PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Leflunomide Treatment for Patients Hospitalised with COVID-19: DEFEAT-COVID Randomised Controlled Trial
AUTHORS	Kralj-Hans, Ines; Li, Kuo; Wesek, Adrian; Lamorgese, Alexia; Omar, Fatima; Ranasinghe, Kapila; McGee, Megan; Brack, Kieran; Li, Shiliang; Aggarwal, Ritesh; Bulle, Ajay; Kodre, Aparna; Sharma, Shashank; Fluck, David; John, Isaac; Sharma, Pankaj; Belsey, Jonathan; Li, Ling; Seshasai, Sreenivasa Rao Kondapally; Li, HongLin; Marczin, N; Chen, Zhong

VERSION 1 – REVIEW

REVIEWER	Cure, Erkan Recep Tayyip Erdoğan School of Medicine, Department of Internal medicine
REVIEW RETURNED	20-Sep-2022

GENERAL COMMENTS	The authors found that leflunomide had no positive effect in treating COVID-19. They also revealed that leflunomide could be used safely in treating rheumatoid arthritis patients with COVID-19. I think that the article will contribute to the literature.

REVIEWER	Depuydt , Pieter Ghent University Hospital, Department of Intensive Care Medicine
REVIEW RETURNED	13-Dec-2022

GENERAL COMMENTS	The manuscript submitted by Dr. Kralj-Hans and colleagues presnets the results of a multicenter, open-label randomized controlled trial in patients with moderate to severe COVID-19, in wich patients are randomized between lefunomide added to standard of care (SOC) versus SOC alone; SOC including corticosteroids, antivirals, anticoagulation and antibiotics. Patients were stratified in 4 groups according to risk for mortality due to comorbidities and severity of COVID-19. Primary endpoint is time to a two point reduction on a 7-step scale of clinical status. The study observed a limited reduction in the time to clinical improvement in the leflunomide-treated group of 1 day, bordering on statistical significance. The authors assign this limited effect to the evolving insights and improvement of SOC while the patients were recruited, diluting the possible effect of leflunomide. The study is well designed and appears to have been appropriately conducted; the major endpoint is clinically meaningful, although some subjectivity may have been introduced based on the open nature of the leflunomide administration. Statistical analysis is lege artis. Interestingly, the authors also include other endpoints such as longer term (post-COVID)

	symptoms and signs, and viral replication. The discussion is interesting and balanced. I have only one trivial remark and one question: p4. Symptoms include pneumonia, a systemic inflammatory response and cardiovascular complications... --> these are not symptoms and another wording should be chosen e.g. (associated clinical) syndromes. Q: Was the effect size different across the four strata with increasing risk for mortality?

REVIEWER	Salton, Francesco
REVIEW RETURNED	13-Dec-2022

GENERAL COMMENTS	Thank you for the opportunity to review this paper. The Authors conducted an open-label RCT to evaluate the efficacy of leflunomide vs standard care on the time to clinical improvement defined as a two-point reduction on a clinical performance scale, finding a statistically significant difference only when excluding patients who were randomized despite not fulfilling inclusion criteria and dropouts. The Authors also evaluated secondary endpoints e.g. all-cause mortality, duration of oxygen dependence as well as the safety profile and the incidence of some post-COVID conditions finding no differences between the treatment arm and standard care. I think the study is quite well designed, scientifically sound and sufficiently novel, as there are only few and smaller other RCTs on the use of leflunomide in COVID, with less relevant outcomes. However, there are some major issues that need to be addressed before the paper is considered potentially suitable for publication: 1. The Authors performed ITT analyses also including 10 patients not fulfilling the inclusion criteria (wrong randomizations) and 3 who withdrew consent before receiving leflunomide treatment (dropouts). However, I think these patients would have been better excluded also from the ITT analyses and reported in the study flow-chart. Given the “modified ITT” analysis showed a statistically significant difference in the primary endpoint between groups, the results and discussion section should be changed accordingly. 2. A “per protocol” sensitivity analysis would be useful instead, including only patients who have completed the assigned treatment, to evaluate the real efficacy of the study protocol. 3. One serious possible bias of this study is the insufficient information provided about the dose of concomitant medications, especially glucocorticoids. In fact, while dexamethasone 6 mg or equivalent for 10 days has become the most used protocol, it is not the only one and different doses and treatment duration may impact the outcome (please see DOI 10.1183/13993003.01514-2022 and related references). Therefore, at least the median glucocorticoid dose and duration (or cumulative prednisone dose) should be indicated for each group. 4. Similarly, this study has been conducted in very different clinical settings. It should be discussed and reported as one major limitation the fact that the availability of noninvasive and invasive ventilation, as well as the time elapsed from clinical worsening to intubation or other internal protocols (e.g. pharmacological therapies, noninvasive ventilation, high-flow nasal cannula etc.) may have diverged and implied different outcomes. Please also consider and discuss that the pressure on single hospitals may have been different between Centers during either the same or subsequent pandemic surges, potentially hindering the ability of

