVID-19 but the results also open the possibility that preprints can be more used in general. 16. On lines 170-174, when listing possible outcomes, viral clearance and time to viral clearance are mentioned twice. 17. On lines 245-6, overall is written twice. 18. On Table 1, do you really need both the yes and no lines for 'trial registered'? As there's only one line for 'inpatient' and 'severity', I'd argue the same can be applied for trial registration. In my opinion, avoiding redundant information improves tables' readability. 19. On Table 1, why is there no statistical comparison in number of patients? 20. Are the results from table 2 in line with other estimates of the literature? This might help discussing the generalizability of the findings. I am aware of two reports (https://doi.org/10.1371/journal.pbio.3000959 for COVID-19 preprints, and https://doi.org/10.1101/833400 for bioRxiv before the pandemic), but there may be others. 21. It is unclear to me whether the example presented in Box 1 is

	representative of a change in description that led to a change in risk of bias assessment. Please clarify. 22. Table 4 is very difficult to read. Please consider whether the full version could be presented as supplementary and a summarised version could be kept in the main manuscript. For example, the column 'risk with standard care' seems mostly uninformative and this information could be easily presented in the table legend. Alternatively, you may want to consider breaking the table in multiple smaller tables. 23. When discussing the risk of fabrication or falsification, you may want to consider the feasibility of these assessments for systematic reviewers. While this is an important consideration, it might be more in the realm of responsibilities of peer reviewers. If this would become more commonplace, you might expect that the results of comparability of meta-analysis including or not preprints would change over time as more preprints than peer-reviewed publications would include false data. 24. The references need some extensive revisions. For example, references 9 and 49 have a version of record that can be preferentially cited. I also found that references 24, 26, 30 and 35 are incomplete.

REVIEWER 5	Reviewer 5
REVIEW RETURNED	26-Jun-2022

GENERAL COMMENTS	Thank you for the opportunity to review this interesting manuscript. In this project, Zeraatkar et al. conduct a large-scale evaluation of COVID-19 trials that were disseminated at preprints, preprints with corresponding publications, and only publications. In particular, the authors set out to determine the trustworthiness of impact of preprint trial reports during the COVID-19 pandemic. This is an interesting and important question, as the trade-offs between the benefits (rapid dissemination) and potential harms (lack of peer-review and editorial oversight) of preprints in clinical sciences have been debated extensively. This current project is a major effort that builds upon the herculean living review published in the BMJ. The authors analyze a number of characteristics – publication timing, the frequency of retractions, the differences between preprints and publications, and the impact of including or excluding preprints in meta-analyses. Overall, the authors conclude that they found no evidence that preprints provide less trustworthy results than published papers. Although this project contains a significant number of interesting findings, at times, the large number of outcomes makes it somewhat challenging to identify the focus. In particular, the conclusions and inferences in the abstract and key points sections primarily seem to relate to the analysis focused on how summary effect estimates and certain of evidence may change after including preprints (I find this to be the most interesting component of the study). Other evaluations have focused on the similarities between COVID-19 preprints and their subsequent journal publications (e.g., https://bmjopen.bmj.com/content/11/7/e051821 and https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.3001285; I was not involved in either of these previous evaluations). One of the key components of the paper is the retraction of preprints and publications. Although reassuring, there are only 4 retracted trials - 3 of which were published in peer-reviewed journals - in this sample. Unfortunately, this sample is not large enough to draw any meaningful inferences. However, the authors conclude that they "…found retractions to occur for both preprints and publications, suggesting that publication in a

	peer-reviewed journal alone does not indicate trustworthiness of a trial". While I completely agree that this is likely the case, I am not convince that the findings from this particular analysis provide suggestive evidence. A second focus of the paper is the concordance between preprints and their corresponding publications. However, these findings are mainly discussed in the main text/tables and are not included in the Abstract or Key findings. The authors note that 57% of trials had discrepancies in the reporting of key methods and results between the preprint and the later published trial report, with details in Supplement 5. However, it is hard to determine the importance of these changes. Overall, this manuscript reflects a significant effort and contains several interesting findings, especially related to the meta-analyses. However, it may be challenging to comment on the overall trustworthiness of preprints without knowing more about the preprints that were submitted but reject by journals. Perhaps the authors could consider making the focus more on how evidence changes, based on numerous factors, vs. the issue of trustworthiness? Please find some additional comments below: Methods: Line 121: it is great to see that the authors shared their full study protocol. Line 124: When were the searches updated? Line 194: The authors note that they "describe the number and types of discrepancies in key methods and results between preprint and published trial reports". Could the authors clarify the specific components that were compared in the methods section of the text? Line 201: How were retractions identified? Line 203: Why did the authors limit their sample to trials that report on interventions up to August 3rd 2021? It seems like this could have missed potentially eligible preprints and publications? Could the authors also clarify why the included vs. excluded evidence from preprints at one, three, and six months after the first trial addressing the drug of interest was made public? Line 217: It appears as if the authors focused on the direction of effect between meta-analyses including vs. excluding preprints. Would it be interesting to consider statistical significance and direction? Line 219: How were the 1%, 2%, and 0.5% values for benefit/harm selected? Results: Line 242: Perhaps clarify that it is August 3rd, 2021? Line 247: Is there a table with the findings reported in lines 247-250? Line 251: How did the authors determine the interim vs. non-interim results, as I don't remember seeing this in the methods section. Line 281: What are 'key results'? Are these primary endpoints? Line 319: The findings in this paragraph can be a little challenging to follow (e.g., the X.X more and X.X fewer). Would it make sense to report the estimates with and without preprints, instead of the differences between the

	two. This may improve clarity. Line 332: 9 cases out of how many total? Discussion: - Please see the comment above about the focus on 'trustworthiness'. - The authors could also consider comparing their findings to a previous evaluation focus on the concordance between preprints and their corresponding publications in the highest impact factor journals: https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2777629. Disclosure: I was an author on that manuscript. I defer to the authors whether they think including this is helpful/necessary. Table on page 22: This is a fascinating table. Is it worth mentioning in the main text that the 1 month analyses are often based on only 1 study vs. meta-analyses? Figures: These figures are terrific.

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1

Recommendation:

Comments:
This is an important paper that adds to the information around Covid-19 and is a matter of public interest, the review is written from a patient and public perspective. This research benefits patients and the public, as it is important to understand how evidence around the pandemic COVID-19 has been translated into clinical management guidelines and policy instruments and adds value by boosting public confidence in transparency and integrity of trusted research and clinical trials that can attract >/ 10% of people and less so from under represented groups in research.

Our response: We thank the reviewer for their positive assessment of our work.

Will it help our readers to make better decisions and, if so, how?

This study is helpful for the public as it provides concrete data around the relative effectiveness, reliability and credibility of pre-prints compared to published papers around results emerging from clinical trials for medications for Covid-19 looking at all stages of disease including severe and hospitalised admissions, during the time period of the analysis of July 2020 and 3 July 2021 looking at results from 356 trials globally, 101 pre-prints and 181 publications. It builds on looking at living systematic reviews 114 and network meta-analyses (SRNMAs) of drug treatments, antiviral antibodies and cellular therapies, and 115 prophylaxis for COVID-19, that provides real-time summaries 116 addressing the comparative effectiveness of treatments and prophylaxis for COVID-19. The article will help readers make better decisions by providing up to date accurate analysis on data from trials that do not solely focus on severity of illness but is wide encompassing and the article is balanced in suggesting that there is a need for evidence users and evidence intermediaries and the broader implementation science community to scrutinize preprints for falsified data and take into account the poor quality of evidence arising from e.g retrospective studies, so a need for sensitivity analysis and thinking subjectively is of paramount importance, the article provides reference to additional resources to help mitigate these concerns. Since many challenges exist in identifying the appropriate evidence, disseminating it to different stakeholders, implementing it and collaborating at break speed across different settings, the need to bridge the gap between what science knows and what is the know-do gap as determined by evidence makers is critically important and there is a need to make better decisions in faster time frames so that future pandemics can be better handled.

Will the article add enough to existing knowledge?

The study sheds light on the rates of precision and estimates in pre-prints compared to published articles on trial results and the effect of pre-prints on meta-analytic studies in relation to evidence quality using GRADE system, and in particular in relation to falsified data and errors addressing issues around publication bias and therefore adds to existing knowledge. It is interesting to see that trials with industry sponsors and government funding and those reporting on severe Covid-19 were published faster although there were discrepancies in the reporting of methods and results between pre-print and published reports although overall effects do not appear to be affected a lot which is re-assuring. I would have liked to have seen the findings discussed in light of recent Cochrane Convenes report and some of the other work with regards to the Global Evidence Commission to address future societal challenges as that may have provided an additional layer of contextual analysis of interests to the target audience and in particular

perhaps helped create better incentives to improve clinical research infrastructure in different countries and equitably distributed capacities to help intermediaries, evidence bodies and government bodies better collaborate to produce, share and use evidence. In addition to perhaps more emphasis on how to mitigate the bias seen itself in some of the results, which is primarily attributed to the fact that more than 2/3 of the trials created had an open label design within the framework of ensuring future sustainability and agility of working practices when another global emergency occurs. The Global Evidence Commission found that "Global commissions are also silent on the need to have the protocols for randomized- controlled trials and other study designs, as well as national evidence-support systems and a broader global evidence architecture, 'ready to go' or already in use" section 7.2 (20) recommendations Global Evidence Commission.

The methods used seem appropriate though as a lay am not fully qualified to comment on this, but it would appear that they do address the nature of both confounding and publication bias sufficiently. It may be useful to refer to another recommendation of the Global Evidence commission here that has called for journal publishers to improve the ways in which they support the use of best evidence. "Journals can mandate the use of reporting guidance and critical-appraisal checklists by reviewers, the placement of single studies in the context of evidence syntheses, and the sharing of anonymized study data. They can also commit to publishing non-positive research reports and replication studies, avoiding 'spin,' and acting quickly when apprised of scientific misconduct. Journals need to find a timely way to publish updates to living evidence products. Journals also need to ensure that publication delays never hinder the public sharing of evidence that is urgently needed for decision-making (and reciprocally that public sharing does not preclude later publication in a journal)" section 7.2 (23) recommendations Global Evidence Commission.

