Peer Review File

**Manuscript Title:** Sex differences in immune responses that underlie COVID-19 disease outcomes

**Editorial Notes:**

**Redactions – unpublished data**

Parts of this Peer Review File have been redacted as indicated to maintain the confidentiality of unpublished data.

**Reviewer Comments & Author Rebuttals**

**Reviewer Reports on the Initial Version:**

Referees' comments:

Referee #1 (Remarks to the Author):

Takahashi et al assess sex-differences in the immune response to SARS-CoV2 and this is an issue of outmost importance given the clear differences seen in disease severity between men and women with COVID-19.

I find the experiments clear and well described, the cohort is a bit small for some of the conclusions drawn but the overall message is very interesting and important.

Study is performed on hospitalized patients which means there was already a selection of symptomatic cases? Or were some patients PCR-tested and found to be positive accidentally and admitted to the hospital for other reasons? Please clarify this.

In Figure 1c, there are some vertical lines of high levels across many cytokines in some selected individuals. I worry about either batch effects or some other technical artifacts in these experiments. Were these cytokine measurements performed in batches? If yes, did the authors perform batch correction? This would be useful to know a bit more about and I see no methods section detailing this.

This sentence "...the median concentrations of virus RNA assayed with nasopharyngeal swabs and saliva were both higher in males, but there was no significant difference by sex." is unclear. Are the viral loads different or not? Please clarify your position here.

Overall heatmaps presented (Figures 1C, 2A, 3A, 3G) are ordered by correlation (dendrogram) on the Y-axis but by group on the X-axis. Clusters are usually not seen, and I understand the purpose of showing comparisons between groups of subjects, but I also wonder if the groups would naturally cluster if also the X-axis was ordered by data alone, ie a dendrogram? I think such unbiased analyses would be very interesting to see.

Figure 2E is very interesting showing the correlation between non-classical monocytes and CCL5 levels. It seems to me the that the male group is split in two, with some males having "female-like" CCL5-to-ncMono relationships, while a few males stand out as very different. Can the authors elaborate anything more about possible other differences among these two groups of male subjects? Could it be that these subjects are sampled at different stages of their infection and therefor display such differences? Interesting observation!
In the analysis of PBMC differences among men and women, I wonder if the overall lymphocyte count and Neutrophil to lymphocyte ratios are different among males and females. There have been several studies suggesting these as possible prognostic markers in COVID-19 (Lagunas-Rangel et al, 2020). I wonder if males patients in this cohort have more pronounced lymphopenia than women? especially given the differences in IL-8 levels seen.

The analysis of progressive vs stable disease is my favorite part of the paper. This controls somewhat for possible differences in timing of sampling during the natural course of infection. Some of the conclusions here are uncertain due to the large inter-individual variation seen. for example, In Figure 4E, the differences in activated CD8+ T-cells between men and women are largely explained by a few women with very high levels, while most other females overlap quite well with the males.

Overall the authors have rapidly addressed a very important aspect of the COVID-19 pandemic and they have done so in a sound manner. The presentations mostly highlight raw data and conclusions are fair. Larger cohort studies will be required to confirm some of the observations but this study represent a very good starting point for such future follow-ups.

Referee #2 (Remarks to the Author):

Gender differences in response to viral infections is well documented and this study builds on these hypothesis. The authors extensively examine the impact of sex difference in a range of virological (viral load) and immunological (antibody responses, plasma cytokines, blood cell phenotypes) in a cohort of patients with COVID-19. They observe that males had higher plasma levels of IL-8, ILL-18 and CCL-5 as well as more non-classical monocytes. Females mount more robust T cell activation which is well retained irrespective of age. They also find that poor T cell responses were correlated with poor outcome. Overall, this is a well written, extensive and systematic analysis of the subject. The male and female patients were well balanced in their baseline characteristics.

One major and critical omission in the analysis is that there is no mention made of time after onset of illness to clinical recruitment? This should be noted in Extended data table 3 and this has to be factored in to the analysis? In theory, all the differences in male vs. female observed may be just a reflection of a trend in difference in time to admission between the sexes?

Specific questions:

Line 98. The health care worker control group. How do you know if they may not have been asymptomatically infected? Was serology done to exclude any SARS-CoV-2 sero-positive health care workers?

Many references in text to differences between males and females that are not statistically significant. E.g. IL-18 (line 161) CXCL10 (Line 164). Given the multiple parameters measured, it is not helpful to keep referring to differences that are not significant.

Line 214 onwards. The patients were categorised into deteriorated group vs stabilised groups if clinical scores at any point was worse that the initial clinical score. In theory, this could include anyone who went from clinical score 1 to 2 as deteriorated, whereas a patient who started at 3 and remained at 3 would be regarded as stabilised? I do not see such an example, but the current strategy would allow for such. While the current analysis may be useful to investigate "trajectory” of illness, is it not also useful to do an additional analysis of patients who consistently remain at level 1 vs those who are level 2 and higher, irrespective of baseline and trajectory? Again, the classification of "deteriorated” becomes meaningless in the lack of data on time after onset. Clearly, most patients would “deteriorate” if you compare symptoms on day of onset of illness vs 7 days later. The lack of a “time-frame” in relation to time after onset of illness is a challenge to
interpreting the data.

Referee #3 (Remarks to the Author):

The primary goal of this study was to systematically evaluate sex differences in immune responses to SARS-CoV-2 in patients with mild to moderate disease in comparison with healthy controls as noted in the title and abstract. The paper is very well written and organized. The healthy controls were healthcare workers (HCW) and a majority of the differences reported in this paper were not actually sex differences in immune responses to SARS-CoV-2 but rather sex-specific differences in the immune responses of patients relative to controls. There were a number of confounds and statistical errors that reduce the quality of the data presented.