	the Center to cope with all the requests for an intensification of care. This might be the reason why the number of patients who underwent NIV was lower among those recruited in India. Indeed, it is very unlikely that the real need for ventilation was that different between Countries. 5. The introduction and methods section should be expanded with some more details about the rationale (introduction) and application (methods) of the standard of care, with special regards to the therapies which have demonstrated a higher efficacy e.g. glucocorticoids, mechanical ventilation etc. 6. Please report the registration number of the study in the methods section. 7. Did the Authors exclude SARS-CoV-2-positive affected by respiratory failure or involvement due to other causes (e.g. congestive heart failure etc.) and patients with conditions (e.g. neurological ones) that could interfere with the success of noninvasive ventilation? If not, this should be discussed among the study limitations; otherwise, it should be reported among the exclusion criteria. 8. The time from hospitalization to study enrollment should be reported. 9. Is the baseline PaO2/FiO2 ratio available? 10. Non-invasive ventilation was required for 14.4% of patients in SOC+L group vs. 16.4% in the SOC group. Do this ratio include patients who were already on NIV at baseline? In any case, the number of patients on NIV at baseline should be reported in baseline data, whereas the number of patients who encountered the need for noninvasive or invasive ventilation (separately) due to clinical worsening should be reported separately in the text. Please also add the p-values where appropriate. 11. It would be useful to perform a stratified analysis for clinical severity at baseline (e.g. no oxygen, oxygen, NIV) with regard to the primary endpoint. 12. The timepoints at which the SpO2/FiO2 ratio has been evaluated should be reported in the methods. 13. The Authors state that the sample size calculation was based on the proportion of patients expected to meet the outcome criteria by 28 days. This is not convincing. Please either report previous literature data used to estimate this proportion or discuss this as a major limitation of the study design. 14. The Authors stated that the primary analysis was stratified by “baseline risk indicators”. However, what is meant for baseline risk indicators is not clear and it is worth further elucidation in the methods section. 15. “During the data cleaning process, 10 patients were flagged as not meeting the inclusion criteria (6 in SOC+L; 4 in SOC), as they did not have moderate COVID-19 symptoms at the time of randomisation.” Having moderate COVID symptoms seems not a prespecified inclusion criterion to me. Please reformulate this statement. 16. Table 1, chronic neurological disorders seem significantly different between groups to me. Please remove the p-value column from this table and discuss eventual differences at baseline in the results section. 17. Table 2, please add percentages and p-values. Furthermore, I don’t understand the first line Adverse events (n) / Patients (n). Please explain. In any case, total AE should be expressed as n. of patients with at least one AE/total n. of patients. 18. In the adverse events section it is reported that 15 vs 9 patients died. Is it meant due to AE? Why is it reported in this

	section? If so, given that the difference seems macroscopically significant, is should be discussed. 19. Please report the unit of measure of data reported in the paper e.g. those regarding viral load and CRP

REVIEWER	McGale, Paul University of Oxford, NDPH
REVIEW RETURNED	16-Jan-2023

GENERAL COMMENTS	This is a phase 3 open label randomised controlled trial, assessing the efficacy of adding leflunomide to standard care of patients hospitalised with COVID-19. Inclusion and exclusion criteria and trial procedures are well described. It would help to add a sentence justifying why the randomised treatment is unblinded. The statistical methods are appropriate, however, the proposed use of the logrank test does not seem to be reported in the main text results. The authors say the CIs of the hazard ratios were used for the significance of the treatment effect, were these different from the logrank p-values? Figures 2, 3, & 4 are rather uninformative without the addition of the results of the statistical tests used to compare the outcomes by randomised treatment. These should be added to the figures.