Scientific reliability

The design of the study includes many confounding variables and the authors address some of the limitations of the study well such as the fact that not all servers of published reports are included in the WHO Covid-19 database and that there may still be errors contains in some that are in the database, but does not give that much detail on approaches that could be taken to eliminate bias so e.g. assigning participants to alternative interventions using randomly generated interventions. I would have liked to see more refence to some of the current rationale behind increasing transparency of reporting requirements as a compulsory requirement for registration with regulatory authorities so the UK MHRA for example has made this a pre-requisite as well as patient inclusion and it would be perhaps interesting to other ways that traditional reporting can better integrate outcomes that matter to patients though this may be outside the scope of the current study. This work is able to meaningfully contribute to the research field around data integrity and transparency of trial results gathered during the Covid-19 pandemic and likely to lead to incremental change in perception of and usage of pre-prints in other fields in the medium term. Given the speed with which some of the early clinical research studies and RCTS took place, it has been previously reported that PPI patient public involvement practices took a bit of a back seat, so any future analysis of other variants in the current time frame of 2022/23 should look at how patient reported outcomes and patient experience data could be better embedded in the design, identification and setting of end points in particular in relation patients that experienced mild disease and patients that now meet criteria for PACS known as LongCOVID.

Importance of the work to general readers

The inclusion of patients is notable, in that they have helped select the outcomes which are listed as being mortality, mechanical ventilation, duration of hospitalization, time to resolution, clinical improvement and virological outcomes and we are told that patients were involved in “generalization of recommendations part of SRNMA” but we are not told more details, which I feel is an omission.

> **Our response:** Since patients were involved primarily in the parallel guidelines, we do not provide additional details. If the reviewers and editors feel that this is important, we can provide additional details. We have revised to provide a citation to the parallel guidelines in which readers can find more information on how patients were involved.

Revision	Page	Line
Patients were involved in outcome selection, interpretation of results, and the generation of parallel recommendations, as part of the parallel SRNMA and guidelines (23).	10	259 to 260

Patients and member of the public want to know what they can present to their doctor to get better advice and medication, but they may not understand the finer nuances in either the publication authorship process of the fact that the totality of the body of evidence is taken into account when forming national clinical guidelines, or how confirmation and information bias contribute to information mismanagement. It may have been useful to think about data visualization in the form of an infographic and other ways of communicating the content of the article as the trial results tables are fairly technical and dense for the average lay reader.

> **Our response:** We would like to clarify that the target audience for our study is not patients. Rather, systematic reviewers and guideline developers. In that sense, infographics for communicating results may not be critical. If, however, reviewers and editors feel differently we can develop infographics.

Does the article read well and make sense? Does it have a clear message?

As the topic and methods of evidence and the broader networks around it such as multilateral organizations, development banks, the Organisation for Economic Co-operation and Development, the G20,national and sub-national government policymakers, evidence intermediaries, including those who do not currently function as evidence intermediaries (such as journalists for the most part), evidence producers (such as units engaged in producing and supporting the use of data analytics, modelling, evaluation, behavioural / implementation research, qualitative insights, evidence syntheses, technology assessment / cost-effectiveness analysis, and guidelines makers) are largely not that well understood by citizens at large or members of the public and sometimes patient advocacy groups, it may have been useful for the patients involved in the outcome selection to have written a short lay summary for the article even though the target audience is primarily healthcare professionals as this may represent an opportunity for broader engagement work. I also feel that given some movements by national and European patient organisations and aggregator of interests such as EURORDIS, representing the voice of rare disease patients in Europe, EFNA representing neurological patient associations and other EU government funded groups in structures programs by the EU IHI and ERAnets it might have been good to try and compare and feed some of their concerns and positions in relation to interruption of clinical trials and to look at further analysis of differences in particular regions and in closer detail around some of the outcomes chosen in relation to clinical improvement and virological outcomes as people with existing conditions may have not met full inclusion criteria for some of the work and/or the faster results in relation to potential therapeutic solutions that meet rigorous scientific standards from e.g RCTs can impact actual health service improvement and delivery of care, for example many patients in clinically vulnerable categories e.g rare disease patients may have been awaiting direction of guidance from appropriate medical bodies e.g. if for example not being able to use a particular type of medication due to it being contra-indicated, so it is important to highlight the extra re-assurance that data can bring to patients and their families to keep up with sometimes conflicting information messages that may not have been tailored

to their condition, due to for example not having access to specialist care as a result of the pandemic or their own clinical trial being halted for allocation of resources in the hospital elsewhere. This demographic of patients may want to see specific results tailored with additional data pertaining to their condition so an additional level of filtering and also reference to what that means in terms of service improvement within the UK picture please see ARDENT by the Cambridge Rare Disease network detailing the effect the pandemic has had on rare disease patients.

Given the international emergency of the pandemic and the imperative for swift evidence-based medicine, it is imperative to ensure quicker results on what works in terms of clinical management. This article has a clear message that data from preprints is reliable and can further enhance trust and transparency in clinical trials reporting, a subject that is becoming increasingly important with reference to patient advocacy groups if one looks at work such as the Good Clinical Trials Initiatives guidelines and advocacy campaigns Transparimed. This study can be used to guide national policies on the subject matter at hand and adds to the knowledge base around the importance of efficient but flexible and agile methods for evidence synthesis that are grounded in methodological rigour but can adapt to ever changing circumstances as those dictated in international emergencies like pandemics. The research team are very strong

Reviewer: 2

Recommendation:

Comments:
This article by Zeraatkar and colleagues describes a methodological study that aimed to assess the trustworthiness and impact of COVID-19 RCT preprints. The paper is original, clearly laid out, well written and interesting. The study appears of high standard, with a protocol included as supplemental material and data underpinning the study available on OSF – it is good to see a methodological study conducted in this fashion. I enjoyed reading the paper but have a concern regarding whether the authors have answered their primary research question and have a few other additional minor comments. These are detailed below:

> **Our response:** We thank the reviewer for their positive assessment of our work.

Major comment

1. Is “trustworthiness” adequately addressed?
This paper hinges around assessing the trustworthiness of preprints. It is first defined at the bottom of page 5, as “complete and consistent reporting of key aspects of the methods and results” and is measured in this study by comparing the preprint methods/results against the published paper methods/results. I am not convinced that this is really trustworthiness, and instead my opinion is that this is looking to see if there is consistency between the preprint and published trial. Is the preprint trustworthy just because the peer-reviewed paper is similar? Surely to be able to answer the key question of “are preprints trustworthy” you would need to assess the data underpinning the preprint. I worry that referring to this as trustworthiness exceeds what the authors have done and readers who see only the headline message/abstract may misinterpret the findings.

> **Our response:** We assumed that important changes in the reporting of key methods and results between preprints and their subsequent publications indicates potential issues with trustworthiness. For example, such situations may indicate inaccurate or incorrect reporting in the preprint that was later corrected during the peer review process. If most preprints have important differences with their later publications, this would suggest that preprints may not be trustworthy.
>
> We also agree, however, that the construct of trustworthiness is broad and includes other aspects of trial design and reporting in addition to differences between preprints and later publications.
>
> We have revised to clarify this issue in the abstract, introduction, and discussion sections of the

manuscript. If the reviewers and editors have alternative terminology to replace 'trustworthiness' and 'impact', we are happy to oblige.

We have revised to explicitly acknowledge these issues.

Revision	Page	Line
Purpose: To assess the trustworthiness (complete and consistent reporting of key aspects of the methods and results between preprint and published trial reports) and impact (effects of preprints on meta-analytic estimates and the certainty of evidence) of preprint trial reports during the COVID-19 pandemic.	3	62 to 64
Preprints that are subsequently published in journals may be the most rigorous and may not represent all trial preprints—particularly those that remain unpublished.	3	83 to 85
To assess preprint trustworthiness, we compared reporting of key aspects of the methods and results between preprint and published trial reports. We assumed that important changes in the reporting of key methods and results between preprints and their subsequent publications may indicate inaccurate or incorrect reporting in the preprint that was later corrected during the peer review process and potential issues with trustworthiness. We acknowledge that differences between preprints and published reports, however, only addresses one component of the construct of trustworthiness and there are other factors, in addition to differences between preprints and published reports that may make a trial untrustworthy.	18	468 to 474
We found no compelling evidence that there are important discrepancies between preprint and published trial reports.	20	529 to 530

Minor comments

2. Prominence given to point estimates

When comparing meta-analysis results from MAs that include/exclude preprints, the authors focus on point estimates (and direction of point estimates) as their approach to determine if conclusions change. I strongly recommend that the authors consider the variability around the point estimate before they claim benefit/harm/no effect.

Our response: For our meta-analyses including/excluding preprints, in addition to reporting on whether the point estimates were different, we report on differences in statistical significance and GRADE ratings, including imprecision, which considers confidence intervals. We used the GRADE minimally contextualized approach to determine whether meta-analyses that include/exclude preprints yield consistent results (Zeng L, Brignardello-Petersen R, Hultcrantz M, Siemieniuk RAC, Santesso N, Traversy G, Izcovich A, Sadeghirad B, Alexander PE, Devji T, Rochwerg B, Murad MH, Morgan R, Christensen R, Schünemann HJ, Guyatt GH. GRADE guidelines 32: GRADE offers guidance on choosing targets of GRADE certainty of evidence ratings. J Clin Epidemiol. 2021 Sep;137:163-175. doi: 10.1016/j.jclinepi.2021.03.026. Epub 2021 Apr 20. PMID: 33857619.). Our approach considered the direction and magnitude of effect and the confidence intervals in the rating of imprecision. We have revised to clarify.

Revision	Page	Line
We used a minimally contextualized approach to make judgements about imprecision (28). This approach considers whether confidence intervals include the null effect and thus does not consider whether plausible effects, captured by confidence intervals, include both important and trivial effects.	10	250 to 252

Except for two cases, all meta-analyses including and excluding results from unpublished preprints produced point estimates that were consistent as to whether they indicated benefit, no appreciable effect, or harm.	14	337 to 339
Four of sixty meta-analyses had results that were statistically significant with preprints and not statistically significant without preprints, or vice versa.	14	351 to 352
We judged nine of 60 meta-analyses to have different ratings of the certainty of evidence when preprints were included versus excluded.	14	355 to 356
Between meta-analyses including versus excluding preprints, judgements related to the GRADE risk of bias domain differed only for one meta-analysis (remdesivir vs. standard care/placebo for mechanical ventilation at 6 months).	14	364 to 366
Between meta-analyses including versus excluding preprints, judgements related to the GRADE imprecision domain differed for 13 of 60 meta-analyses.	15	370 to 371

3. Greater discussion around meta-analysis

Here are some things you may consider commenting on in the discussion that the paper touches on, but that I think could come out clearly in the discussion and (in my opinion) would be of interest to readers.

• Unpublished data in a meta-analysis. A number of meta-analyses of COVID-19 trials included data from unpublished trials (i.e., not even published as preprints). Whilst this is a step even further, it may be an interesting discussion point. See for example the IL-6 prospective meta-analysis (Shankar-Hari et al. JAMA 2021).

