PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (**http://bmjopen.bmj.com/site/about/resources/checklist.pdf**) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Understanding COVID-19 reporting behavior to support political decision-making: A retrospective cross-sectional study of COVID-19 data reported to the World Health Organization
AUTHORS	Abbood, Auss; Ullrich, Alexander; Denkel, Luisa

VERSION 1 – REVIEW

REVIEWER	Leng, AnLi Shandong University
REVIEW RETURNED	31-Mar-2022

GENERAL COMMENTS	This paper provides a new overview of behavior of countries based on COVID-19 case count data submitted to the World Health Organization. The study was meaningful. However, I have some questions. Methods: This manuscript introduced a score to quantify the discrepancy between reporting behavior and epidemiological situation in the Reporting Score Section. Please provide more evidence to prove that the method is scientific and reasonable. Discussion: This study showed a “weekend effect” for most countries.Please discuss the reasons in the Discussion Section. Besides, please provide some evidences or references for the conclusion “COVID-19 case counts reported in the middle of the week seemed more reliable and should provide the basis of decision-making”.

REVIEWER	Turati, Gilberto Università Cattolica del Sacro Cuore, of Economics and Finance
REVIEW RETURNED	24-Apr-2022

GENERAL COMMENTS	Review on “Understanding COVID-19 reporting behavior to support political decision-making: An analysis of COVID-19 data reported to the world health organization” The aim of this paper is to explore the reporting behaviour of world countries with respect to the Covid-19 pandemic. To this end, the author use case counts provided by the WHO and relative to 236 “countries, territories, or areas” from Jan 3rd 2020 to June 14th 2021. They then develop different measures of reporting behaviour (related to the frequency of reporting) to be compared with incidence rates. They also consider a cluster analysis to group different countries according to their reporting behaviour and their incidence rate. Conclusions suggest that most countries provide high quality data given the consistency between incidence rates

	and frequency of reporting. In addition, case counts reported in the middle of the week are likely to provide a better representation of current incidence than data provided at the end of the week. As a general observation, I think that studying the reporting behaviour of countries is very important from a policy perspective. The importance is not limited to define travel-restrictions across countries on more solid basis as the authors seem to suggest. It is also a matter of defining homogeneous data collection systems allowing country comparisons. In many countries, the policy discussion and the evaluation of government performance in tackling the pandemic has been based exactly on country comparisons. Many studies on the effectiveness of restriction policies are based on these data; poor input data will likely result in poor conclusions for policy making. This general observation has two effects on the paper as it currently stands: first, the motivation is poor and needs to be improved making reference to studies using WHO data and studies discussing the appropriate governance for the pandemic; there are only five references quoted by the authors, which is somewhat surprising in my view. Challenges relative to the global governance of health systems from the perspective of economists are discussed in a very simple way in, e.g., the book by Costa-Font, Turati and Batinti (2020). Odone, Delmonte, Scognamiglio, and Signorelli (2020) provides a discussion of Covid-19 data and the challenges they pose from the perspective of epidemiologists. But these are just two examples out of many by now. A brief discussion of this literature should find its way in the introduction of the paper to motivate the analysis by the authors. An additional general issue concerns the main idea of the paper of using the frequency in reporting behaviour as a way to assess the quality of data. Why frequency should be related to data quality is never spelled out clearly in the paper. A possible story that came to my mind is that lower frequency can imply a better control and verification of the data, which implies better data. This is independent on the number of cases and fatalities and it relates to the checks and effort that authorities in charge of collecting data provide before communicating the data. But this story is the opposite of what the authors are pushing in the paper. This lack of conceptualization can somewhat explain the lack of comments for the results of the statistical analysis, which remains mostly descriptive. Further to this, it would be really important to discuss the data generating process and the standards adopted to count Covid-19 cases (and fatalities). How do countries have to report data to the WHO? Are there defined standards to be applied? How does it work within countries? More specific issues: 1) “Yet, epidemiological indicators used - including 7-day incidence rate – loses robustness in case of strongly irregular reporting behaviour”: can you provide examples of how irregular reporting behaviour by countries might affect decisions? 2) “7-day incidence rates were scaled to a range between 0 and 1 to make them comparable to the binary reporting rate”: the need for rescaling can be gauged later when defining the reporting score. I would anticipate the definition of the score. 3) “the reporting score discriminates against countries with no/few COVID-19 cases since this leads to a binary reporting rate of or close to 0. For such countries, we applied an imputation on the reporting rate”: I am always doubtful with respect to this procedures, especially when robustness checks are not provided