VERSION 1 – AUTHOR RESPONSE

Comment from Reviewer 1
Dr. Erkan Cure, Recep Tayyip Erdoğan School of Medicine, Department of Internal medicine
Comments to the Author:
Comment 1: *The authors found that leflunomide had no positive effect in treating COVID-19. They also revealed that leflunomide could be used safely in treating rheumatoid arthritis patients with COVID-19. I think that the article will contribute to the literature.*

Response: We thank the reviewer for the positive comment.

Comments from Reviewer: 2
Prof. Pieter Depuydt , Ghent University Hospital
Comments to the Author:
General Comments: *The manuscript submitted by Dr. Kralj-Hans and colleagues presents the results of a multicenter, open-label randomized controlled trial in patients with moderate to severe COVID-19, in which patients are randomized between leflunomide added to standard of care (SOC) versus SOC alone; SOC including corticosteroids, antivirals, anticoagulation and antibiotics. Patients were stratified in 4 groups according to risk for mortality due to comorbidities and severity of COVID-19. Primary endpoint is time to a two-point reduction on a 7-step scale of clinical status. The study observed a limited reduction in the time to clinical improvement in the leflunomide-treated group of 1 day, bordering on statistical significance. The authors assign this limited effect to the evolving insights and improvement of SOC while the patients were recruited, diluting the possible effect of leflunomide. The study is well designed and appears to have been appropriately conducted; the major endpoint is clinically meaningful, although some subjectivity may have been introduced based on the open nature*

of the leflunomide administration. Statistical analysis is lege artis. Interestingly, the authors also include other endpoints such as longer term (post-COVID) symptoms and signs, and viral replication. The discussion is interesting and balanced.

Response: We thank the reviewer for the encouraging comments and recognition of the merits of the study. At the study design stage, the COVID-19 pandemic has reached a critical turning point following various measures to curb the spread of the disease. Thus, a simplified study protocol facilitated a rapid roll out of the study in multiple centres. An open label design avoided the need for patients in the control arm to be administered a placebo which would have added a level of complexity to the pharmacy and direct care delivery staff and resources. The needs of the trial had to be balanced with clinical care so as not to increase the burden on the overstretched resources. An open-label design could potentially introduce bias; however, it would also allow early detection of significant adverse events and a potential outcome benefit. This was an important consideration when testing an off-label use of a medication in the disease associated with high morbidity and mortality.

We recognise the concerns with open label study design and have highlighted this in the Study Limitation section:

"In order to balance the needs of the trial with clinical care and to minimise disruption to already overstretched clinical resources during COVID-19 pandemic, we chose to adopt an open label study design. This design may have affected the data collection and clinical management of the patients and potentially introduced a bias. However, it also allowed early detection of significant adverse events and a potential outcome benefit. This was an important consideration when testing an off-label use of a medication in COVID-19, a disease with high morbidity and mortality."

Comment 1: *Symptoms include pneumonia, a systemic inflammatory response and cardiovascular complications... --> these are not symptoms and another wording should be chosen e.g. (associated clinical) syndromes.*

Response: We have corrected the word from "Symptoms…" to "Associated clinical syndromes…" in the first paragraph of the Introduction section:

"Associated clinical syndromes include pneumonia, systemic inflammatory response and cardiovascular….."

Comment 2: *Was the effect size different across the four strata with increasing risk for mortality?*

Response: In this study there were four pre-specified risk groups that allow statistically valid subgroup analyses to be carried out. The highest risk group (Group 1) included patients with a high NEWS2 score (which captures baseline clinical severity) and the presence of comorbidities considered to be indicators of high risk for an adverse outcome. The lowest risk group (Group 4) included patients with low NEWS2 scores and no significant comorbidities. Groups 2 and 3 were intermediate. The table below documents the proportion of patients achieving the primary outcome in each of these four stratification groups. As expected, there is an improvement in outcome observed as one progresses from high risk to low risk groups. However, there is no numerical or statistical difference in treatment effect within each risk group, nor is there evidence of a between-groups difference in treatment effect.

Stratification group	Patients (n=214)		Patients achieving primary outcome (n)		Patients achieving primary outcome (%)			p*
	SOC+L	SOC	SOC+L	SOC	SOC+L	SOC	Total	
Group 1 (n=29)	13	16	9	11	69.2	68.8	69.0	0.670
Group 2 (n=33)	15	18	11	13	73.3	72.2	72.7	0.448
Group 3 (n=101)	50	51	46	49	92.0	96.1	94.1	0.683
Group 4 (n=51)	26	25	25	25	96.2	100	98.0	0.261

Comments from Reviewer: 3
Francesco Salton
General comments: *Thank you for the opportunity to review this paper. The Authors conducted an open-label RCT to evaluate the efficacy of leflunomide vs standard care on the time to clinical improvement defined as a two-point reduction on a clinical performance scale, finding a statistically significant difference only when excluding patients who were randomized despite not fulfilling inclusion criteria and dropouts. The Authors also evaluated secondary endpoints e.g. all-cause mortality, duration of oxygen dependence as well as the safety profile and the incidence of some post-COVID conditions finding no differences between the treatment arm and standard care.*
I think the study is quite well designed, scientifically sound and sufficiently novel, as there are only few and smaller other RCTs on the use of leflunomide in COVID, with less relevant outcomes.