Major:
1. Confounding variables.
   a. HCW are not an appropriate control group for the patients, given age, BMI, and lack of information on pre-existing conditions. If this is the only control group available, the baseline differences need to be thoroughly accounted for in analysis.
   b. For many analyses (e.g., cytokines and immune cells), changes in responses or cell counts could just be reflective of the older age or greater BMI of patients compared with controls as both of these factors contribute to a greater baseline inflammatory state. Interestingly, both of these factors explained a lot of variability when comparing stabilized to deteriorated patients, but were never considered in the comparisons of patients vs. controls. For the T cell counts that appear to be reduced in patients relative to controls (Fig 2b), for example, it could be that aging rather that COVID-19 is explaining these differences.
   c. Lacking discussion of discordance in findings between Cohorts A and B; Cohort B included more severe patients that Cohort A – this should be acknowledged and could explain some of the discordance. It also should be made clear when data from one or both Cohorts are included in analyses.
   d. Lines 464 – 466: dropped values in top and bottom 1st percentile. Needs to be made explicitly clear how many values were dropped for each cytokine (how many of the 38 patients?). In Figure 1 legend, the sample size indicated suggests it was the same for all cytokines, but this contradicts what is written in the methods. May also want to include in supplemental how this affected analysis – were findings very different if the outliers were included?
   e. In many cases (e.g., Figure 1) samples sizes don’t add up (i.e., counting dots vs the sample size lists in the figure legend).
   f. For all measures, what was the timepoint when samples were collected and was that based on time since PCR+ test, time since hospital admission, or something else?
2. Statistical analyses.
   a. Many comparisons of male HCW vs male patients and of female HCW vs female patients, and then looking at “difference of differences” A more interesting comparison would be male vs female HCW and male v female patients, and then, potentially, the difference in differences.
   b. Figure 4a: I think all these figures would benefit from a third panel that is not sex-disaggregated, or some other way to account for differences between males and females in the study. A study of this size is very vulnerable to groups being unequal at baseline (e.g.: males being older than females), and I see no evidence that is accounted for in these comparisons.
   c. Extended data table 6: lacking comparison between males and females within each disease state, as opposed to disease state within each sex.
   d. Used propensity scores to adjust for age as a confounder (lines 480 – 481), but given the importance of age as a risk factor for disease, I would think you want this as a main effect in the model, and potentially an interaction term between age and sex.
   e. For analyses of viral loads, how can viral loads be detected below the limit of detection and then how can those values be included in the statistical analyses?
   f. Uniform management of outliers as there are cases where outliers were removed (see above)
and in other cases (e.g., Fig 3d and f) they are included and driving the effect.

3. Overinterpretation.
a. Lines 165 – 167 “innate inflammatory cytokines and chemokines are more robustly elevated early and throughout disease course in male over female patients”. From my understanding, they measured 71 cytokines and found significantly higher levels in males compared to females for one in each cohort (IL-8 in cohort A and CCL5 in cohort B). I don’t think this supports the conclusion above.
b. Classification as stable vs deteriorated (lines 212 -216). I could not find any explanation of when enrollment and the baseline sampling occurred relative to hospital admission. Was there an attempt to control for differences in disease severity at enrollment? The disease process of deteriorating from score 1 to 2 may not be the same as from 3 to 4.
c. Figure 2c has a representative FACS image for Figure 2d. However, the FACS image does not appear representative of the monocyte graph in Figure 2d. The percentages of ncMono in male and female HCW are 7.82 and 2.43, respectively, but the graph shows that the mean % ncMono of male and female HCW is similar. The percentage of cMono in HCW is about 60% in both males and females, which is about 2-fold higher than cMono % in patients in Figure 2c. However, Figure 2d shows that % cMono is slightly higher in the patients compared to HCW.
d. In the abstract, the authors state that the plasma level of CCL-5 along with IL-8 and IL-18 is higher in male patients compared to female patients. However, Figure 4b and Extended Data Figure 2 clearly show that the plasma level of CCL-5 in male patients is not higher than that in female patients. Extended Data Table 4 shows that there is a significant difference of CCL-5 between male and female patients in Cohort B. In the same manner, IL-8 level did not show any difference between male and female patients in Cohort B. However, Figure 1d shows that IL-8 level in male patients is significantly higher than that in female patients in Cohort A. The authors appear cherry-picked the data in Cohorts A and B to support their theory.

4. Cohorts.
a. The authors have divided up patients into two cohorts based on whether or not they were baseline versus longitudinal or whether they had not received antiviral or steroidal interventions. A third cohort may be useful in directly comparing patients who were and were not treated with antivirals and corticosteroids at baseline, just to see how this affected viral loads and innate immune responses.
b. It is unclear how these “mild” patients are being considered mild since they are still admitted to the hospital. The authors would simplify this by comparing their own clinical scoring to that of the WHO for COVID19 disease severity, in order to give the audience a better reference of clinical diagnosis.
c. The cohort used should be referred to in each figure legend.

Minor:
1. Incorporate tables, including at least one demographic table comparing males and females in each cohort into the main text
2. Need more information on what “first” time point means – is this at hospital admission? Any control or acknowledgement that people may have been at very different points in disease progression at “baseline”? Unclear if or how this was controlled for.
a. Some of this information is included in extended table 3, but nowhere in main body of the text.
3. Description of longitudinal analysis is not clear – needs some detail about how many time points per person and over what length of time (days? Weeks?)
4. Number of samples taken per subject is included in extended data table 1, but no information about timing of samples is included.
5. For each of the heatmaps, the “color key” seems to be missing color, as it is showing the scale for Z scores but there is no color in the boxes. I am assuming this is meant to show color intensity and that there was just an error in copy/pasting or formatting.
6. These heatmaps show quite a bit of heterogeneity and no clear pattern and may be better suited to supplementary figures and replaced with some other graphs from supplemental.
7. The limit of detection or cutoff value is missing from the antibody graphs (Figure 1b, 4a) as well as how that limit was set.

8. Why is the race and ethnicity information missing from all controls?

**Author Rebuttals to Initial Comments:**

Referees' comments:

Referee #1 (Remarks to the Author):

Takahashi et al assess sex-differences in the immune response to SARS-CoV2 and this is an issue of outmost importance given the clear differences seen in disease severity between men and women with COVID-19.

I find the experiments clear and well described, the cohort is a bit small for some of the conclusions drawn but the overall message is very interesting and important.

We are very grateful for the positive and constructive suggestions by this referee. We have addressed all issues raised by the reviewer, and believe that the revised manuscript is much improved as a result.

Study is performed on hospitalized patients which means there was already a selection of symptomatic cases? Or were some patients PCR-tested and found to be positive accidentally and admitted to the hospital for other reasons? Please clarify this.

All patients necessitated hospitalization for their symptoms (Patients were admitted at the discretion of the emergency room doctor who evaluates the patient. WHO score ≥ 3 (=Hospitalized Mild disease, WHO COVID-19 Therapeutic Trial Synopsis, February 18, 2020)).