	in the text. One way to circumvent the problem would be to provide results excluding these countries with no/few Covid-19 cases. 4) Robustness tests should be provided also with respect to the different waves of the pandemic. One can easily think that from Jan 3rd 2020 to June 14th 2021 reporting behaviour has changed: one can expect that – during the first wave – uncertainty in counting the number of cases was higher than during the second wave. 5) "When excluding tiny island states with almost no COVID-19 cases, we identified the highest scores for China (0.98), Egypt (0.97) and Greenland (0.97), while the lowest scores were found in Andorra (-0.15), San Marino (-0.07), and Seychelles (0.05).": Andorra, San Marino and Seychelles are also tiny states. 6) "A correlation of the population size and reporting score on the country-level can be found in Error! Reference source not found.. The Spearman rank correlation coefficient for all countries is 0.28.": there is any comments on the correlation and it is unclear what we can learn from here. Notice that all references are not reported correctly in the manuscript. 7) "Using this count, we created a histogram visualizing that most countries (close to 140) have no (zero) weekdays with a binary reporting rate below 50%. Interestingly, the second most frequent group of countries had a binary reporting rate of less than 50% for all weekdays": interesting, but what do we learn from here? 8) Figures are also poor and can be improved (too many lines in the Cluster pictures make these hard to read, despite the red line). It would be nice to see graphically the results of the cluster analysis. References Costa Font J., Turati G., Batinti A. (2020), The political economy of health and healthcare, Cambrdige University Press Odone A., D. Delmonte, T. Scognamiglio, C. Signorelli (2020), COVID-19 deaths in Lombardy, Italy: data in context, The Lancet Public Health, 5(6), e310

VERSION 1 – AUTHOR RESPONSE

Reviewer Reports:

Reviewer: 1
Dr. AnLi Leng, Shandong University

Comments to the Author:
This paper provides a new overview of behavior of countries based on COVID-19 case count data submitted to the World Health Organization. The study was meaningful. However, I have some questions.

Methods: This manuscript introduced a score to quantify the discrepancy between reporting behavior and epidemiological situation in the Reporting Score Section. Please provide more evidence to prove that the method is scientific and reasonable.

We agree that we didn't make it clear enough how we came up with this score. Therefore, we added an example and a more thorough explanation to the methods section. Herein, we explained how we found a discrepancy between binary reporting rate and incidence during data inspection which

inspired the score. We added an example in Appendix 9. However, to our knowledge, scientific evidence to support our approach is scarce. Publications[i,ii] with scores reported that proof their validity (proper scoring rules) are not transferable to our research question as we are not quantifying any probabilities.

Discussion: This study showed a “weekend effect” for most countries. Please discuss the reasons in the Discussion Section. Besides, please provide some evidences or references for the conclusion “COVID-19 case counts reported in the middle of the week seemed more reliable and should provide the basis of decision-making”.

We added references to define and discuss the “weekend effect” in more detail. Further, we added references that support our conclusion. Please find the revised sections in the discussion.

Reviewer: 2
Gilberto Turati, Università Cattolica del Sacro Cuore

Comments to the Author:
Review on “Understanding COVID-19 reporting behavior to support political decision-making: An analysis of COVID-19 data reported to the world health organization”

The aim of this paper is to explore the reporting behaviour of world countries with respect to the Covid-19 pandemic. To this end, the author use case counts provided by the WHO and relative to 236 “countries, territories, or areas” from Jan 3rd 2020 to June 14th 2021. They then develop different measures of reporting behaviour (related to the frequency of reporting) to be compared with incidence rates. They also consider a cluster analysis to group different countries according to their reporting behaviour and their incidence rate. Conclusions suggest that most countries provide high quality data given the consistency between incidence rates and frequency of reporting. In addition, case counts reported in the middle of the week are likely to provide a better representation of current incidence than data provided at the end of the week.
As a general observation, I think that studying the reporting behaviour of countries is very important from a policy perspective. The importance is not limited to define travel-restrictions across countries on more solid basis as the authors seem to suggest. It is also a matter of defining homogeneous data collection systems allowing country comparisons. In many countries, the policy discussion and the evaluation of government performance in tackling the pandemic has been based exactly on country comparisons. Many studies on the effectiveness of restriction policies are based on these data; poor input data will likely result in poor conclusions for policy making.
This general observation has two effects on the paper as it currently stands: first, the motivation is poor and needs to be improved making reference to studies using WHO data and studies discussing the appropriate governance for the pandemic; there are only five references quoted by the authors, which is somewhat surprising in my view. Challenges relative to the global governance of health systems from the perspective of economists are discussed in a very simple way in, e.g., the book by Costa-Font, Turati and Batinti (2020). Odone, Delmonte, Scognamiglio, and Signorelli (2020) provides a discussion of Covid-19 data and the challenges they pose from the perspective of epidemiologists. But these are just two examples out of many by now. A brief discussion of this literature should find its way in the introduction of the paper to motivate the analysis by the authors.