Our response: We have revised to include a discussion of prospective meta-analyses.

Revision	**Page**	**Line**
Review authors who are concerned about publication bias or are wanting to make decisions within timeframes that may not be conducive to peer review and publication may also consider conducting prospective meta-analyses—meta- analyses that are conducted using an inception cohort of registered trials and incorporate unpublished data from investigators (48). During the COVID-19 pandemic, investigators have conducted such prospective meta-analyses to address the effectiveness of corticosteroids and IL-6 receptor blockers for COVID-19 (49, 50).	17	428 to 433

• Timing of when to conduct a MA. The authors focus around 4 crude timepoints for demonstration and it is interesting to see how the evidence accumulates over time. Do the authors have any comments on when you should conduct a meta-analysis? There are various approaches that tackle timing such as living systematic reviews, framework for prospective adaptive meta-analysis (FAME, Tierney et al. PLoS Med 2021) and ALL-IN meta-analysis (ter Schure and Grunwald 2021).

Our response: The COVID-19 pandemic instigated a surge of research activity and pressures to make decisions very quickly and within timeframes that were not compatible with the traditional timeframes for peer review and publication. Our choice of timepoints were informed by the experiences of the linked living guidelines (Agarwal A, Rochwerg B, Lamontagne F, Siemieniuk R A, Agoritsas T, Askie L et al. A living WHO guideline on drugs for covid-19 BMJ 2020; 370 :m3379 doi:10.1136/bmj.m3379) and the timeframes within which the panel made recommendations. We agree, however, that these timepoints were arbitrary. In our experience, the timepoint at which meta-analyses are conducted should be guided by the timepoint at which the guideline panels need to make recommendations or decision-makers need to make decisions. We have revised to clarify.

Revision	**Page**	**Line**

The choice of timepoints was informed by timeframes within which guideline developers needed to issue recommendations (23).	9	219 to 221

4. Justification for search date

The last search was carried out on August 3rd 2021. Given this is linked to the living systematic review and network meta-analysis, I would have expected that the search would be far more up to date. The authors do comment that these searches are carried out daily, so it is not clear to me why they have chosen this date. Updating this or providing justification for this date would be useful.

Our response: The living COVID-19 NMA performs daily searches for relevant COVID-19 literature. This study, however, required the extraction of additional data from trials that are not routinely collected for the living COVID-19 NMA. For feasibility, we were only able to perform these extractions for trials that we had identified up to August 3rd, 2021. We have revised to clarify.

Revision	Page	Line
While the parallel living SRNMA performs ongoing daily searches, we pragmatically limited our search because it was no longer feasible to continue to collect additional data from preprints beyond this timepoint.	7	156 to 158

5. Risk of bias

Risk of bias was specified as high risk for many of these trials for an outcome of mortality due to the open-label nature. Personally, I disagree with this judgement, and do not consider that the open-label nature could introduce bias into an outcome as objective as mortality. I do recognize that the authors are consistent with their previous approach but still suggest they reconsider.

Our response: The living COVID-19 NMA team considered whether unblinded trials should be rated at high or low risk of bias. Ultimately, the team decided to consider such trials at high risk of bias due to potential for imbalances in co-interventions that may influence even objective outcomes like mortality.

Specific comments

1. Retractions

Is there a typo in the results section? think there is a typo and it should say six instead of four

Our response: There were four retracted trials. The numbers do not add to four because there is overlap in the numbers. For example, one of the trials favipiravir with ivermectin and so reported on both favipiravir and ivermectin.

2. Supplement 3 – Is some text missing here? The RHS column is empty for some fields – e.g., p45 Probably low risk of bias

Our response: We do not see any empty columns in supplement 3.

3. Supplement 4 – There is some missing text here, “43,849 records excluded for not being…”

Our response: We have revised to address the missing text.

4. Supplement 6 – fixed effect models have also been carried out. Please specify what model this is based on in the methods.

Our response: Supplement 6 presents results for both fixed effect and random effects models.

Reviewer: 3

Recommendation:

Comments:
The main new contribution of this paper is the evaluation of the effect of preprints on meta-analytic estimates. This paper did not assess "trustworthiness" of the trials. This term usually refers to the underlying data, not the reports (see, for a summary, https://www.cochranelibrary.com/cdsr/doi/10.1002/14651858.ED000152/full). The current paper assessed only reports (preprints and manuscripts). The objective of this study is more accurately described as assessing the reporting of preprints and publications, not their trustworthiness. The word trustworthiness should not be used in the title or conclusion of the manuscript. One could also argue that the word "impact" in the title is misleading as this paper studies a narrow impact of preprints – on meta-analytic estimates and GRADE ratings, not on social media indicators or guidelines or decisions, as other studies have done.

Our response: We assumed that important changes in the reporting of key methods and results between preprints and their subsequent publications indicates potential issues with trustworthiness. For example, such situations may indicate inaccurate or incorrect reporting in the preprint that was later corrected during the peer review process. If most preprints have important differences with their later publications, this would suggest that preprints may not be trustworthy.

We agree, however, that the construct of trustworthiness is broad and includes other aspects of trial design and reporting in addition to differences between preprints and later publications.

We also agree that the construct of 'impact' is broad, and we only considered 'impact' on meta-analytic estimates and the certainty of evidence. We have revised to clarify this issue in the limitations.

We have revised to clarify this issue in the abstract, introduction, and discussion sections of the manuscript. If the reviewers and editors have alternative terminology to replace 'trustworthiness' and 'impact', we are happy to oblige.

We have revised to explicitly acknowledge these issues.

Revision	**Page**	**Line**
Purpose: To assess the trustworthiness (complete and consistent reporting of key aspects of the methods and results between preprint and published trial reports) and impact (effects of preprints on meta-analytic estimates and the certainty of evidence) of preprint trial reports during the COVID-19 pandemic.	3	62 to 64
Preprints that are subsequently published in journals may be the most rigorous and may not represent all trial preprints—particularly those that remain unpublished.	3	83 to 85
To assess preprint trustworthiness, we compared reporting of key aspects of the methods and results between preprint and published trial reports. We assumed that important changes in the reporting of key methods and results between preprints and their subsequent publications may indicate inaccurate or incorrect reporting in the preprint that was later corrected during the peer review process and potential issues with trustworthiness. We acknowledge that differences between preprints and published reports, however, only addresses one component of the construct of trustworthiness and there are other factors, in addition to differences between preprints and published reports that may make a trial untrustworthy.	18	468 to 474

We found no compelling evidence that there are important discrepancies between preprint and published trial reports.	20	529 to 530
Our assessment of the impact of preprints focuses only on the impact of preprints on meta-analytic estimates, the certainty of evidence, and decision-making and does not consider other aspects of impact, such as number of citations or Altmetrics.	19	495 to 497

INTRODUCTION

The introduction to the paper needs to be updated as it does not summarize prior research evaluating COVID-19 preprints or comparing COVID-19 preprints to their final publications. See below, and there are likely more recent studies. (The preprint of one of these studies is cited in the discussion section, but, oddly, the final publication is not).

Bero L, Lawrence R, Leslie L, et al. Cross-sectional study of preprints and final journal publications from COVID-19 studies: discrepancies in results reporting and spin in interpretation. BMJ Open 2021;11:e051821. doi:10.1136/ bmjopen-2021-051821

Nicolalde B, Anazco D, Mushtaq M, et al. Citations and publication rate of preprints on pharmacological interventions for COVID-19: the good, the bad and, the ugly. Res Sq 2020;version 2.

Kataoka Y, Oide S, Ariie T, et al. COVID-19 randomized controlled trials in medRxiv and PubMed. Eur J Intern Med 2020;81:97–9.

Our response: We have revised to include the listed references. The study by Nicolalde and colleagues primarily addresses basic science research, and its findings may not be applicable to randomized trials. We have still cited it.

METHODS

The protocol is submitted as an appendix but was not published. Why not publish in OSF or on some other open access platform? These platforms also allow comments and are publicly accessible.

> **Our response:** We have now uploaded our protocol to OSF, in addition to sharing it as a supplement to the manuscript. See: https://osf.io/ambnk

Line 135-136: How were “concerns regarding research integrity” monitored via Epistemonikos and the WHO database? The referenced supplement 2 does not provide any information on searching for concerns regarding research integrity. Furthermore, no data are reported on these concerns, so this phrase should be deleted from the methods section. As noted in https://www.cochranelibrary.com/cdsr/doi/10.1002/14651858.ED000152/full, retractions are only the tip of the iceberg for identifying problematic studies.

Our response: Epistemonikos and the WHO database were only used to monitor for retraction notices or for publications. We have revised to clarify.

Revision	Page	Line
Our search is supplemented by ongoing surveillance of living evidence retrieval services, including the Living Overview of the Evidence (L-OVE) COVID-19 platform by the Epistemonikos Foundation (https://app.iloveevidence.com/loves/5e6fdb9669c00e4ac072701d) and the Systematic and Living Map on COVID-19 Evidence by the Norwegian Institute of Public Health (https://www.fhi.no/en/qk/systematic-reviews-hta/map/). Using the above sources, we monitor for retraction notices. Supplement 2 includes additional details of our search strategy.	6	132 to 137

The search for preprints does not include a comprehensive list of preprint servers (Appendix 2). This is mentioned as a limitation in the discussion section.

Our response: We have maintained our living SRNMAs of COVID-19 trials for over two years. During this time, we have been informed of any trials—in preprint or in publication—that were not identified by our search strategy. Our living SRNMA team and the parallel guideline team includes trialists, research librarians, and information scientists have vetted the comprehensiveness of our search strategy. The search strategy has also been peer-reviewed by the BMJ through all iterations of updates. Although we did not search all preprint servers, our search includes preprint servers that may plausibly include COVID-19 therapeutic trials. We have revised to clarify.

Revision	Page	Line
Although the WHO COVID-19 database is a comprehensive source of published and preprint literature, it does not include all preprint servers—though preprint servers not covered by our search address other subjects and are unlikely to include COVID-19 trials.	18	465 to 467

The method for linking preprints with subsequent publications needs to be clarified. Does the living systematic review use a study-based register, rather than a records based register? The value of a study based registers is that all versions of all records linked to a publication would be identified. What about multiple versions of preprints and articles? What were the selection criteria for forming the pairs? Line 173 (in the data collection section) states “For preprints with more than one version, we extracted data from the first version of the preprint, which is the least likely to have been modified in response to peer review” How were multiple versions of manuscripts handled?