As the initial screening, clinical PCR tests were performed in CLIA-certified laboratory and only the PCR-positive patients were enrolled. Only after the confirmation of PCR-positivity, the patients were enrolled into this study and the first time point samples were collected for each patient. This description has now been added to the Main text and the Methods section.

In Figure 1c, there are some vertical lines of high levels across many cytokines in some selected individuals. I worry about either batch effects or some other technical artifacts in these experiments. Were these cytokine measurements performed in batches? If yes, did the authors perform batch correction? This would be useful to know a bit more about and I see no methods section detailing this.
The plasma samples for ELISA were sent out and measured at Eve technologies (Calgary, Canada) in two batches, but these two batches were measured with the same kit using the same standard curves.

For example, Tedesco et al. (Convenience versus Biological Significance: Are PMA-Differentiated THP-1 Cells a Reliable Substitute for Blood-Derived Macrophages When Studying in Vitro Polarization?, Front Pharmacol, 2018; PMID: 29520230) used the same luminex ELISA assay of this company in their paper, and they discussed how they took into consideration the batch effect of the assays from different assay runs. They discuss as follows:

“(samples) were analyzed in different runs, a quantitative comparison for single analytes was performed only when the data fulfilled the following criteria: (a) the standard curves of the two data sets had overlapping shapes, and (b) the fluorescence intensity (FI) values obtained from all the samples of at least one cell type fell within the central part of the standard curve…”

In our case, two batches of measurements were done with the exactly same standard curves, and we do not expect the significant batch effects between these two batches.

For this point, we added the following description in the Method:

“The shipment of the samples and measurements were done in two separate batches, but the measurements were performed with the same assay kits using the same standard curves, therefore minimizing the batch effects between the measurements.”

This sentence “…the median concentrations of virus RNA assayed with nasopharyngeal swabs and saliva were both higher in males, but there was no significant difference by sex.” is unclear. Are the viral loads different or not? Please clarify your position here.

There was no statistical difference between male and female patients. We modified the description to clarify this point.

Overall heatmaps presented (Figures 1C, 2A, 3A, 3G) are ordered by correlation (dendrogram) on the Y-axis but by group on the X-axis. Clusters are usually not seen, and I understand the purpose of showing comparisons between groups of subjects, but I also wonder if the groups would naturally cluster if also the X-axis was ordered by data alone, ie a dendrogram? I think such unbiased analyses would be very interesting to see.
We modified all the heatmaps to the ones with X-axis ordered by dendrogram, and per suggestion from Reviewer #3, we moved heatmaps to the Extended Data Fig. 2.

Figure 2E is very interesting showing the correlation between non-classical monocytes and CCL5 levels. It seems to me that the male group is split in two, with some males having “female-like” CCL5-to-ncMono relationships, while a few males stand out as very different. Can the authors elaborate anything more about possible other differences among these two groups of male subjects? Could it be that these subjects are sampled at different stages of their infection and therefore display such differences? Interesting observation!

We are grateful for this insightful suggestion. According to this suggestion, we divided male patients into two groups: “high” group who had high ncMono percentages (upper quartile, 4 patients in 17 patients, all had > 5% of ncMono in % of Live) and “low-int” group (others). While there was no difference in terms of age, BMI, and DFSO, “high” group had significantly low T cell levels in addition to higher plasma IL-18 and CCL5 levels. We added these results in the revised figure (Fig. 2d) and some description in the manuscript as follows:

“We then divided 17 Cohort A male patients into two groups, namely, "high" group who had high percentages of ncMono (upper quartile 4 patients, all had > 5% of ncMono) and “low-int” group (others, 13 patients). We compared age, BMI, DFSO, T cells, and plasma IL-18, and CCL5 levels. While we found no difference in age, BMI, DFSO (Fig. 2d), we noted that high-ncMono group had significantly lower T cell levels and higher plasma CCL5 levels (Fig. 2d)."

In the analysis of PBMC differences among men and women, I wonder if the overall lymphocyte count and Neutrophil to lymphocyte ratios are different among males and females. There have been several studies suggesting these as possible prognostic markers in COVID-19 (Lagunas-Rangel et al, 2020). I wonder if males patients in this cohort have more pronounced lymphopenia than women? especially given the differences in IL-8 levels seen.

First, we compared the B cell and T cell count in Cohort A. The results are shown in Fig. R1a (M_HCW:F_HCW:M_Pt:F_Pt = 6 : 42 : 10 : 13; because of the missing data for cell counts in Cohort A patients, sample size is smaller than total Cohort A size). While B cell numbers were comparable among all groups, T cell numbers were markedly decreased in patients compared to HCW controls. However, there was no difference between male and female patients, in parallel with the panel on % of Live in Fig. 2a.
Next, we compared T cell number and B cell number in Cohort B, and found that T cell number is significantly lower in male patients compared with female patients (Extended Data Table 4). Therefore, the results collectively suggested that elevation of important cytokines/chemokines including IL-8 might be correlated with the T cell number decline in male patients longitudinally, and we added description on this point in the results.

The analysis of progressive vs stable disease is my favorite part of the paper. This controls somewhat for possible differences in timing of sampling during the natural course of infection. Some of the conclusions here are uncertain due to the large inter-individual variation seen. for example, In Figure 4E, the differences in activated CD8+ T-cells between men and women are largely explained by a few women with very high levels, while most other females overlap quite well with the males.

Indeed, there are some female patients who had high percentages for CD38+HLA-DR+CD8 T cells. In order to see if indeed these women skewed our overall analysis, we excluded 4 female patients who had CD38+HLA-DR+CD8 T cells > 3% (Fig. 4f) and redid the analysis. The result is shown in Fig. 4b. Even with the exclusion of these patients, there was no statistically significant correlation between age and CD38+HLA-DR+CD8 T cells (p = 0.375, R = -0.22) in female patients, which still support our observation described in the manuscript.

[Redacted]

Overall the authors have rapidly addressed a very important aspect of the COVID-19 pandemic and they have done so in a sound manner. The presentations mostly highlight raw data and conclusions are fair. Larger cohort studies will be required to confirm some of the observations but this study represent a very good starting point for such future follow-ups.

We are very grateful for the insightful and constructive criticisms of this referee.
Referee #2 (Remarks to the Author):

Gender differences in response to viral infections is well documented and this study builds on these hypothesis. The authors extensively examine the impact of sex difference in a range of virological (viral load) and immunological (antibody responses, plasma cytokines, blood cell phenotypes) in a cohort of patients with COVID-19. They observe that males had higher plasma levels of IL-8, IL-18 and CCL-5 as well as more non-classical monocytes. Females mount more robust T cell activation which is well retained irrespective of age. They also find that poor T cell responses were correlated with poor outcome. Overall, this is a well written, extensive and systematic analysis of the subject. The male and female patients were well balanced in their baseline characteristics.