Response: We thank the reviewer for the positive feedback.

Comment 1: *The Authors performed ITT analyses also including 10 patients not fulfilling the inclusion criteria (wrong randomizations) and 3 who withdrew consent before receiving leflunomide treatment (dropouts). However, I think these patients would have been better excluded also from the ITT analyses and reported in the study flow-chart. Given the "modified ITT" analysis showed a statistically significant difference in the primary endpoint between groups, the results and discussion section should be changed accordingly.*

Response: We share the sentiment conveyed by the reviewer but respectfully disagree with the comment made. Excluding those patients would have gone against the principle of "intention-to-treat" analysis to which we committed from the start. The ITT analysis provides an assessment of all the patients taking part in a trial, based on the group they were initially and randomly allocated to. This is regardless of whether they dropped out, fully adhered to the treatment or switched to an alternative. We recognise that ITT analysis could underestimate the observed efficacy of the treatment.

Comment 2. *A "per protocol" sensitivity analysis would be useful instead, including only patients who have completed the assigned treatment, to evaluate the real efficacy of the study protocol.*

Response: Following on from the response to the comment above, we indeed felt the importance of performing additional sensitivity analysis which we referred to as a "modified ITT analysis", described accordingly in the Results section under Primary outcomes and Secondary Outcomes. Both our trial statistician and the Data Monitoring Committee members supported this approach.

In accordance with the Reviewer's suggestion, we have further clarified the Statistical Analyses section under Methods: "…We also present a modified intention to treat analysis for the primary and secondary outcomes, as a sensitivity analysis, to account for study participants who were randomised in error and those who withdrew consent prior to the intervention."

Comment 3. *One serious possible bias of this study is the insufficient information provided about the dose of concomitant medications, especially glucocorticoids. In fact, while dexamethasone 6 mg or equivalent for 10 days has become the most used protocol, it is not the only one and different doses and treatment duration may impact the outcome (please see DOI 10.1183/13993003.01514-2022 and related references). Therefore, at least the median glucocorticoid dose and duration (or cumulative prednisone dose) should be indicated for each group.*

Response: We thank the reviewer for this insightful comment and confirm that there were variations in the regime of steroid therapy between the participating study centres (dexamethasone 4mg for 3 days, dexamethasone 6mg for a duration of 7 – 10 days; methylprednisolone 80mg for 7days; methylprednisolone 120 mg per day for 5 days).

We have now included these details in the manuscript in Treatment Assignment and Compliance paragraph under Results section; "…Overall, steroid uptake was >95% in both treatment arms with different protocols used at participating study centres: dexamethasone 4 mg/day for 3 days; dexamethasone 6 mg/day for 7-10 days; methylprednisolone 80 mg/day for 7 days and methylprednisolone 120 mg/day for 5 days. However, there was no difference in the steroid treatment assigned between the control and the treatment groups."

The study protocol did not specify one particular corticosteroid regime within the "standard of care" given that the SOC continued to evolve. As we reported, the uptake of corticosteroids was greater than 95% in both control and treatment arms in all participating sites. The randomisation process would have ensured equal number of patients received a particular corticosteroid regime in each arm and avoided bias between the study groups.

It is noteworthy that the citation the reviewer directed us to (F Salton et al, Europ Resp J, 2022) reported no difference in mortality outcome comparing prolonged higher dose methylprednisolone with the equivalent of dexamethasone (6mg, 7-10 days course) in treating COVID-19 pneumonia in a multi-centre randomised trial, conferring that the prognostic benefit was due to the drug class effect rather than dose dependent effect.