Our response: We use a record-based register. We document, however, when records report on the same trial. We did not identify multiple versions of identical manuscripts. We only identified post-hoc analyses of previously identified trials. To link preprints and publications, two reviewers read the preprint and manuscript independently and in duplicate and collected information on the following variables: authorship list, trial registration, trial name, country of recruitment, recruitment dates, interventions investigated, and baseline patient characteristics. These variables were subsequently used to link preprints and publications. To avoid errors, we avoid any automated algorithms for identifying preprint and publication pairs. We have revised to clarify.

Revision	Page	Line
Reviewers also link preprint reports with their subsequent publications based on trial registration numbers, the names of investigators, recruiting centres and countries, dates of recruitment, and baseline patient characteristics. When links between preprints and subsequent publications are unclear, we contact trial authors for confirmation. Reviewers resolve discrepancies by discussion or, when necessary, by adjudication with a third-party reviewer.	6	143 to 147

A major gap is that information on harm outcomes reported was not collected. Lines 169-174 list the included outcomes. Although the “direction” of an outcome was assessed (e.g., a drug study designed to

determine if a drug decreased mortality would measure an increase in mortality or a decrease in mortality), this is not the same as outcomes that specifically assess harms. Previous studies have found discrepancies in reporting harm outcomes between COVID 19 preprints and their final publications (eg, Bero 2021).

> **Our response:** The objective of our study was to look for discrepancies in preprints and publications that may affect the interpretation of trials and how trials inform decision-making. For results, we focused on differences in the reporting of key outcomes that may affect decision-making. The choice of outcomes that we examined was informed by the outcomes that the parallel guideline panel (Agarwal A, Rochwerg B, Lamontagne F, Siemieniuk R A, Agoritsas T, Askie L et al. A living WHO guideline on drugs for covid-19 BMJ 2020; 370 :m3379 doi:10.1136/bmj.m3379) selected as being important or critical to decision-making.
>
> Harm outcomes are most often specific to the intervention being investigated and it would not have been feasible for our team to examine discrepancies for specific harm outcomes for each type of intervention.
>
> We have revised to acknowledge this issue as a limitation.

Revision	Page	Line
Our assessment of differences in key results between preprints and publications was limited to the outcomes that were included in our living SRNMAs. While these outcomes were identified as being important or critical to decision-making by co-authors of the living SRNMA and the parallel guideline, they do not include adverse events. It is possible that there may be differences in such outcomes between preprints and publications (52). Further, it is also possible that there may be other aspects of the reporting of results (e.g., baseline characteristics of patients) that may be different between preprints and publications.	19	484 to 490

The patient involvement description does not appear relevant to this study, but rather to the living SRNMA and guidelines.

> **Our response:** Patients were only involved as part of the linked SRNMAs and guidelines. Since patients' involvement informed our choice of outcomes and magnitudes of effect (both of which are relevant to this study), we have retained the statement on patient involvement. If the editors agree, we can remove this statement.

The analysis focuses on comparing the reporting of preprints and their corresponding publications, as well as the calculation of meta-analytic estimates that exclude or include data from preprints at different time intervals. "Key methods" appear to be the same as risk of bias criteria "Key methods include description of the randomization process and allocation concealment, blinding of patients and healthcare providers, extent of and handling of missing outcome data, blinding of outcome assessors and adjudicators, and prespecification of outcomes and analysis. The "key results" analyzed focus on the meta-analytic estimates but do not include other aspects of outcome reporting that could vary (see Bero 2021).

> **Our response:** We compare the key results reported in preprint reports of trials and their later publications. These results are independent of our meta-analyses. Our meta-analyses investigate whether including versus excluding evidence from unpublished preprints may affect decision-making.
>
> The objective of our study was to look for discrepancies in preprints and publications that may affect the interpretation of the trial and how the trial informs decision-making. We focused on aspects of the study that we judged to be critical to decision-making. For methods, we judged these aspects to be related to the risk of bias criteria. Discrepancies in the descriptions of methods that result in different risk of bias judgements have the potential to importantly affect the interpretation of the trial. For results, we focused on differences in the reporting of key outcomes that may affect decision-making. The choice of outcomes that we examined was informed by the outcomes that the parallel guideline panel, including patient representatives, selected as being important or critical to decision-making. We agree, however, that there may other results (e.g., baseline characteristics of patients) that may also be different between preprints and publications. We have revised to clarify

and to acknowledge this limitation.

Revision	Page	Line
We focused on the same outcomes as our living SRNMA and linked guidelines that were identified as being important or critical for decision-making by the review authors and authors of the parallel guidelines, including patient partners: mortality, mechanical ventilation, adverse events leading to discontinuation, admission to hospital, viral clearance, hospital length of stay, ICU length of stay, duration of mechanical ventilation, time to symptom resolution or clinical improvement, days free from mechanical ventilation, and time to viral clearance (18-21).	7	175 to 181
Our assessment of differences in key results between preprints and publications was limited to the outcomes that were included in our living SRNMAs. While these outcomes were identified as being important or critical to decision-making by co-authors of the living SRNMA and the parallel guideline, they do not include adverse events. It is possible that there may be differences in such outcomes between preprints and publications (52). Further, it is also possible that there may be other aspects of the reporting of results (e.g., baseline characteristics of patients) that may be different between preprints and publications.	19	484 to 490

RESULTS

Line 253. How was the determination that a preprint reported on interim results made? Did this have to be stated in the preprint or were the Ns for the outcomes compared between preprints and final publications? If n's were compared, were increases or decreases for n's for reported outcomes both considered to indicate interim results?

Our response:

Revision	Page	Line
Our assessment of differences in key results between preprints and publications was limited to the outcomes that were included in our living SRNMAs. While these outcomes were identified as being important or critical to decision-making by co-authors of the living SRNMA and the parallel guideline, they do not include adverse events. It is possible that there may be differences in such outcomes between preprints and publications (52). Further, it is also possible that there may be other aspects of the reporting of results (e.g., baseline characteristics of patients) that may be different between preprints and publications.	19	484 to 490

The differences in outcome reporting should be described. An example is given (box 2), but no information on the details of these discrepancies. For example, did the reported outcomes differ in the metrics used, measures, time point at which assessments were made, population (subgroup) analyzed, N, etc? Comparing just the statistics and numerical results reported could miss key discrepancies in outcome reporting.

> **Our response:** We report this information in several layers of details depending on the degree of information in which readers are interested. We report the number of discrepancies in outcomes in-text. We report differences in numerical results for mortality and mechanical ventilation (the two most commonly reported outcomes) in figure 1. Supplement 5 presents more detailed categorization of the types of discrepancies. We report detailed discrepancies in our master data file that is freely available. Using our master data file, interested readers can also identify the specific trials with discrepancies. Our master data file is available on OSF.

Line 292. 11 publication / preprint pairs showed reporting of “an outcome” in the publication that was not included in the preprint. Clarify if this was one, at least one, or multiple outcomes between each preprint-publication pair. Also, what was the nature of these outcomes? Additional time points, harm, different metrics or cutoffs used? This critical information to determine if preprints could be missing important information compared to subsequent publications is not reported.

> **Our response:** The reporting of additional outcomes had to do with a publication reporting one or more of our outcomes of interest that were not reported in the preprint. Our outcomes of interest included mortality, mechanical ventilation, adverse events leading to discontinuation, admission to hospital, viral clearance, hospital length of stay, ICU length of stay, duration of mechanical ventilation, time to symptom resolution or clinical improvement, days free from mechanical ventilation, and time to viral clearance. We have revised to clarify.

Revision	Page	Line
Other differences between preprints and publications in key results included the publication reporting at least one additional key outcome that was not included in the preprint (n=11; 14.9%).	13	310 to 312

It is not surprising that inclusion of preprints in the MA would decrease imprecision as this criteria primarily relates to the quantity and heterogeneity of data included in the meta-analysis. If less data are included in the meta-analysis, one would expect wider confidence intervals and / or greater heterogeneity.

> **Our response:** The GRADE imprecision domain is affected by the quantity of data, indicated by the width of confidence intervals (Guyatt GH, Oxman AD, Kunz R, Brozek J, Alonso-Coello P, Rind D, Devereaux PJ, Montori VM, Freyschuss B, Vist G, Jaeschke R, Williams JW Jr, Murad MH, Sinclair D, Falck-Ytter Y, Meerpohl J, Whittington C, Thorlund K, Andrews J, Schünemann HJ. GRADE guidelines 6. Rating the quality of evidence--imprecision. J Clin Epidemiol. 2011 Dec;64(12):1283-
> 93. doi: 10.1016/j.jclinepi.2011.01.012. Epub 2011 Aug 11. Erratum in: J Clin Epidemiol. 2021 Sep;137:265. PMID: 21839614.). Heterogeneity or inconsistency in results across trials would

impact the inconsistency domain (Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, Alonso-Coello P, Glasziou P, Jaeschke R, Akl EA, Norris S, Vist G, Dahm P, Shukla VK, Higgins J, Falck-Ytter Y, Schünemann HJ; GRADE Working Group. GRADE guidelines: 7. Rating the quality of evidence--inconsistency. J Clin Epidemiol. 2011 Dec;64(12):1294-302. doi: 10.1016/j.jclinepi.2011.03.017. Epub 2011 Jul 31. PMID: 21803546.). We did not observe any evidence of important inconsistency. We had anticipated that including preprints in meta- analyses would reduce precision because meta-analyses including preprints would be informed by more data and thus have narrower confidence intervals.

We have revised to clarify.

Revision	Page	Line
We used the GRADE approach to assess the certainty of evidence across risk of bias (limitations in trial design leading to systematic under- or over-estimation of treatment effects), inconsistency (heterogeneity in results reported across trials), indirectness (differences between the question addressed in trials and the question of interest), imprecision (width of confidence intervals), and publication bias (propensity for studies with statistically significant results, interesting results, or results that support a particular hypothesis to be published, published faster, or published in journals with higher visity) and assessed whether including versus excluding preprint reports led to differences in ratings of the overall certainty of evidence, judgments related to specific GRADE domains, and whether differences in ratings are likely to impact decision making (i.e., evidence rated as high/moderate versus low/very low) (27).	9 to 10	241 to 250

Another relevant question would be to determine if final publications rather than preprints should be included in meta-analyses. This could be done by comparing the analytic estimates between meta-analyses that include preprints vs. meta-analyses that include the final publications of these preprints.

Our response: While addressing the question of whether final publications or preprints of the same trial should be included in systematic reviews was outside the scope of our work, we have revised to acknowledge this important remaining question in the discussion.

Revision	Page	Line
In situations where there are discrepancies between preprints and publications, systematic reviewers and guideline developers will also need to consider whether the results of preprints or publications are more trustworthy and should be	20	519 to 525
incorporated in the meta-analysis. In such situations, systematic reviewers and guideline developers may assume that changes between preprints and publications are due to errors or inaccuracies in the reporting in the preprint that were later corrected during peer review. Future research should address whether including results from preprints or final publications impacts overall findings and decisions.		