We are grateful to receive positive and constructive suggestions from this referee. We have addressed all issues raised by the reviewer, and believe that the revised manuscript is much improved as a result.

One major and critical omission in the analysis is that there is no mention made of time after onset of illness to clinical recruitment? This should be noted in Extended data table 3 and this has to be factored in to the analysis? In theory, all the differences in male vs. female observed may be just a reflection of a trend in difference in time to admission between the sexes?

We regret that this information was not clearly explained in our original description. We had included the days after onset of illness as “DFSO” column (days from symptom onset) in Extended Data Table 3 (now Extended Data Table 2 in revised version).

As you can see from Extended Data Table 2 (demographic for Cohort A patients), there was no statistical difference in terms of clinical recruitment since DFSO between men and women in Cohort A patients. Since this is an important point as pointed out by this referee, we made a new panel to show this and highlighted it in Extended Data Fig. 1a.

Specific questions:
Line 98. The health care worker control group. How do you know if they may not have been asymptomatically infected? Was serology done to exclude any SARS-CoV-2 sero-positive health care workers?

Yes. HCWs were tested every 2 weeks for PCR and serology. For the control group, the PBMCs and plasma analysis were done when both tests were negative. In other words, if either or both of these tests were positive, these samples were excluded from the analyses. We have now included this description in the Methods section.

Many references in text to differences between males and females that are not statistically significant. E.g. IL-18 (line 161) CXCL10 (Line 164). Given the multiple parameters measured, it is not helpful to keep referring to differences that are not significant.
We understand and agree with this argument. However, some of the differences with p-values very close to 0.05 might in fact reflect the significant differences between sexes, but did not reach significant due to the small size, especially for Cohort A.

For example, in the new Fig. 4a illustrating sex-aggregated panels according to the suggestion from Reviewer #3, the p-value of difference in age between stabilized group and deteriorated group was p=0.051. However, we do not think this difference is meaningless. Therefore, while reducing the remarks on the differences that are not significant, we kept some of these remarks.

Line 214 onwards. The patients were categorised into deteriorated group vs stabilised groups if clinical scores at any point was worse that the initial clinical score. In theory, this could include anyone who went from clinical score 1 to 2 as deteriorated, whereas a patient who started at 3 and remained at 3 would be regarded as stabilised? I do not see such an example, but the current strategy would allow for such. While the current analysis may be useful to investigate “trajectory” of illness, is it not also useful to do an additional analysis of patients who consistently remain at level 1 vs those who are level 2 and higher, irrespective of baseline and trajectory? Again, the classification of “deteriorated” becomes meaningless in the lack of data on time after onset. Clearly, most patients would “deteriorate” if you compare symptoms on day of onset of illness vs 7 days later. The lack of a “time-frame” in relation to time after onset of illness is a challenge to interpreting the data.

The information for time frame (days from symptom onset) at the first sampling was included in previous Extended Data Table 3 (DFSO) as mentioned in the response to the first comment. To clarify the temporal relationship between the first sampling and the timing that patient reached Cmax (maximum clinical score recorded after the first sampling), we included one additional column, “DFSO at the first day of Cmax” for deteriorated patients in new Extended Data Table 2.

Regarding the temporal relationships between Cmax and C1, we calculated (DFSO at the first day of Cmax) – (DFSO at C1) in “deteriorated” groups, and the result is shown in Fig. c, and there was no statistical difference between sexes. For male, 3.7 ± 4.1 and for female, 4.2 ± 2.7, p = 0.81).
In addition, we realized that our original description of how we assigned “stabilized” vs. “deteriorated” disease in this analysis was unclear.

Original: The patient was categorized into the deteriorated group if the patient marked a worse, or higher, clinical score at any point compared to the patient’s initial clinical score (see Extended Data Table 3. Cmax > C1), and were otherwise categorized as stabilized.

We have modified this sentence to clarify our classification better, and also added some descriptions about the above points as follows:

Revised: The clinical scores at the first sample collection (C1) were 1 or 2 for all of the Cohort A patients. The patient was categorized into the “deteriorated” group if the patient marked a score of 3 or higher after the first sample collection date as their maximum clinical scores during admission (Cmax), and if the patient maintained the score of 1 or 2, they categorized as stabilized (see Extended Data Table 2 for the detailed information for each patient). Both in male (N=17) and female (N = 22) Cohort A patients, 6 patients in each group deteriorated in their disease course (35.3% and 27.3%, respectively), and the intervals between the dates on which the patients reached Cmax (DFSO at Cmax) and the first sample collection (DFSO at C1) were not significantly different between deteriorated male patients and female patients (mean ± SD = 3.7 ± 4.1 and 4.2 ± 2.7, respectively. p = 0.81).
Referee #3 (Remarks to the Author):

The primary goal of this study was to systematically evaluate sex differences in immune responses to SARS-CoV-2 in patients with mild to moderate disease in comparison with healthy controls as noted in the title and abstract. The paper is very well written and organized. The healthy controls were healthcare workers (HCW) and a majority of the differences reported in this paper were not actually sex differences in immune responses to SARS-CoV-2 but rather sex-specific differences in the immune responses of patients relative to controls. There were a number of confounds and statistical errors that reduce the quality of the data presented.

We are grateful to receive careful and constructive reviews from this referee. We have now addressed all issues related to confounds and statistical analyses in the revised manuscript.

Major:
1. Confounding variables.
a. HCW are not an appropriate control group for the patients, given age, BMI, and lack of information on pre-existing conditions. If this is the only control group available, the baseline differences need to be thoroughly accounted for in analysis.

For the analyses in the main figures, we agree with this comment that age and BMIs are not matched between HCWs and patients and is one of the limitations. However, the most important comparisons and findings with the main figures (elevated IL-8, IL-18, nCMI in male patients, elevated T cell responses in female patients, and comparisons between stabilized/deteriorated group within patients) are on the comparisons between male and female patients, whose age/BMI are well-matched (Extended Data Fig. 1a and Extended Data Table 1). Therefore, we do not think that this limitation fundamentally undermines our main conclusions.