Comment 4. *Similarly, this study has been conducted in very different clinical settings. It should be discussed and reported as one major limitation the fact that the availability of and invasive ventilation, as well as the time elapsed from clinical worsening to intubation or other internal protocols (e.g. pharmacological therapies, ventilation, high-flow nasal cannula etc.) may have diverged and implied different outcomes. Please also consider and discuss that the pressure on single hospitals may have*

been different between Centers during either the same or subsequent pandemic surges, potentially hindering the ability of the Center to cope with all the requests for an intensification of care. This might be the reason why the number of patients who underwent NIV was lower among those recruited in India. Indeed, it is very unlikely that the real need for ventilation was that different between Countries.

Response: The reviewer is correct to raise a concern about the difference in the provision of care between the institutions given the international multicentre nature of the study. The participating centres were selected because of the similarities in the provision of clinical care (including pharmacological therapies, non-invasive ventilation, invasive ventilation, provision of oxygen therapy). We agree that recruiting centres from different countries results in a degree of heterogeneity in clinical practices, as highlighted in the Results section. As the reviewer rightly concluded it is very unlikely that the real need for ventilation was markedly different between countries. The degree of heterogeneity between the centres may have eroded the power of the study to detect a statistical difference but added strength to it by reflecting the real-life situation. We have certainly recognised this as a potential Limitation of the study:

"…..Although patient characteristics and medications received as part of SOC did not differ between the randomised arms, the more heterogeneous population, milder COVID-19 disease, and more effective standard of care treatments most likely impacted on the hypothesised effect size and the ability of finding a difference in our recruited sample."

Comment 5. *The introduction and methods section should be expanded with some more details about the rationale (introduction) and application (methods) of the standard of care, with special regards to the therapies which have demonstrated a higher efficacy e.g. glucocorticoids, mechanical ventilation etc.*

Response: To address the reviewer's comment, we would first like to clarify that at the time of the study conception, the standard of care for COVID-19 was not yet established. As we stated, the primary objective of the study was to assess the impact of leflunomide in addition to continually evolving SOC. Hence, we did not define the SOC at the outset, but simply reported on the SOC used during the study. Accordingly, our Introduction explained the rationale of the proposed therapy, focussing on dual properties of leflunomide (anti-inflammatory and antiviral). We considered it more appropriate to discuss the effect of leflunomide with the background of emerging and established SOC in the Results and the Discussion sections. We hope this explains why the Introduction and Methods were written in this way. Rewriting these sections with the results of the studies published during or after active recruitment in our trial would shift the balance in the premise of our study.

Comment 6. *Please report the registration number of the study in the methods section.*

Response: The trial registration number is listed in a separate section under Trial registration as per journal submission format.

Comment 7. *Did the Authors exclude SARS-CoV-2-positive affected by respiratory failure or involvement due to other causes (e.g. congestive heart failure etc.) and patients with conditions (e.g. neurological ones) that could interfere with the success of noninvasive ventilation? If not, this should be discussed among the study limitations; otherwise, it should be reported among the exclusion criteria.*

Response: We did not specifically exclude patients in whom non-invasive ventilation was contraindicated due to neurological conditions but can confirm that none of the participants were in this category. This was a pragmatic trial selecting participants from a patient cohort admitted to the hospital for having moderate to severe symptoms of COVID-19 infection. It is also well recognised that acutely decompensated heart failure may be a complication from COVID-19 infection and patients would benefit from CPAP and non-invasive ventilation in an acute decompensation setting. It is difficult to conclusively differentiate between heart failure vs Covid progression as the cause of clinical deterioration because of an overlap in clinical picture when chronic pulmonary disease and heart failure coexist. Therefore, such exclusion could have led to bias due to inter-observer (clinician) variability.

Comment 8. *The time from hospitalization to study enrolment should be reported.*

Response: The median time from hospitalisation to study enrolment was 2 days (IQR 1-3) for both SOC+L and SOC groups. We have now added the information in baseline characteristics table (Table 1).

Comment 9. *Is the baseline PaO2/FiO2 ratio available?*

Response: The study was set out to enrol patients with hypoxia on room air, needing oxygen therapy and having at least moderate symptoms of COVID-19 infections. The baseline median $PaO_2/Fi0_2$ ratios were 213 mmHg (IQR: 122 - 253) for SOC group and 206 mmHg (IQR: 137 - 266) for the SOC+L group. The PaO_2/FiO2 data were not uniformly available for all study participants, thus we chose to report SpO_2/FiO_2 ratio instead.