DISCUSSION

The findings and implications are overstated as this paper does not assess trustworthiness, and only narrow measures of impact.

Our response: We assumed that important changes in the reporting of key methods and results between preprints and their subsequent publications indicates potential issues with trustworthiness. For example, such situations may indicate inaccurate or incorrect reporting in the preprint that was later corrected during the peer review process. If most preprints have important differences with their later publications, this would suggest that preprints may not be trustworthy.

We agree, however, that the construct of trustworthiness is broad and includes other aspects of

trial design and reporting in addition to differences between preprints and later publications.

We also agree that the construct of 'impact' is broad, and we only considered 'impact' on meta-analytic estimates and the certainty of evidence. We have revised to clarify this issue in the limitations.

We have revised to clarify this issue in the abstract, introduction, and discussion sections of the manuscript. If the reviewers and editors have alternative terminology to replace 'trustworthiness' and 'impact', we are happy to oblige.

We have revised to explicitly acknowledge these issues.

Revision	**Page**	**Line**
Purpose: To assess the trustworthiness (complete and consistent reporting of key aspects of the methods and results between preprint and published trial reports) and impact (effects of preprints on meta-analytic estimates and the certainty of evidence) of preprint trial reports during the COVID-19 pandemic.	3	62 to 64
Preprints that are subsequently published in journals may be the most rigorous and may not represent all trial preprints—particularly those that remain unpublished.	3	83 to 85
To assess preprint trustworthiness, we compared reporting of key aspects of the methods and results between preprint and published trial reports. We assumed that important changes in the reporting of key methods and results between preprints and their subsequent publications may indicate inaccurate or incorrect reporting in the preprint that was later corrected during the peer review process and potential issues with trustworthiness. We acknowledge that differences between preprints and published reports, however, only addresses one component of the construct of trustworthiness and there are other factors, in addition to differences between preprints and published reports that may make a trial untrustworthy.	18	468 to 474
We found no compelling evidence that there are important discrepancies between preprint and published trial reports.	20	529 to 530
Our assessment of the impact of preprints focuses only on the impact of preprints on meta-analytic estimates, the certainty of evidence, and decision-making and does not consider other aspects of impact, such as number of citations or Altmetrics.	19	495 to 497

The recommendation that systematic reviewers and guideline developers consider evidence from preprints (line 386-387) is supported by the data presented in this paper. The discussion of false and fabricated data (lines 392-400) appears to be an add on that is not directly related to this paper. The cursory mention of methods for detecting fabrication and falsification in clinical trials (box 3), does not cover other detritmental research practices which may make a study actually untrustworthy.

Our response: We thought it may be useful to readers to be directed to some tools for assessing data fabrication and falsification issues. This issue became very relevant in COVID-19 due to some false and fabricated trials being used to support the treatment of COVID-19 patients with hydroxychloroquine and ivermectin. If the reviewers and editors feel strongly, we can remove details on detecting false or fabricated data.

The section on "relation to previous work" is inadequate. It is incomplete (see mention above re other relevant studies) and inaccurately describe one of the studies as assessing spin, when it actually assessed characteristics of outcome reporting, as well as spin.

Our response: We have revised based on the previous suggested made by the reviewer.

Revision	Page	Line
Three studies have reported on differences between COVID-19 preprint and published study reports and citations and Altmetric attention metrics (51-53). One study addressed publication characteristics and dissemination of COVID-19 preprints, one study addressed outcome reporting and spin in interpretation of results, and one study addressed risk of bias and spin. These studies were, however, restricted to only publications up to August and October 2020—which is not representative of the current landscape of COVID-19 research and which does not include the majority of evidence being currently used to guide COVID-19 care, including critical trials addressing the effects of corticosteroids and IL-6 receptor blockers (1, 2). These studies did not compare the effects of including preprints on meta-analytic estimates and the certainty of the body of evidence, which is particularly important because evidence users use the totality of the body of evidence, rather than single studies, to make treatment decisions and recommendations (51). One study has addressed publication rates and citations for COVID-19 research but it primarily addresses basic science research and so its findings may not be applicable to COVID-19 clinical trials (54).	17 to 18	437 to 449

It is accurate that a major contribution of this study (as other have assessed reporting and spin) is that

assessed the effect of preprints on analytic estimates, but, as noted above, a limitation of this analysis is that it did not compare meta-analytic estimates when preprints vs final publications were included.

Our response: While addressing the question of whether final publications or preprints of the same trial should be included in systematic reviews was outside the scope of our work, we have revised to acknowledge this important remaining question in the discussion.

Revision	Page	Line
In situations where there are discrepancies between preprints and publications, systematic reviewers and guideline developers will also need to consider whether the results of preprints or publications are more trustworthy and should be incorporated in the meta-analysis. In such situations, systematic reviewers and guideline developers may assume that changes between preprints and publications are due to errors or inaccuracies in the reporting in the preprint that were later corrected during peer review. Future research should address whether including results from preprints or final publications impacts overall findings and decisions.	20	519 to 525

Reviewer: 4

Recommendation:

Comments:
This is a very good study, that should be of interest to all evidence users, particularly stakeholders involved in data synthesis. It presents an analysis of key differences from preprints to peer-reviewed publications of COVID-19 clinical trials and changes in meta-analytic effect estimates and confidence assessments by including preprints or not. This is a substantial contribution to the decision-making process of planning and conducting meta-analyses. Below I list some major and minor points that, in my opinion, need revision.

Major:

1. I do not see how the analyses presented inform on the trustworthiness of preprints. The study presents the changes in methods and results reported between preprint and peer-reviewed publication, but only some of the items assessed can be interpreted as improving trust by improving transparency.

As this interpretation may be subjective and the authors do not wish to change the title, then I'd recommend adding clarification to the 'Purpose' section of the abstract (1).

Our response: We assumed that important changes in the reporting of key methods and results between preprints and their subsequent publications indicates potential issues with trustworthiness. For example, such situations may indicate inaccurate or incorrect reporting in the preprint that was later corrected during the peer review process. If most preprints have important differences with their later publications, this would suggest that preprints may not be trustworthy.

We also agree, however, that the construct of trustworthiness is broad and includes other aspects of trial design and reporting in addition to differences between preprints and later publications.

We have revised to clarify this issue in the abstract, introduction, and discussion sections of the manuscript. If the reviewers and editors have alternative terminology to replace 'trustworthiness', we are happy to oblige.

We have revised to explicitly acknowledge these issues.

Revision	Page	Line
Purpose: To assess the trustworthiness (complete and consistent reporting of key aspects of the methods and results between preprint and published trial reports) and impact (effects of preprints on meta-analytic estimates and the certainty of evidence) of preprint trial reports during the COVID-19 pandemic.	3	62 to 64
Preprints that are subsequently published in journals may be the most rigorous and may not represent all trial preprints—particularly those that remain unpublished.	3	83 to 85
To assess preprint trustworthiness, we compared reporting of key aspects of the methods and results between preprint and published trial reports. We assumed that important changes in the reporting of key methods and results between preprints and their subsequent publications may indicate inaccurate or incorrect reporting in the preprint that was later corrected during the peer review process and potential issues with trustworthiness. We acknowledge that differences between preprints and published reports, however, only addresses one component of the construct of trustworthiness and there are other factors, in addition to differences between preprints and published reports that may make a trial untrustworthy.	18	468 to 474
We found no compelling evidence that there are important discrepancies between preprint and published trial reports.	20	529 to 530

2. Similarly, the abstract should include the clarification that the purpose of assessing impact is related to changes in meta-analysis as the impact on clinical practice or outcomes of the pandemic are not directly and systematically assessed.

Our response: We have revised as suggested.

Revision	Page	Line
Purpose: To assess the trustworthiness (complete and consistent reporting of key aspects of the methods and results between preprint and published trial reports) and impact (effects of preprints on meta-analytic estimates and the certainty of evidence) of preprint trial reports during the COVID-19 pandemic.	3	62 to 64
Our assessment of the impact of preprints focuses only on the impact of preprints on meta-analytic estimates and the certainty of evidence and does not consider other aspects of impact, such as number of citations or Altmetrics.	19	495 to 497

3. An important limitation not mentioned is regarding the subjectivity of the GRADE framework for assessing certainty of evidence, which is one of the most relevant outcomes of the study.

Our response: We agree and have revised to acknowledge this limitation.

Revision	Page	Line
We used the GRADE approach to assess the certainty of evidence (2). While the GRADE framework provides a transparent and systematic framework of all factors that may bear on the certainty of evidence, its application is subjective.	19	498 to 500

4. On Table 1, single center and multicentre do not sum up to the total. Is there another option? There seems to be 10 studies missing in published without preprints, 14 missing in preprint only, and 4 missing in preprint first.

Our response: We were unable to classify a small number of trials as either single or multicentre. For example, some trials recruited and randomized participants through the internet. These odd trials are missing from the classification.

5. On Table 1, it is unclear from which statistical test the p values are obtained. This information is not anywhere of the results or methods sections.

Our response: We have revised to report this information.

Revision	Page	Line
We compare the characteristics and risk of bias of trials with preprints, trials with publications, and trials first posted as a preprint and subsequently published by calculating differences in proportions, associated confidence intervals, and z tests to test for differences in independent proportions. To compare the number of participants in trials with preprints, trials with publications, and trials first posted as a preprint and subsequently published, we performed Mann Whitney U tests.	8	191 to 195

6. Did you consider corrections for type 1 error rates, given the multiple comparisons performed? While these comparisons may not be the focus of the paper, I believe this is still an important consideration to be made even if corrections are not used.

Our response: We do not draw any inferences from the statistical tests and only interpret descriptive characteristics. For this reason, we did not adjust for multiple testing. If the reviewers and editors feel strongly, however, we will oblige.

7. Why are there so many missing values (reported as undefined or NA) on Table 3? While upper bound confidence intervals may not be exact and some software yield such results, some of the missing values refer to medians (even though p values are reported). This warrants a revision of all results in this table. If there really is the case for inexact values, it should be stated in the legend.

Our response: The missing values are because there is insufficient follow-up of the COVID-19 preprints to be able to calculate the upper bound of the confidence interval.

8. Is it possible to present the median and interquartile range (or other summary) of the number of changes per preprint? The results on the most frequent types of changes are very interesting, but it led me to wonder whether there is an influence of a few poorly reported preprints as the overall number of changes are very low for most of the categories.

Our response: We have revised as suggested.