However, in order to address this comment, we performed an additional analysis for Cohort A, adjusting age and BMIs, and the data are now presented as Extended Data Table 3. As expected from the very small sample sizes for these statistical adjustments, many of the differences seen with the raw data analyses in main figures are not significant (difference-in-difference). Nonetheless, the directions of differences and their magnitudes generally coincide with the raw data analysis, and even more, we still see statistically significant sex differences in IL-8 and CXCL10, which are known to be critical cytokine and chemokine in this disease, supporting our raw data analyses. We added some descriptions on the findings in Extended Data Table 3 in the Results in parallel with the description on raw data analyses, and also added the following sentence in the “Overview of the study (baseline analysis)”
“However, there was significant differences in age and BMI between HCW controls and patients (patients had higher ages and BMIs, Extended Data Table 1), and age- and BMI-adjusted analysis was also performed in parallel (Extended Data Table 3).”

And as the general remark of the limitation of this study, we added this paragraph in the Discussion:

“It is important to note that there are some limitations to the analyses presented in this manuscript. The healthy healthcare worker controls were not matched to patients based on age, BMI or underlying risk factors. To account for this, we performed adjusted analyses for the baseline and longitudinal comparisons between patients (Cohort A and the full patient population, Cohort B) and healthcare workers, controlling for age and BMI. However, there may still be residual confounding; information on underlying risk factors was not available for the healthcare worker controls.”

b. For many analyses (e.g., cytokines and immune cells), changes in responses or cell counts could just be reflective of the older age or greater BMI of patients compared with controls as both of these factors contribute to a greater baseline inflammatory state. Interestingly, both of these factors explained a lot of variability when comparing stabilized to deteriorated patients, but were never considered in the comparisons of patients vs. controls. For the T cell counts that appear to be reduced in patients relative to controls (Fig 2b), for example, it could be that aging rather than COVID-19 is explaining these differences.

Male patient and female patient groups themselves are completely age-and BMI-matched (Extended Data Table 1, Extended Data Fig. 1a). Therefore, our analyses on the differences inside the patient group are valid.

With respect to the T cell counts, both in Cohort A (Extended Data Table 3) and Cohort B (Extended Data Table 4), T cells are decreased in patients compared to HCWs following age- and BMI-adjusted analyses.

c. Lacking discussion of discordance in findings between Cohorts A and B; Cohort B included more severe patients that Cohort A – this should be acknowledged and could explain some of the discordance. It also should be made clear when data from one or both Cohorts are included in analyses.
We added following description in “Overview of Study design, longitudinal analysis”:

“Since Cohort B included patients with more severe patients in ICU, the clinical scores were on average higher in Cohort B compared to Cohort A (mean ± SD: 1.3 ± 0.5 (female) and 1.4 ± 0.5 (male) in Cohort A and 2.5 ± 1.5 (female) and 2.7 ± 1.3 (male) in Cohort B, Extended Data Table 1).”

Also, in every titles of the legends of figures and tables, we further specified the cohort characteristics.

d. Lines 464 – 466: dropped values in top and bottom 1st percentile. Needs to be made explicitly clear how many values were dropped for each cytokine (how many of the 38 patients?). In Figure 1 legend, the sample size indicated suggests it was the same for all cytokines, but this contradicts what is written in the methods. May also want to include in supplemental how this affected analysis – were findings very different if the outliers were included?

We re-examined the ELISA data, for which the outlier management was applied in the original manuscript. In this re-examination, we found that one patient and one HCW had outlier values (beyond 1.5x interquartile range) in more than half of the 71 cytokines/chemokines measured. We think this was likely due to poor sample quality, and decided to exclude these two samples. And upon the exclusion of these samples, we decided not to use outlier exclusion in ELISA or in any other assay to unify the handling of the data.

e. In many cases (e.g., Figure 1) samples sizes don’t add up (i.e., counting dots vs the sample size lists in the figure legend).

Not every patient has all types of samples, and for Cohort A, the data types available for each patient could be found in Extended Data Table 2.

To clearly indicate this point, but we added the following in the main text:

“For both baseline analysis and longitudinal analysis, types of samples available from each patient and each time point were variable, and information on the sample size (n) for each assay is shown in the respective figure legends or tables. For Cohort A, sample types obtained from each patient at the first time point can be found in Extended Data Table 3.”
We also re-checked that the sample size information in each figure legend and corrected if there was any mistake. (Please note that during this revision period a couple of mistakes/errors were found in our database and there are small changes in Ns from the previous manuscript)

f. For all measures, what was the time point when samples were collected and was that based on time since PCR+ test, time since hospital admission, or something else?

All the sample data used in this study and their time-point information defined as DFSO (days from symptom onset) are included in the Extended Data Table 2 (Cohort A first time point) and Supplemental Information Table (all samples, all time points). Further, we summarized and graphed DFSO for Cohort A first time point samples in Extended Data Fig. 1a.

2. Statistical analyses.
   a. Many comparisons of male HCW vs male patients and of female HCW vs female patients, and then looking at “difference of differences” A more interesting comparison would be male vs female HCW and male v female patients, and then, potentially, the difference in differences.

To address this comment, we added male vs. female HCW and male vs. female patient comparisons to the analyses and the results are incorporated in Extended Data Table 3 and 5. (Please note that BMI is now also included to the model and adjusted in addition to the age)

b. Figure 4a: I think all these figures would benefit from a third panel that is not sex-disaggregated, or some other way to account for differences between males and females in the study. A study of this size is very vulnerable to groups being unequal at baseline (e.g.: males being older than females), and I see no evidence that is accounted for in these comparisons.

According to this comment, we added sex-aggregated panels in Fig. 4a.

In order to emphasize that there is no difference in terms of ages, BMIs, or disease phase (DFSO at the first sampling) in Cohort A patients between sexes, we added the panels to show that there is no difference between male and female patients (Extended Data Fig. 1a).
c. Extended data table 6: lacking comparison between males and females within each disease state, as opposed to disease state within each sex.

We have now added the comparison between males and females within each disease state to our analyses, and added the results to the Extended Data Table 6.

d. Used propensity scores to adjust for age as a confounder (lines 480 – 481), but given the importance of age as a risk factor for disease, I would think you want this as a main effect in the model, and potentially an interaction term between age and sex.

Age was initially addressed as a confounder and included in the propensity score because we are primarily interested in isolating sex differences in immune response, not necessarily differences based on age. However, per this suggestion we did run the models with an interaction term. The results are shown as the Table d below. There is generally not statistically significant interaction between age and sex, which combined with the fact that our focus is not on the relationship between age and immune response, we decided not to include an interaction term.