Comment 10. *Non-invasive ventilation was required for 14.4% of patients in SOC+L group vs. 16.4% in the SOC group. Do this ratio include patients who were already on NIV at baseline? In any case, the number of patients on NIV at baseline should be reported in baseline data, whereas the number of patients who encountered the need for noninvasive or invasive ventilation (separately) due to clinical worsening should be reported separately in the text. Please also add the p-values where appropriate.*

Response: Those ratios included patients who were already on non-invasive ventilation at baseline. The % of patients needing non-invasive ventilation at the time of study enrolment was 4.8% in SOC+L group vs. 7.3% in SOC group, p=0.451. The % patients who then went on needing non-invasive ventilation support was 9.6% in SOC+L group vs. 9.1% in SOC group, p = 1.000.

The % of patients needing invasive ventilation at the time of study enrolment was 1.0% in SOC+L group vs. 1.8% in SOC group, p = 1.000. The % patients who then went on needing invasive ventilation support was 3.9% in SOC+L group vs. 5.5% in SOC group; p = 1.000.

We have updated the baseline characteristics table to include these observations..

Comment 11. *It would be useful to perform a stratified analysis for clinical severity at baseline (e.g. no oxygen, oxygen, NIV) with regard to the primary endpoint.*

Response: Patients underwent stratified randomisation, based on NEWS2 score (which captures baseline clinical severity) and the presence of comorbidities considered to be indicators of high risk

for an adverse outcome. This approach allowed us to carry out a statistically valid stratified analysis for the primary outcome. The table below documents the proportion of patients achieving the primary outcome in each of the four stratification groups. Groups 1 and 3 represent those patients with a high NEWS2 score. As can be seen, there is no numerical or statistical difference in treatment effect between the groups. Based on these results, any post hoc re-stratification exercise looking at sub-metrics of the NEWS2 score would simply increase parameter uncertainty and would not be expected to yield a better understanding of the study results.

Stratification group	Patients (n=214)		Patients achieving primary outcome (n)		Patients achieving primary outcome (%)			p*
	SOC+L	SOC	SOC+L	SOC	SOC+L	SOC	Total	
Group 1 (n=29)	13	16	9	11	69.2	68.8	69.0	0.670
Group 2 (n=33)	15	18	11	13	73.3	72.2	72.7	0.448
Group 3 (n=101)	50	51	46	49	92.0	96.1	94.1	0.683
Group 4 (n=51)	26	25	25	25	96.2	100	98.0	0.261

Comment 12. *The timepoints at which the SpO2/FiO2 ratio has been evaluated should be reported in the methods.*

Response: It is standard clinical practice that SpO_2 is monitored every 4 hours in a clinically stable patient. The frequency would increase to continuous SpO_2 monitoring in a patient with oxygen requirement or ventilation support. Our protocol did not deviate from this widely accepted clinical practice. We have now added this clarification in the Methods section as the reviewer suggested: "SpO_2/FiO_2 data were monitored daily. The frequency of SpO_2 monitoring varied with FiO_2 administration. It is standard clinical practice that SpO_2 is monitored every 4 hours in a clinically stable patient. The frequency increases to continuous SpO_2 monitoring in a patient with oxygen requirement or ventilation support. Where multiple daily values were recorded we selected the SpO_2/FiO_2 ratio reflecting increased oxygen demand."

Comment 13. *The Authors state that the sample size calculation was based on the proportion of patients expected to meet the outcome criteria by 28 days. This is not convincing. Please either report previous literature data used to estimate this proportion or discuss this as a major limitation of the study design.*

Response: The reviewer is correct that the calculation of the sample size was subject to substantial uncertainty, given that the study was designed in April 2020, at which time there was

extremely little data available for how treatments might affect the clinical path of COVID-19. Below is the relevant section from the SAP:

Given the still-evolving nature of our understanding of COVID-19, input assumptions for the sample size calculation are subject to a significant degree of uncertainty. For this reason, an initial estimate of sample size was made, subject to review and adjustment at an early stage in the study, once a more reliable estimate could be derived. The analytical strategy for this early interim analysis is described in section 4h below.

The primary outcome measure is a time-to-event analysis, based on an assessment of TTCI. The results of the comparison will be assessed using a Cox-derived hazard ratio, with p-value being based on the logrank method. As we do not have any indication from prior studies of the likely hazard ratio, we have used the method described by Machin et al[1], which is based purely on the proportion of patients expected to achieve the outcome criteria by 28 days.