Revision	Page	Line
We identified a median of 1 [IQR: 0 to 2] discrepancies per pair of preprint and publication reports.	11	286 to 287

9. In lines 283-5, there's a statement that changes in outcome data are likely due to accumulating events over time. From these 20 trials with changes, how many of them had the preprint reporting that they are presenting interim analyses and that the outcome would continue to be evaluated? I think it could help you corroborate this claim.

Our response: We agree with the reviewer's suggestion. Unfortunately, however, we are unable to provide exact numbers since we were unable to contact all trial investigators to determine whether differences in results were due to accumulating events.

10. In the text, when describing the two cases of changes in conclusions (results from Table 4), it would be helpful to also have at hand how many preprint trials are added to lead to such changes. In one case it was just one additional trial with proportionally very few patients added while in the other, there are 3 trials more with more substantial additions in patient.

Our response: We agree and have revised as suggested.

Revision	Page	Line
The meta-analysis without preprints included one trial with 5,418 participants and the meta-analysis with preprints included two trials with 5,472 participants.	14	342 to 344
The meta-analysis without preprints included seven trials with 1,826 participants and the meta-analysis with preprints included nine trials with 4,000 participants.	14	348 to 350

11. You conclude that including preprints may affect the results of meta-analyses and encourage the use of preprints. However, the changes found can be positive or negative (improving or worsening certainty of evidence) and I believe it is important to highlight this also in the conclusion section.**Our response:** We agree with the reviewer's assessment. When systematic reviewers, however, consider evidence from preprints, they have the option to perform sensitivity analyses excluding preprints and to scrutinize preprints for potential problems with data fabrication or falsification. If systematic reviewers judge that including preprints lowers the certainty of evidence, they can disregard evidence from all or some preprints. Systematic reviewers, however, cannot determine what the certainty of evidence would be with preprints if they do not include preprints in the systematic review. We have revised to clarify.

Revision	Page	Line
While we found that preprints may both increase or reduce the certainty of evidence, we encourage systematic reviewers and guideline developers to consider evidence from preprints, appraise preprint reports, and to only consider excluding preprints in situations where there are data fabrication or integrity issues.	16	408 to 411

12. Please consider openly sharing the raw data collected for the study.

Our response: The raw data is available on Open Science Framework.

Revision	Page	Line

Data: Data is available at https://osf.io/9adxb/.	2	48

Minor:

13. Not all abbreviations are defined on first use, such as IQR and SD. While it might seem obvious to those used to them, it might not be so to others, particularly those for which English is a second/foreign language.

Our response: We have revised as suggested.

Revision	Page	Line
Other differences in the reporting of key methods were the publication reporting one or more additional statistics important for meta-analysis (e.g., interquartile ranges (IQR) or standard deviations (SD)) that were not previously reported in the preprint (n=6; 8.1%), the preprint reporting on interim results and the publication on completed trial results (n=4; 5.4%), and the publication including a protocol and/or statistical analysis plan as a supplementary that was not previously included with the preprint (n=3; 4.1%).	12	294 to 298

14. In the abstract, it may be useful to mention that certainty of evidence was assessed using the GRADE framework.

Our response: We have revised as suggested.

Revision	Page	Line
For the effects of eight therapies on mortality and mechanical ventilation, we performed meta-analyses including preprints and excluding preprints at 1 month, 3 months, and 6 months after the first trial addressing the therapy became available either as a preprint or publication (120 meta-analyses in total, 60 of which included	3	69 to 73
preprints and 60 of which excluded preprints) and assessed the certainty of evidence using the GRADE framework.		

15. In many parts of the manuscript the authors state that preprints are relevant for health emergencies. I'd argue this is not the case and would like to ask the authors to consider if this is also not true for any disease, which will be an urgent issue for those affected by it. Of course, the generalizability of the results found is limited to COVID-19 but the results also open the possibility that preprints can be more used in general.

Our response: We agree and have revised to acknowledge this point.

Revision	Page	Line
We show that preprints remain the only source of findings of many trials for several months—a length of time that is unacceptable in a health emergency and is not conducive to treating patients with timely evidence.	4	87 to 89
We show that preprints remain the only source of findings of many trials for several months. Half of all preprints, for example, remain unpublished at six months and a third at one year—a length of time that may be unacceptable in a health emergency or to patients who may expect that their care is guided by the most recent and best available evidence.	15	379 to 382

16. On lines 170-174, when listing possible outcomes, viral clearance and time to viral clearance are

mentioned twice.

Our response: We have revised as suggested.

Revision	Page	Line
We focused on the same outcomes as our living SRNMA: mortality, mechanical ventilation, adverse events leading to discontinuation, admission to hospital, viral clearance, hospital length of stay, ICU length of stay, duration of mechanical ventilation, time to symptom resolution or clinical improvement, days free from mechanical ventilation, and time to viral clearance.	7	175 to 181

17. On lines 245-6, overall is written twice.

Our response: We have revised as suggested.

Revision	Page	Line
Supplement 4 presents additional details on the results of the search and Table 1 presents overall trial characteristics and trial characteristics stratified by publication status.	10	265 to 266

18. On Table 1, do you really need both the yes and no lines for 'trial registered'? As there's only one line for 'inpatient' and 'severity', I'd argue the same can be applied for trial registration. In my opinion, avoiding redundant information improves tables' readability.

Our response: We have revised to remove the redundant line.

Revision	Page	Line
Table 1.	21	536

19. On Table 1, why is there no statistical comparison in number of patients?

Our response: We have revised to include a statistical comparison using the Wilcoxon Rank Sum test.

Revision	Page	Line
Table 1.	21	536

20. Are the results from table 2 in line with other estimates of the literature? This might help discussing the generalizability of the findings. I am aware of two reports (https://doi.org/10.1371/journal.pbio.3000959 for COVID-19 preprints, and https://doi.org/10.1101/833400 for bioRxiv before the pandemic), but there may be others.

Our response: Time to publication of COVID-19 preprints is unfortunately not reflective of preprints in other fields. For example, a pre-COVID-19 estimation of the median time to publication of completed clinical trials was 21 months (Ross JS, Mocanu M, Lampropulos JF, Tse T, Krumholz HM. Time to publication among completed clinical trials. JAMA Intern Med. 2013 May 13;173(9):825-8. doi: 10.1001/jamainternmed.2013.136. PMID: 23460252; PMCID: PMC3691813.). Conversely, we found median time to publication to be closer to 6 months. During the COVID-19 pandemic, journals expedited the publication of COVID-19 research and were publishing more prolifically on COVID-19 than in other areas—which likely led to a shorter time between preprints being posted on preprint servers to eventually becoming published in peer-reviewed journals. We have revised to clarify.

Revision	Page	Line
The generalizability of our results is, however, limited to COVID-19. Journals have expedited the publication of COVID-19 research and have been publishing more prolifically on COVID-19 than in other areas, which may reduce opportunity for revisions between preprints and their subsequent publications and may mean time to and predictors of publication may be different than in other research areas. For these reasons, our estimate of the time to publication of COVID-19 preprints is different from estimates from before COVID-19 (53).	18	459 to 464

21. It is unclear to me whether the example presented in Box 1 is representative of a change in description that led to a change in risk of bias assessment. Please clarify.

Our response: The example presented in table 1 resulted in a change to the rating of risk bias due to randomization from 'probably high risk of bias' to 'definitely low risk of bias'. We have revised to clarify.

Revision	Page	Line
The PANAMO trial, which was initially available as a preprint on SSRN and later published in Lancet Rheumatology, provided additional details in the publication on	11	292
allocation concealment. The publication describes central randomization with an online tool and the development of the randomization list by a third party—all of which were not reported in the preprint (29, 30). This resulted in a change of the rating of the risk of bias due to randomization of 'probably high risk of bias' to 'definitely low risk of bias'.		

22. Table 4 is very difficult to read. Please consider whether the full version could be presented as supplementary and a summarised version could be kept in the main manuscript. For example, the column 'risk with standard care' seems mostly uninformative and this information could be easily presented in the table legend. Alternatively, you may want to consider breaking the table in multiple smaller tables.

Our response: If the size of the table cannot be accommodated, we can break up the table or present in the supplement. We will defer to the preferences of the editors and production team.

23. When discussing the risk of fabrication or falsification, you may want to consider the feasibility of these assessments for systematic reviewers. While this is an important consideration, it might be more in the realm of responsibilities of peer reviewers. If this would become more commonplace, you might expect that the results of comparability of meta-analysis including or not preprints would change over time as more preprints than peer-reviewed publications would include false data.

Our response: We agree and have revised to address feasibility in discussion.

Revision	Page	Line
We acknowledge, however, that it may not always be feasible for systematic reviewers to be responsible for applying these methods. We also encourage editors and peer reviewers to also consider applying these methods.		

24. The references need some extensive revisions. For example, references 9 and 49 have a version of record that can be preferentially cited. I also found that references 24, 26, 30 and 35 are incomplete.

Our response: We have revised the problematic references.

Reviewer: 5

Recommendation:

Comments:
Thank you for the opportunity to review this interesting manuscript. In this project, Zeraatkar et al. conduct a large-scale evaluation of COVID-19 trials that were disseminated at preprints, preprints with corresponding publications, and only publications. In particular, the authors set out to determine the trustworthiness of impact of preprint trial reports during the COVID-19 pandemic. This is an interesting and important question, as the trade-offs between the benefits (rapid dissemination) and potential harms (lack of peer-review and editorial oversight) of preprints in clinical sciences have been debated extensively. This current project is a major effort that builds upon the herculean living review published in the BMJ. The authors analyze a number of characteristics – publication timing, the frequency of retractions, the differences between preprints and publications, and the impact of including or excluding preprints in meta-analyses. Overall, the authors conclude that they found no evidence that preprints provide less trustworthy results than published papers.

Although this project contains a significant number of interesting findings, at times, the large number of outcomes makes it somewhat challenging to identify the focus. In particular, the conclusions and inferences in the abstract and key points sections primarily seem to relate to the analysis focused on how summary effect estimates and certain of evidence may change after including preprints (I find this to be the most interesting component of the study). Other evaluations have focused on the similarities between COVID-19 preprints and their subsequent journal publications (e.g., https://bmjopen.bmj.com/content/11/7/e051821 and https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.3001285; I was not involved in either of these previous evaluations).

Our response: We thank the reviewer for their positive assessment of our work. We have revised the abstract and introduction of our manuscript to clarify the objectives of this study.