[Redacted]

e. For analyses of viral loads, how can viral loads be detected below the limit of detection and then how can those values be included in the statistical analyses?

We put “0” values for the ND samples below the detection limit and used for statistical analyses. We included this information to the Method.

f. Uniform management of outliers as there are cases where outliers were removed (see above) and in other cases (e.g., Fig 3d and f) they are included and driving the effect.

In this revised manuscript, we decided not to do outlier exclusion in any of the data presented, as described in the response to comment d above.
3. Overinterpretation.

a. Lines 165 – 167 “innate inflammatory cytokines and chemokines are more robustly elevated early and throughout disease course in male over female patients”. From my understanding, they measured 71 cytokines and found significantly higher levels in males compared to females for one in each cohort (IL-8 in cohort A and CCL5 in cohort B). I don’t think this supports the conclusion above.

We modified the expression as follows:

“These data indicated that, while levels of the most of the innate inflammatory cytokines and chemokines are comparable, there are a few exceptions that are more robustly elevated at the baseline (IL-8 and IL-18) and during disease course (CCL5) in male patients over female patients.”

b. Classification as stable vs deteriorated (lines 212 -216). I could not find any explanation of when enrollment and the baseline sampling occurred relative to hospital admission. Was there an attempt to control for differences in disease severity at enrollment? The disease process of deteriorating from score 1 to 2 may not be the same as from 3 to 4.

Upon admission to the hospital, the patients were clinically tested with PCR and only the patients tested positive were eligible for this study. Then the informed consent for this study was obtained, the patient was enrolled in the study, and the first time point samples were collected.

In most of the cases, the first time point samples were collected 2-7 days after the hospital admission.

Information on the timing of hospitalization and collection of samples (all patients, all samples, including multiple time point samples) as defined with days from symptom onset (DFSOs) can be found in Supplementary Information Table 1.

About the "deteriorated" group, all the patients in this group got at least score 3 as their maximum scores. (> Oxygen 3L/min (oxygen mask), or CRP > 70). Patients remained between score 1 to 2 are classified as stable group. To clarify this point, we revised the text as described in the response to the last comment from Reviewer #2.

Original: The patient was categorized into the deteriorated group if the patient marked a worse, or higher, clinical score at any point compared to the patient’s initial clinical score (see Extended Data Table 3. Cmax > C1), and were otherwise categorized as stabilized.
We have modified this sentence to clarify our classification better, and also added some descriptions about the above points as follows:

Revised: The clinical scores at the first sample collection (C1) were 1 or 2 for all of the Cohort A patients. The patient was categorized into the “deteriorated” group if the patient marked a score of 3 or higher after the first sample collection date as their maximum clinical scores during admission (Cmax), and if the patient maintained the score of 1 or 2, they categorized as stabilized (see Extended Data Table 2 for the detailed information for each patient). Both in male (N=17) and female (N = 22) Cohort A patients, 6 patients in each group deteriorated in their disease course (35.3% and 27.3%, respectively), and the intervals between the dates on which the patients reached Cmax (DFSO at Cmax) and the first sample collection (DFSO at C1) were not significantly different between deteriorated male patients and female patients (mean ± SD = 3.7 ± 4.1 and 4.2 ± 2.7, respectively. p = 0.81).

c. Figure 2c has a representative FACS image for Figure 2d. However, the FACS image does not appear representative of the monocyte graph in Figure 2d. The percentages of ncMono in male and female HCW are 7.82 and 2.43, respectively, but the graph shows that the mean % ncMono of male and female HCW is similar. The percentage of cMono in HCW is about 60% in both males and females, which is about 2-fold higher than cMono % in patients in Figure 2c. However, Figure 2d shows that % cMono is slightly higher in the patients compared to HCW.

This discrepancy is simply because the numbers indicated in the panel are the % in the parent gate, but not % in the live cells as shown in Fig 2d. We added a description in the figure legend that these numbers are % of parent gate.

d. In the abstract, the authors state that the plasma level of CCL-5 along with IL-8 and IL-18 is higher in male patients compared to female patients. However, Figure 4b and Extended Data Figure 2 clearly show that the plasma level of CCL-5 in male patients is not higher than that in female patients. Extended Data Table 4 shows that there is a significant difference of CCL-5 between male and female patients in Cohort B. In the same manner, IL-8 level did not show any difference between male and female patients in Cohort B. However, Figure 1d shows that IL-8 level in male patients is significantly higher than that in female patients in Cohort A. The authors appear cherry-picked the data in Cohorts A and B to support their theory.

Indeed, this part is the remarks on the baseline analysis, so we deleted description on CCL5 from here and modified as follows:

"By focusing our analysis on patients with moderate disease who had not received immunomodulatory medications, our results revealed that male patients had higher plasma
levels of innate immune cytokines such as IL-8 and IL-18 along with more robust induction of non-classical monocytes.”

4. Cohorts.
a. The authors have divided up patients into two cohorts based on whether or not they were baseline versus longitudinal or whether they had not received antiviral or steroidal interventions. A third cohort may be useful in directly comparing patients who were and were not treated with antivirals and corticosteroids at baseline, just to see how this affected viral loads and innate immune responses.

We agree the analysis of this cohort is definitely interesting, but we also think that these analyses would not be very relevant to the sex difference. In Cohort A, none of these patients were on these interventions, and in Cohort B, almost exactly same proportion of patients in male group and female group have undergone these treatments (Extended Data Table 1).

This means that our two cohorts are well controlled in terms of these interventions upon the analysis of the sex differences.

b. It is unclear how these “mild” patients are being considered mild since they are still admitted to the hospital. The authors would simplify this by comparing their own clinical scoring to that of the WHO for COVID19 disease severity, in order to give the audience a better reference of clinical diagnosis.

All patients necessitated hospitalization for their symptoms (Patients were admitted at the discretion of the emergency room doctor who evaluates the patient. WHO score ≥ 3 (=Hospitalized Mild disease, WHO COVID-19 Therapeutic Trial Synopsis, February 18, 2020)).

Indeed, although there were patients who had “Hospitalized Mild disease”, the usage of “mild here might be misleading as you pointed out. Therefore, we modified the expression “mild and moderate” to “moderate” in the revised text.

In terms of the correspondence between our clinical score and WHO score, there 2 major differences:

1) Tocilizumab usage is not relevant in WHO staging, but we include it for staging.