In the study by Cao et al[2] investigating lopinavir-ritonavir in COVID-19, 22% of patients died, while 74% had achieved clinical improvement by day 28, the remaining 4% requiring ongoing care. Median TTCI was 16 days and median length of stay was 13 days. By contrast, an analysis of 16,749 patients admitted to UK hospital with COVID-19 between 6/2/20 and 18/4/20 showed an overall mortality rate of 33%, with 17% requiring ongoing care at the time of the analysis[3]. TTCI was not assessed in this study, but median length of stay was 7 days. Of note, the minimum period of post admission follow-up in the UK study was 14 days, which will tend to affect patients recruited in April, so it is likely that the proportion of patients requiring ongoing care is only an approximation of the 28 day figure.

These data suggest that there are significant differences in both patient characteristics and clinical practice between the UK and Wuhan. In this context, it is probably unreliable to directly use the Cao data to guide our sample size.

Based on the 33% figure for UK overall mortality and a further assumption that the proportion of patients requiring ongoing care will be 17% at 28 days, we have estimated that the proportion of patients in the control arm meeting the primary outcome TTCI criteria at 28 days will be 50%.

We have further assumed that the use of leflunomide will increase this proportion to 72.5%, based on an appraisal of data seen in the pilot study.

Based on this, and assuming: α = 0.05; β = 0.20; allocation ratio =1:1, the required number of patients per treatment arm is estimated as 74. Assuming a 20% attrition rate, the total number of patients required in the study will therefore be 178 – representing 89 patients in each arm.

1. *Machin D, Campbell M, Fayers, P, Pinol A (1997) Sample Size Tables for Clinical Studies. Second Ed. Blackwell Science IBSN 0-86542-870-0 p. 176-177*
2. *Cao B, Wang Y, Wen D, et al; A trial of lopinavir-ritonavir in adults hospitalized with severe Covid-19. N Engl J Med 2020; 382:1787-99*
3. *Docherty AB, Harrison EM, Green CA, et al. Features of 16,749 hospitalised UK patients with COVID-19 using the ISARIC WHO Clinical Characterisation Protocol. Available at: https://www.medrxiv.org/content/10.1101/2020.04.23.20076042v1.full.pdf*

The power calculation assumed that 33% of the patients in the control arm would die and a further 17% would require ongoing care at 28 days. The true figures from the study were 9.1% and 0% - presumably reflecting various factors including a change in the prevalent viral variant and adoption of more established standard of care therapy by the time the study was recruiting. We therefore agree that the study was underpowered and have included a statement to this effect in the "Study limitations" section:*"…*Although patient characteristics and medications received as part of SOC did not differ between the randomised arms, the more heterogeneous population, milder COVID-19

disease, and more effective standard of care treatments most likely impacted on the hypothesised effect size and the ability of finding a difference in our recruited sample."

Comment 14. *The Authors stated that the primary analysis was stratified by "baseline risk indicators". However, what is meant for baseline risk indicators is not clear and it is worth further elucidation in the methods section.*

Response: We have now defined "baseline risk indictors" in the Statistical section as "age </≥ 70; co-morbidities, clinical status based on NEWS2 scores" to conform to the same description used in the Randomisation section of the Methods.

Comment 15. *"During the data cleaning process, 10 patients were flagged as not meeting the inclusion criteria (6 in SOC+L; 4 in SOC), as they did not have moderate COVID-19 symptoms at the time of randomisation." Having moderate COVID symptoms seems not a prespecified inclusion criterion to me. Please reformulate this statement.*

Response: We have defined that "… patients with respiratory compromise and blood oxygen saturation (SpO_2) <93% on room air detected on pulse oximeter were considered to fulfil the moderate COVID-19 infection." in the Participants section under the Methods.

Comment 16. *Table 1, chronic neurological disorders seem significantly different between groups to me. Please remove the p-value column from this table and discuss eventual differences at baseline in the results section.*

Response: We thank the reviewer for carefully reviewing the manuscript. There were indeed significantly more patients with chronic neurological disorders in the SOC+L group, due to a history of cerebral vascular events. This difference is most likely due to chance as randomisation could not control for all the baseline patient characteristics. We have added a statement in the Results section to account for this notable difference: "…There was a significant difference with a higher proportion of patients with chronic neurological disorders in the SOC+L group, mostly due to a history of cerebral vascular events. None of the patients were contraindicated to have NIV due to numerological disorders."

We have further removed the p-values from the table, as suggested by the reviewer.