Revision	Page	Line
Purpose: To assess the trustworthiness (complete and consistent reporting of key aspects of the methods and results between preprint and published trial reports) and impact (effects of preprints on meta-analytic estimates and the certainty of evidence) of preprint trial reports during the COVID-19 pandemic.	3	62 to 64
Knowledge of the extent to which preprints may accelerate the dissemination of findings, the frequency and nature of discrepancies between pre-prints and subsequent published reports, and the impact preprints on meta-analytic estimates could inform the trade-off that evidence users face. Our study capitalizes on our living systematic reviews and network meta-analyses (SRNMAs) of drug treatments, antiviral antibodies and cellular therapies, and prophylaxis for COVID-19—an initiative launched in July 2020 that provides real-time summaries addressing the comparative effectiveness of treatments and prophylaxis for COVID-19—to report on the characteristics, trustworthiness—that is, complete and consistent reporting of key aspects of the methods and results between preprint and published trial reports—and impact of COVID-19 trial preprint reports (18-20).	5	110 to 117

One of the key components of the paper is the retraction of preprints and publications. Although reassuring, there are only 4 retracted trials - 3 of which were published in peer-reviewed journals - in this sample. Unfortunately, this sample is not large enough to draw any meaningful inferences.
However, the authors conclude that they "…found retractions to occur for both preprints and publications, suggesting that publication in a peer-reviewed journal alone does not indicate trustworthiness of a trial". While I completely agree that this is likely the case, I am not convince that

the findings from this particular analysis provide suggestive evidence.

Our response: We have revised to acknowledge that there were too few retractions to be able to draw confident inferences.

Revision	Page	Line
We report on the number of publications and preprints that were retracted. Preprints, however, may be less likely to be retracted because they may draw less attention and because preprint servers may be less likely than journals to have formal policies addressing research integrity. Further, there were too few retractions to be able to draw confident conclusions.	19	491 to 494

A second focus of the paper is the concordance between preprints and their corresponding publications. However, these findings are mainly discussed in the main text/tables and are not included in the Abstract or Key findings. The authors note that 57% of trials had discrepancies in the reporting of key methods and results between the preprint and the later published trial report, with details in Supplement 5. However, it is hard to determine the importance of these changes.

Our response: We have revised to describe these findings in the abstract and key findings. Please note that it was difficult to provide a comprehensive overview of the results of the study while adhering to the word limit.

We communicate the importance of differences in reporting of key methods by reporting whether they resulted a change in the ratings of risk of bias. Thirty trials had one or more discrepancies in the rating of key methods. The overall trial rating of risk of bias, however, changed only for one trial based on additional information provided in the published report—suggesting that the discrepancies in the reporting of key methods were mostly inconsequential in making judgements about the validity of the trials.

We communicate the importance of differences in reporting of key results by reporting on the magnitude of differences and precisions in figure 1.

Revision	Page	Line
There were few important differences in key methods and results between trial preprints and their subsequent published reports.	3	76 to 77
We did not find compelling evidence of important differences between preprint and published reports of trials—though preprint reports of trials that are subsequently published in journals may not be representative of all trial preprints.	15	384 to 386
The overall trial rating of risk of bias, however, changed only for one trial based on additional information provided in the published report.	12	299 to 300
Despite discrepancies in outcome data being common, results were similar between preprints and publications both in magnitude and precision. Figure 1 shows differences in results on mortality and mechanical ventilation between preprints and publications. Among all preprints with differences in outcomes, differences in relative effects did not exceed 15%, except for one trial with very few events that included just one additional event in the publication (32, 33). Other differences between preprints and publications in key results included the publication reporting at least one additional key outcome that was not included in the preprint (n=11; 14.9%).	13	306 to 312

Overall, this manuscript reflects a significant effort and contains several interesting findings, especially related to the meta-analyses. However, it may be challenging to comment on the overall trustworthiness of preprints without knowing more about the preprints that were submitted but reject by journals. Perhaps

the authors could consider making the focus more on how evidence changes, based on numerous factors, vs. the issue of trustworthiness?

Our response: We assumed that important changes in the reporting of key methods and results between preprints and their subsequent publications indicates potential issues with trustworthiness. For example, such situations may indicate inaccurate or incorrect reporting in the preprint that was later corrected during the peer review process. If most preprints have important differences with their later publications, this would suggest that preprints may not be trustworthy.

We also agree, however, that the construct of trustworthiness is broad and includes other aspects of trial design and reporting in addition to differences between preprints and later publications.

We have revised to clarify this issue in the abstract, introduction, and discussion sections of the manuscript. If the reviewers and editors have alternative terminology to replace 'trustworthiness', we are happy to oblige.

We have revised to explicitly acknowledge these issues.

Revision	Page	Line
Purpose: To assess the trustworthiness (complete and consistent reporting of key aspects of the methods and results between preprint and published trial reports) and impact (effects of preprints on meta-analytic estimates and the certainty of evidence) of preprint trial reports during the COVID-19 pandemic.	3	62 to 64
Preprints that are subsequently published in journals may be the most rigorous and may not represent all trial preprints—particularly those that remain unpublished.	3	83 to 85
To assess preprint trustworthiness, we compared reporting of key aspects of the methods and results between preprint and published trial reports. We assumed that important changes in the reporting of key methods and results between preprints and their subsequent publications may indicate inaccurate or incorrect reporting in the preprint that was later corrected during the peer review process and potential issues with trustworthiness. We acknowledge that differences	18	468 to 474
between preprints and published reports, however, only addresses one component of the construct of trustworthiness and there are other factors, in addition to differences between preprints and published reports that may make a trial untrustworthy.		
We found no compelling evidence that there are important discrepancies between preprint and published trial reports.	20	529 to 530

Please find some additional comments below:

Methods:

Line 121: it is great to see that the authors shared their full study protocol.

Our response: We thank the reviewer for the encouraging comment.

Line 124: When were the searches updated?

Our response: We have revised to clarify.

Revision	Page	Line

This study includes trials identified up to August 3rd, 2021.	6	138

Line 194: The authors note that they “describe the number and types of discrepancies in key methods and results between preprint and published trial reports". Could the authors clarify the specific components that were compared in the methods section of the text?

Our response: We have revised to clarify.

Revision	Page	Line
Key methods included description of the randomization process and allocation concealment, blinding of patients and healthcare providers, extent of and handling of missing outcome data, blinding of outcome assessors and adjudicators, and prespecification of outcomes and analyses.	7	169 to 173

Line 201: How were retractions identified?

Our response: We used our search strategy to monitor for retraction notices. Courtesy of our work on the parallel living SRNMAs that informed linked living WHO guidelines, we capitalized on our network of collaborators that included COVID-19 trialists and guideline developers who communicated retractions or concerns about data integrity with our team.

We have revised to clarify.

Revision	Page	Line
Using the above sources, we monitor for retraction notices. We also capitalize on our team that includes COVID-19 clinical trialists and guideline developers to flag trials with potential data integrity issues.	6	136

Line 203: Why did the authors limit their sample to trials that report on interventions up to August 3rd 2021? It seems like this could have missed potentially eligible preprints and publications? Could the authors also clarify why the included vs. excluded evidence from preprints at one, three, and six months after the first trial addressing the drug of interest was made public?

Our response: The living COVID-19 NMA performs daily searches for relevant COVID-19 literature. The preprint study, however, required the extraction of additional data from trials that are not routinely collected for the living COVID-19 NMA. For feasibility, we were only able to perform these extractions for trials that we had identified up to August 3rd, 2021. We have revised to clarify.

The COVID-19 pandemic instigated a surge of research activity and pressures to make decisions very quickly and within timeframes that were not compatible with the traditional timeframes for peer review and publication. Our choice of timepoints were informed by the experiences of the linked living guidelines and the timeframes within which the panel made recommendations. We agree, however, that these timepoints were arbitrary. We have revised to clarify.

Revision	Page	Line
While the parallel living SRNMA performs ongoing daily searches, we pragmatically limited our search because it was no longer feasible to continue to collect additional data from preprints beyond this timepoint.	7	156 to 158
The choice of timepoints was informed by timeframes within which guideline developers needed to issue recommendations (23).	9	219 to 221

Line 217: It appears as if the authors focused on the direction of effect between meta-analyses including

vs. excluding preprints. Would it be interesting to consider statistical significance and direction?

Our response: We have revised to report the proportion of meta-analyses including versus excluding preprints that produced results with differences in direction and statistical significance.

Revision	Page	Line
Except for two cases, across all meta-analyses including and excluding results from unpublished preprints, the point estimates were consistent as to whether they indicated benefit, no appreciable effect, or harm.	14	337 to 339
Four of sixty meta-analyses had results that were statistically significant with preprints and results that were not statistically significant without preprints, or vice versa.	14	351 to 352

Line 219: How were the 1%, 2%, and 0.5% values for benefit/harm selected?

Our response: These thresholds come from our parallel living systematic reviews of drug for COVID-19 (Siemieniuk RA, Bartoszko JJ, Ge L, Zeraatkar D, Izcovich A, Kum E, Pardo-Hernandez H, Qasim A, Martinez JPD, Rochwerg B, Lamontagne F, Han MA, Liu Q, Agarwal A, Agoritsas T, Chu DK, Couban R, Cusano E, Darzi A, Devji T, Fang B, Fang C, Flottorp SA, Foroutan F, Ghadimi M, Heels-Ansdell D, Honarmand K, Hou L, Hou X, Ibrahim Q, Khamis A, Lam B, Loeb M, Marcucci M, McLeod SL, Motaghi S, Murthy S, Mustafa RA, Neary JD, Rada G, Riaz IB, Sadeghirad B, Sekercioglu N, Sheng L, Sreekanta A, Switzer C, Tendal B, Thabane L, Tomlinson G, Turner T, Vandvik PO, Vernooij RW, Viteri-García A, Wang Y, Yao L, Ye Z, Guyatt GH, Brignardello-Petersen R. Drug treatments for covid-19: living systematic review and network meta-analysis. BMJ. 2020 Jul 30;370:m2980. doi: 10.1136/bmj.m2980. Update in: BMJ. 2020 Sep 11;370:m3536. Update in: BMJ. 2020 Dec 17;371:m4852. Update in: BMJ. 2021 Mar 31;372:n858. Erratum in: BMJ. 2021 Apr 13;373:n967. PMID: 32732190; PMCID: PMC7390912.). The thresholds were informed by a survey of the authors. We have revised to clarify.

Revision	Page	Line
Our thresholds for beneficial and harmful effects were informed by surveys of the co-authors in the parallel living SRNMAs (18-20).	9	239 to 240

Results:

Line 242: Perhaps clarify that it is August 3rd, 2021?

Our response: The living COVID-19 NMA performs daily searches for relevant COVID-19 literature. The preprint study, however, required the extraction of additional data from trials that are not routinely collected for the living COVID-19 NMA. For feasibility, we were only able to perform these extractions for trials that we had identified up to August 3rd, 2021. We have revised to clarify.