So our score 2 and 3 would go into WHO 4.
2) In our categorization, we did not delineate between non-invasive ventilation and mechanical ventilation. So our score 5 corresponds to WHO 6 and 7.

| WHO score | Our score |
|-----------|-----------|
| 0         | x         |
| 1         | x         |
| 2         | x         |
| 3         | 1         |
| 4         | 2,3       |
| 5         | 4         |
| 6         | 5         |
| 7         | 5         |
| 8         | x         |

We now added some description about the comparison between WHO scoring and our scoring system in the Method.

c. The cohort used should be referred to in each figure legend.

In all of the figures (both in main and extended), Cohort A is used. We referred to this point in the titles of every figure legends.

Minor:
1. Incorporate tables, including at least one demographic table comparing males and females in each cohort into the main text

Due to the significant space constraint of Nature format, we are unable to include the demographic table in the main figures. However, we added some additional remarks on demographics or timing of sample collection (DFSO) about patients as described in the responses to other comments, and also added some panels in Extended Data Fig. 1a about Cohort A.
2. Need more information on what “first” time point means – is this at hospital admission? Any control or acknowledgement that people may have been at very different points in disease progression at “baseline”? Unclear if or how this was controlled for.

a. Some of this information is included in extended table 3, but nowhere in main body of the text.

Upon hospital admission, the patients were clinically tested with PCR and only the patients tested positive were eligible for this study. Then the informed consent for this study was obtained, enrolled in the study, and the first time point samples were collected. In most of the cases, the first time point samples were collected 2-7 days after the admission. (Please also refer to the response to the first comment of Reviewer #1).

The first time point samples were collected on average 11 days after the symptom onset, and well-matched between sexes (information found in Extended Data Table 1). We added following description in the Method, “Patients and HCWs” section:

“All patients necessitated hospitalization for their symptoms and had WHO score ≥ 3 at admission (=Hospitalized Mild disease, WHO COVID-19 Therapeutic Trial Synopsis, February 18, 2020). At the initial screening, clinical PCR tests were performed in CLIA-certified laboratory and only the PCR-positive patients were enrolled. Only after the confirmation of PCR-positivity, the patients were enrolled and the first time point samples for this study were collected for each patient. The first time point samples were collected at 11.4 ± 8.1, 10.2 ± 6.3, 11.7 ± 7.2, and 12.1 ± 7.3 (mean ± SD) days after the symptom onset (DFSO) in Cohort A female, Cohort A male, Cohort B female, and Cohort B male, respectively (Extended Data Fig 1a right panel for Cohort A and Extended Data Table 1).”

3. Description of longitudinal analysis is not clear – needs some detail about how many time points per person and over what length of time (days? Weeks?)

The numbers of time points were variable between patients, and the information is included in Extended Data Table 1. The intervals between each time point were in general 3 to 7 days, but sometimes more than 7 days. The information on the exact time points for all the patient samples used in this study can be found in the Supplementary Information Table 1 as defined with days from symptom onset (DFSO).

4. Number of samples taken per subject is included in extended data table 1, but no information about timing of samples is included.
included the detailed information (DFSO for all the sample collection) in the Supplemental Information Table 1.

5. For each of the heatmaps, the “color key” seems to be missing color, as it is showing the scale for Z scores but there is no color in the boxes. I am assuming this is meant to show color intensity and that there was just an error in copy/pasting or formatting.

In the manuscript we submitted, the color was not missing. This is probably simply due to technical trouble with Acrobat. We have contacted the Nature editorial team to ensure that all colors are transmitted as the submitted version.

6. These heatmaps show quite a bit of heterogeneity and no clear pattern and may be better suited to supplementary figures and replaced with some other graphs from supplemental.

We moved all the heatmaps to the Extended Data Fig. 2.

7. The limit of detection or cutoff value is missing from the antibody graphs (Figure 1b, 4a) as well as how that limit was set.

We have now evaluated the cutoff values for positivity and are 0.392 and 0.436 for anti-S1-IgG and anti-S1 IgM, respectively. Eighty pre-pandemic plasma samples were assayed to establish the negative baseline. The cutoff values were statistically determined with confidence level of 99%. We included this information in the Method, and also added dotted line in the figures for anti-S1-IgG/IgM (Figure 1b and 4a, 4b). This information is now included in the Method.

8. Why is the race and ethnicity information missing from all controls?

In this revised version, the ethnicity information for HCW is included to the demographic tables. (Extended Data Table 1)

Reviewer Reports on the First Revision:

Referees' comments:

Referee #1 (Remarks to the Author):

I find the revised manuscript much improved and the additional analyses have convinced me that the author’s conclusions are robust and provide important novel insights.
Referee #2 (Remarks to the Author):

The authors have done a good job of addressing the reviewer concerns.
No further comments.

Referee #3:

UW: The reviewer has not supplied remarks to the author.

The reviewer notes in comments to the editor that previous points have not been addressed, and remains concerned that the authors compare COVID-19 patients with ‘healthy’ control health care workers who were around 20 years younger and with a significantly lower BMI.
The reviewer notes that when the authors controlled for the differences in BMI and age between their controls and COVID-19 patients, the sex differences went away.
The reviewer notes furthermore that in most cases, the authors do not have statistically significant differences between male and female COVID-19 patients, aside from a few differences in some cell populations and chemokines.

Referee #4 (Remarks to the Author):

Overall, the data is clearly important and very well described, but possibly overinterpreted (see my line 170 comment below).

Re: reviewer 3, I think this reviewer would have been appeased if they had directly adjusted for potential confounders instead of using a propensity score approach. Philosophically, propensity scores don’t make much sense because sex at birth is immutable. You don’t have a propensity to be born female, unlike having a propensity to smoke, for example. Reviewer 3 essentially asked for this in comment 2d, but it is impossible to tell what the authors did from the response. They could have left the propensity score and just include the age interaction. But it would be good to know why they chose propensity score adjustment versus inclusion of the confounders as covariates and how would the results have changed if they just included the confounders as covariates in the regression model.

I think they could reword some things to be less causally strong. For example, “Male patients have higher levels of key innate immune cytokines” could be “Higher levels of … were observed in male patients.”
Or “Female COVID-19 patients induce more robust T cell response than male patients” is a particularly strong causal statement. I think even if it was put in past tense, it would be better. Or “A more robust T cell response was observed in female COVID-19 patients”. I understand wanting to avoid passive tense, but implying causality from a small sample size should also be avoided.