Comment 17. *Table 2, please add percentages and p-values. Furthermore, I don't understand the first line Adverse events (n) / Patients (n). Please explain. In any case, total AE should be expressed as n. of patients with at least one AE/total n. of patients.*

Response: In Table 2: "Adverse events (n)" indicated the total number of AE; "Patients (n)" indicated the total number of patients with at least one AE. Reporting both these values is more informative in the context of the present study as it shows that some patients experienced multiple adverse events. The calculation of p-values for AEs in clinical trials is generally considered to represent spurious accuracy, given the way that data are collected and post-processed. Whilst we can carry out the calculation, the results would not be informative.

We have updated the Table 2 and its legend stating the definition of the terms/data to make it clearer.

Comment 18. *In the adverse events section it is reported that 15 vs 9 patients died. Is it meant due to AE? Why is it reported in this section? If so, given that the difference seems macroscopically significant, is should be discussed.*

Response: We would like to point out that in Table 2, we stated 9 vs. 10 patients died in the interventional and control arm, respectively, and this was not significantly different. As is a standard practice in clinical trials all deaths are reported as serious AE, whether or not related to the treatment.

Comment 19. *Please report the unit of measure of data reported in the paper e.g. those regarding viral load and CRP*

Response: The viral load is expressed as $\log^{10}$ copies/ml, and CRP is expressed as mg/L. These units have been described in the respective results section. We have now also added the respective units to Table 1.

Comments from Reviewer: 4
Dr. Paul McGale, University of Oxford
Comment 1: *This is a phase 3 open label randomised controlled trial, assessing the efficacy of adding leflunomide to standard care of patients hospitalised with COVID-19. Inclusion and exclusion criteria and trial procedures are well described. It would help to add a sentence justifying why the randomised treatment is unblinded.*

Response: We provided a detailed response to a similar comment from Reviewer 2 explaining why we adopted an open label study. The following paragraph was added to the Limitations section as the reviewer suggested:

"In order to balance the needs of the trial with clinical care and to minimise disruption to already overstretched clinical resources during COVID-19 pandemic, we chose to adopt an open label study. An open label design could potentially introduce data analysis bias. However, it would also allow early detection of significant adverse events and a potential outcome benefit. This was an important consideration when testing an off-label use of a medication in COVID-19, a disease with high morbidity and mortality."

Comment 2. *The statistical methods are appropriate, however, the proposed use of the logrank test does not seem to be reported in the main text results. The authors say the CIs of the hazard ratios were used for the significance of the treatment effect, were these different from the logrank p-values?*
Response: The results of the log rank estimates for the primary outcome in both analysis sets (p=0.070 and p=0.028) have already been reported in the relevant paragraph in the Results section. They are qualitatively consistent with the results based on assessment of the confidence intervals.

Comment 3. Figures 2, 3, & 4 are rather uninformative without the addition of the results of the statistical tests used to compare the outcomes by randomised treatment. These should be added to the figures.

Response: We have updated Figures 2 and 3 with the respective hazard ratios, 95% confidence interval and p values.

The viral load data (Figure 4) did not lend itself to between-group comparisons across time. Undertaking this analysis would have involved comparing different number of samples per randomisation group at every time point, which we feel would have ultimately misrepresented the actual changes in viral load because the samples were not available from every patients and every time point. To minimise this type of error, we deemed it more appropriate to analyse viral load across time through within-groups paired samples analysis.

By clustering data over specific time points, we could better represent the overall distribution across time. Additional quantitative comparisons were deemed not appropriate for these data. We made the following update to the description of the Viral load section findings under Results section:

"Quantitative SARS-COV-2 PCR measurements from nasopharyngeal swabs at baseline showed no difference in median log viral loads between the two groups, SOC+L 4.68 (IQR 4.45-4.85) vs SOC 4.76 (IQR 4.48-4.92), ($p = 0.272$) We clustered the serial samples to reflect the crucial time intervals during the hospital stay: time coinciding with finishing leflunomide loading dose (by Day 4), time to 75% patients being discharged from hospital (by Day 7), time to finishing leflunomide maintenance dose (by Day 11) and beyond (Figure 4). Viral loads were significantly reduced in both treatment arms by Day 7, $p<0.001$; and by Day 11, $p <0.030$. The rate of viral load reduction between groups by Day 11 appeared to be similar."

VERSION 2 – REVIEW

REVIEWER	Salton, Francesco
REVIEW RETURNED	04-Mar-2023

GENERAL COMMENTS	The Authors have addressed most part of my concerns and they have given sufficient explanation for those they have not implemented in the manuscript. I have no other comments.