Revision	Page	Line
While the parallel living SRNMA performs ongoing daily searches, we pragmatically limited our search because it was no longer feasible to continue to collect additional data from preprints beyond this timepoint.	7	156 to 158

Line 247: Is there a table with the findings reported in lines 247-250?

Our response: We have revised as suggested.

Revision	Page	Line
As of August 3rd 2021, we identified 356 eligible trials, 101 of which were only available as preprints, 181 only available as journal publications, and 74 first available as preprints and subsequently published as journal articles.	10	263 to 265

Line 251: How did the authors determine the interim vs. non-interim results, as I don't remember seeing this in the methods section.

Our response: We contacted trial investigators. We could not, however, contact all trial investigators.

Line 281: What are 'key results'? Are these primary endpoints?

Our response: We have revised to clarify.

Revision	Page	Line
Key results included number of participants analyzed and number of events in each trial arm for dichotomous outcomes and number of participants analyzed, means or medians and measures of variability for continuous outcomes. We focused on the same outcomes as our living SRNMA: mortality, mechanical ventilation, adverse events leading to discontinuation, admission to hospital, viral clearance, hospital length of stay, ICU length of stay, duration of mechanical ventilation, time to symptom resolution or clinical improvement, days free from mechanical ventilation, and time to viral clearance.	7	173 to 181

Line 319: The findings in this paragraph can be a little challenging to follow (e.g., the X.X more and X.X fewer). Would it make sense to report the estimates with and without preprints, instead of the differences between the two. This may improve clarity.

Our response: We compare the results of meta-analyses including versus excluding preprints. To improve interpretability, we report the results of the meta-analyses as absolute rather than relative effects (i.e., the number of events per 1,000 patients rather than relative risks). This is consistent with GRADE guidance and guidance from the Cochrane collaboration for presenting result of systematic reviews (Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, Norris S, Falck-Ytter Y, Glasziou P, DeBeer H, Jaeschke R, Rind D, Meerpohl J, Dahm P, Schünemann HJ. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. J Clin Epidemiol. 2011 Apr;64(4):383-94. doi: 10.1016/j.jclinepi.2010.04.026. Epub 2010 Dec 31. PMID: 21195583.). The absolute numbers represent the risk difference with and without the intervention rather than differences between meta-analyses including versus excluding preprints.

The advantage of presenting results in absolute numbers is to allow judgements about whether the magnitude of differences between the results of meta-analyses including versus excluding preprints is important or meets the minimally important difference for that outcome.

In addition to absolute effects, we present the relative effects in Table 4.

Line 332: 9 cases out of how many total?

Our response: We have revised to clarify the denominator for the results in this section.

Revision	Page	Line
There were nine cases of 60 meta-analyses for which the rating of the certainty of evidence was different when preprints were included versus excluded.	3	79 to 80
Between meta-analyses including versus excluding preprints, judgements related to the GRADE imprecision domain differed for 13 of 60 meta-analyses.	15	370 to 371

Discussion:

- Please see the comment above about the focus on 'trustworthiness'.

Our response: We assumed that important changes in the reporting of key methods and results between preprints and their subsequent publications indicates potential issues with trustworthiness. For example, such situations may indicate inaccurate or incorrect reporting in the preprint that was later corrected during the peer review process. If most preprints have important differences with their later publications, this would suggest that preprints may not be trustworthy.

We also agree, however, that the construct of trustworthiness is broad and includes other aspects of trial design and reporting in addition to differences between preprints and later publications.

We have revised to clarify this issue in the abstract, introduction, and discussion sections of the manuscript. If the reviewers and editors have alternative terminology to replace 'trustworthiness', we are happy to oblige.

We have revised to explicitly acknowledge these issues.

Revision	Page	Line
Purpose: To assess the trustworthiness (complete and consistent reporting of key aspects of the methods and results between preprint and published trial reports) and impact (effects of preprints on meta-analytic estimates and the certainty of evidence) of preprint trial reports during the COVID-19 pandemic.	3	62 to 64
Preprints that are subsequently published in journals may be the most rigorous and may not represent all trial preprints—particularly those that remain unpublished.	3	83 to 85
To assess preprint trustworthiness, we compared reporting of key aspects of the methods and results between preprint and published trial reports. We assumed that important changes in the reporting of key methods and results between preprints and their subsequent publications may indicate inaccurate or incorrect reporting in the preprint that was later corrected during the peer review process and potential issues with trustworthiness. We acknowledge that differences between preprints and published reports, however, only addresses one component of the construct of trustworthiness and there are other factors, in addition to differences between preprints and published reports that may make a trial untrustworthy.	18	468 to 474
We found no compelling evidence that there are important discrepancies between preprint and published trial reports.	20	529 to 530

- The authors could also consider comparing their findings to a previous evaluation focus on the concordance between preprints and their corresponding publications in the highest impact factor journals: https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2777629. Disclosure: I was an author on that manuscript. I defer to the authors whether they think including this is helpful/necessary.

Our response: We have revised to include a discussion of this paper.

Revision	Page	Line

Our results are also aligned with previous assessments before COVID-19 that study interpretations and other study details do not change importantly between preprints and their later publications in high impact journals (53).	18	454 to 465

Figures: These figures are terrific.

Our response: We thank the reviewer for their positive feedback.

VERSION 2 – REVIEW

REVIEWER 1	**Kirkham, Jamie;** The University of Manchester, Biostatistics. Competing Interest: None
REVIEW RETURNED	25-Jul-2022

GENERAL COMMENTS	As per my original review during the BMJ manuscript meeting I do feel as though the impact of the findings from this study are overstated in places. Most comments have now been addressed but here are my reflections on the revision. It's a study well worthy of publication and adds greater depth to what has been done on this topic already. 1) I was quite surprised that so many reviewers suggested this work is not about trustworthiness in yet the authors chose not to make a change. I also agree that this is not about trustworthiness despite the authors attempt to define this in the revision. Trustworthiness is when you suspect foul play – the authors discuss some retracted trials due to concerns about data fabrication or falsification – this is about 'trustworthiness' and affects a minority in this study. Most of this article is about discrepancies. Can the authors please change this wording. Use 'discrepancies' if an alternative suggestion is needed. My rationale is that I thought all the discrepancies found we very minor in most cases - this seems to stem from 2 possible reasons; 1) in the published article, possible changes could have been requested under peer review; 2) there seems to be some suggestion that some researchers may preprint before final results are confirmed which may result in small changes in numbers presented. Larger changes are perhaps only observed when the results are more sensitive to change, e.g. the event is rare. This is potentially impactful and deserves notable caution (and for this reason the study has merit), but the authors did not find any serious changes in results between preprint and published article - this had been confirmed in other similar studies which are cited. 2) Table 3 – where the upper confidence interval is 'NA' – this needs defining as a footnote what this means as does 'undefined'. 3) Box 1 – I disagree that the preprint in this scenario should be

	judged as high risk and the full publications as low risk – the risk stays the same in my view – it's a reporting deficiency which is different to RoB. I think this phrasing needs to be changed throughout.

VERSION 2 – AUTHOR RESPONSE

Comments to the Author

As per my original review during the BMJ manuscript meeting I do feel as though the impact of the findings from this study are overstated in places. Most comments have now been addressed but here are my reflections on the revision. It's a study well worthy of publication and adds greater depth to what has been done on this topic already.

Our response: We thank the review for their positive evaluation of our work.

1) I was quite surprised that so many reviewers suggested this work is not about trustworthiness in yet the authors chose not to make a change. I also agree that this is not about trustworthiness despite the authors attempt to define this in the revision. Trustworthiness is when you suspect foul play – the authors discuss some retracted trials due to concerns about data fabrication or falsification – this is about 'trustworthiness' and affects a minority in this study.

Most of this article is about discrepancies. Can the authors please change this wording. Use 'discrepancies' if an alternative suggestion is needed.

My rationale is that I thought all the discrepancies found we very minor in most cases - this seems to stem from 2 possible reasons; 1) in the published article, possible changes could have been requested under peer review; 2) there seems to be some suggestion that some researchers may preprint before final results are confirmed which may result in small changes in numbers presented. Larger changes are perhaps only observed when the results are more sensitive to change, e.g. the event is rare.

Our response: We have revised as suggested. We have replaced most references to 'trustworthiness' with references to 'consistency' and 'discrepancies'. We retained some references to trustworthiness in the discussion in the context of falsified or fabricated trials as suggested by the editors.

Revision	Page	Line
Title: COVID-19 Preprints: Consistency with later publications and impact for decision-making	1	1-2
We found no compelling evidence that preprints provide results that are inconsistent with published papers.	4	104-105
We use these reviews to assess the degree of discrepancies between COVID-19 trial preprints and their later publications and the effects of considering evidence from preprints on meta-analytic estimates, the certainty (quality) of evidence, and decision-making.	6	138-140
Our study presents a detailed assessment of the degree of discrepancies between COVID-19 trial preprints and their later publications and the impact of trial preprints meta-analytic estimates, the certainty of evidence, and decision-making.	22	407-409

Our findings have implications for evidence users, such as clinicians, who are concerned with the quality of preprints and for systematic reviewers and guideline developers deciding whether to consider preprint reports in systematic reviews and guideline recommendations.	23	435-437

This is potentially impactful and deserves notable caution (and for this reason the study has merit), but the authors did not find any serious changes in results between preprint and published article - this had been confirmed in other similar studies which are cited.

Our response: We have revised as suggested.

2) Table 3 – where the upper confidence interval is 'NA' – this needs defining as a footnote what this means as does 'undefined'.

Our response: We have revised as suggested. See footnote of Table 3. "The upper bounds of confidence intervals could often not be estimated due to insufficient follow-up of preprints."

3) Box 1 – I disagree that the preprint in this scenario should be judged as high risk and the full publications as low risk – the risk stays the same in my view – it's a reporting deficiency which is different to RoB. I think this phrasing needs to be changed throughout.

Our response: We used the risk of bias criteria and assessments from our parallel reviews addressing treatments and prophylaxis for COVID-19. In these reviews, only studies that are certainly at low risk of bias (that is with complete and clear reporting allowing reviewers to judge the trial to be at low risk of bias) are categorized as being low risk of bias. Studies with unclear reporting are categorized at high risk of bias. Our rationale for this approach was that risk of bias assessments are meant to identify studies that are at "risk" of bias. Studies with unclear reporting are at "risk" of bias even though they may be at low or high risk of bias—the lack of clear reporting, however, does not allow us to definitely say that a study is trustworthy and thus at low risk of bias.

This approach is also conservative for our study because even preprints with unclear reporting are classified as having discrepancies between preprints and publications. Hence, we are likely overestimating the extent to which risk of bias assessments may change between preprints and later publications.