Other comments:
Line 170 This comparison is the difference between male patients to male controls and female patients to female controls. It is difficult to interpret given that the HCWs may not be an appropriate control group. So while the differences between males and females are statistically significant unadjusted (fig 1c). This goes away once adjusted and the statistically significant difference is the interaction between sex and HCW/patient (assuming this is a multiplicity adjusted p-value). So the finding really is that the difference between male patients and male controls is larger than the difference between female patients and female controls. I think this is the heart of what the reviewer is getting at but seems to be glossed over in the manuscript. So, if the difference is no longer significant after adjustment, is it a power issue or is the observed sex
difference due to something else (the other requested analyses)? I can’t really tell the answer to that, but it is probably in there somewhere. This investigation wasn’t really designed for that question. It was designed to generate hypotheses and guide future investigations (the authors may disagree with this statement).

Line 608: this is not a multivariate trajectory analysis. Perhaps they are saying this because it has multiple covariates. I would remove the word multivariate.

Pvalues: I’m not sure what Nature’s standards are on this, but many medical journals would not accept asterisks for some p-values and stating the p-value for others. The authors like showing the exact value when .05<p<.1 as if this is somehow interpretable.

Author Rebuttals to First Revision:

Referee #4:

There is a lot of data to digest in this manuscript, so I will stick to a careful examination of your question below while taking in the manuscript as a whole. Overall, I wish they had included a statistician in their team. I think much of the statistical methods would be more clear if they had. There is really clunky language used to describe their statistical methods and results. Overall, the data is clearly important and very well described, but possibly overinterpreted (see my line 170 comment below).

We are grateful for the constructive feedback and have updated the manuscript to reflect the reviewer’s comments. We hope that our updates will provide clarity on the methods and results and allow for clear interpretation of these important findings.

Re: reviewer 3, I think this reviewer would have been appeased if they had directly adjusted for potential confounders instead of using a propensity score approach. Philosophically, propensity scores don’t make much sense because sex at birth is immutable. You don’t have a propensity to be born female, unlike having a propensity to smoke, for example. Reviewer 3 essentially asked for this in comment 2d, but it is impossible to tell what the authors did from the response. They could have left the propensity score and just include the age interaction. But it would be good to know why they chose propensity score adjustment versus inclusion of the confounders as covariates and how would the results have changed if they just included the confounders as covariates in the regression model.

We have updated the results and the manuscript to reflect the reviewer’s request to include the confounders as covariates in the models rather than using a propensity score for covariate adjustment. Please see the revised Extended Data Table 4 and the corresponding methods (Lines 662-676). The effect estimates and precision are comparable for both methods.

I think they could reword some things to be less causally strong. For example, “Male patients have higher levels of key innate immune cytokines” could be “Higher levels of ... were observed in male patients.”

Or “Female COVID-19 patients induce more robust T cell response than male
patients” is a particularly strong causal statement. I think even if it was put in past tense, it would be better. Or “A more robust T cell response was observed in female COVID-19 patients”. I understand wanting to avoid passive tense, but implying causality from a small sample size should also be avoided.

We agree with the reviewer’s assessment of potential over-interpretation or inaccurately causal language in some portions of the manuscript. While we strongly believe that the findings of this study are important and provide a strong foundation for additional investigation into the relationship between sex and immune response in COVID-19 infection, this study and the accompanying analyses were designed to an initial assessment of these hypotheses. We have carefully reviewed the manuscript and updated causal language as well as updated the limitations section of our discussion to reflect this (Results and Discussion).

Other comments:
Line 170 This comparison is the difference between male patients to male controls and female patients to female controls. It is difficult to interpret given that the HCWs may not be an appropriate control group. So while the differences between males and females are statistically significant unadjusted (fig 1c). This goes away once adjusted and the statistically significant difference is the interaction between sex and HCW/patient (assuming this is a multiplicity adjusted p-value). So the finding really is that the difference between male patients and male controls is larger than the difference between female patients and female controls. I think this is the heart of what the reviewer is getting at but seems to be glossed over in the manuscript. So, if the difference is no longer significant after adjustment, is it a power issue or is the observed sex difference due to something else (the other requested analyses)? I can’t really tell the answer to that, but it is probably in there somewhere. This investigation wasn’t really designed for that question. It was designed to generate hypotheses and guide future investigations (the authors may disagree with this statement).

In regards to the results summarized in Line 170, thank you for calling the lack of clarity in the presentation and interpretation of these results to our attention. We have updated the text at Line 170 and throughout the manuscript to clarify the interpretation of the direct female-to-male patient comparisons versus the difference-in-differences comparisons (Lines 180-195).

The reviewer is correct in that differences between male and female patients in IL-8 and IL-18, for example, are significant in unadjusted analyses, but not statistically significant in the adjusted analyses comparing male and female patients in Cohort A (Extended Data Table 3). However, if you examine the magnitude and direction of the differences (the first column of Extended Data Table 3), they are very similar between the unadjusted and adjusted analyses for Cohort A. The additional information provided by the difference-in-difference analysis demonstrating that IL-8, for example, appears to be significantly more elevated in male patients (compared to male healthcare workers) than in female patients (compared to female healthcare workers) informed our conclusions regarding the potential role of IL-8 at baseline.

As mentioned above and by the reviewer, we have now included language indicating that these analyses were designed to be initial explorations of these hypotheses and
are impacted by the small sample size and subsequent lack of power. Additionally, we agree that for the comparisons between patients and healthcare workers explored in this analysis there is the possibility for residual confounding (as now mentioned in our discussion).

Line 608: this is not a multivariate trajectory analysis. Perhaps they are saying this because it has multiple covariates. I would remove the word multivariate.

Thank you for bringing this to our attention. The intention was to write ‘multivariable’ not multivariate and the manuscript has been updated accordingly.

P-values: I’m not sure what Nature’s standards are on this, but many medical journals would not accept asterisks for some p-values and stating the p-value for others. The authors like showing the exact value when .05<p<.1 as if this is somehow interpretable.

Since there were some apparent instances that the difference did not reach significance likely due to the small cohort size for Cohort A (for example, Fig 4a. Age difference between “deteriorated” and “stable” patients, p=0.051), we think it might be beneficial to show the “tendency”, too, with p-value between 0.05 and 0.1.

In this revision, we deleted all the asterisks and included all the raw p-values < 0.1 in all the main and extended data figures.