Thank you for these important comments. We extended the motivation of this study from the perspective of travel restrictions only to other public health measures. We revised and extended the introduction accordingly. Further, we revised the discussion (as also suggested by the editor), explained our findings in the context of background literature, re-structured the discussion and added more references (including Odone et al. 2020). We hope that we could resolve your concerns.

An additional general issue concerns the main idea of the paper of using the frequency in reporting behaviour as a way to assess the quality of data. Why frequency should be related to data quality is never spelled out clearly in the paper. A possible story that came to my mind is that lower frequency can imply a better control and verification of the data, which implies better data. This is independent on the number of cases and fatalities and it relates to the checks and effort that authorities in charge of collecting data provide before communicating the data. But this story is the opposite of what the authors are pushing in the paper. This lack of conceptualization can somewhat explain the lack of comments for the results of the statistical analysis, which remains mostly descriptive.

We added this important issue to limitations in the discussion:
In this study, we interpret the frequency of reporting as a surrogate for accuracy and reliability of data. However, for some countries, this assumption might be incorrect. Less frequent reporting could also be an indicator of high data quality, as counts might be very thoroughly validated before reporting. However, WHO encourages its member states to report case counts daily and countries may revise incorrect reporting retrospectively[iii]. Thus, we assume our assumption eligible for most countries.

Further to this, it would be really important to discuss the data generating process and the standards adopted to count Covid-19 cases (and fatalities). How do countries have to report data to the WHO? Are there defined standards to be applied? How does it work within countries?

According to WHO, since 22 March 2020, global data are compiled through WHO region-specific dashboards (with count data provided by countries), and/or aggregate count data reported to WHO headquarters daily. Counts primarily represent laboratory-confirmed cases and deaths as defined by WHO case definitions. However, some differences may exist due to local adaptations. We included this paragraph to the discussion (limitations).

More specific issues:
1) “Yet, epidemiological indicators used - including 7-day incidence rate – loses robustness in case of strongly irregular reporting behaviour”: can you provide examples of how irregular reporting behaviour by countries might affect decisions?

Several countries including Germany used 7-day-incidence rates as one indicator of implementing travel restrictions[iv]. This represents one example for which reporting lags or gaps might have an impact on political decisions.

2) “7-day incidence rates were scaled to a range between 0 and 1 to make them comparable to the binary reporting rate”: the need for rescaling can be gauged later when defining the reporting score. I would anticipate the definition of the score.

We added the definition of the reporting score to the methods section: “[…] the reporting score is defined as country-specific measure within the range of -1 and +1 that is based on the normalized means of the 7 days incidence rates and binary reporting behaviors.”
We consider our approach suitable as scaling the 7 days-incidence rate before subtracting allows both metrics having similar weights in the score. We hope that we could dispel your concerns.

3) “the reporting score discriminates against countries with no/few COVID-19 cases since this leads to a binary reporting rate of or close to 0. For such countries, we applied an imputation on the reporting rate”: I am always doubtful with respect to this procedures, especially when robustness checks are not provided in the text. One way to circumvent the problem would be to provide results excluding these countries with no/few Covid-19 cases.

Thank you for this hint. We followed your advice and uploaded three plots: One without imputation (Appendix 7), one with imputation (Figure 1), and one with countries being dropped that could never obtain a good score due to no/few cases (Appendix 8). We added them as part of a sensitivity analyses. The results confirmed our imputation strategy being appropriate. The plots with imputation were more similar to the plots without imputation with only a few changes in the higher ranges of the score. At the same time, plots with dropped data showed stronger distortions strongly suggesting that we picked a useful imputation strategy.

4) Robustness tests should be provided also with respect to the different waves of the pandemic. One can easily think that from Jan 3rd 2020 to June 14th 2021 reporting behaviour has changed: one can expect that – during the first wave – uncertainty in counting the number of cases was higher than during the second wave.

This is a great advice. We plotted the global binary reporting rate over the experiment time frame (Appendix 10). The data suggests that reporting become more reliable over time. We added this finding to the results: “Plotting the global binary reporting rate over time suggests that reporting have become more reliable since April 2020 (Appendix 10). However, performing sensitivity analyses by

excluding early data (January – March 2020) did not fundamentally change our findings (Appendix 11).

5) “When excluding tiny island states with almost no COVID-19 cases, we identified the highest scores for China (0.98), Egypt (0.97) and Greenland (0.97), while the lowest scores were found in Andorra (-0.15), San Marino (-0.07), and Seychelles (0.05).”: Andorra, San Marino and Seychelles are also tiny states.

Thank you for this hint. We applied a clearer definition and simply excluded countries with less than 500k inhabitants and less then 100km² area which usually are small islands that had an advantage in keeping COVID out of their countries. We changed the results accordingly.

6) “A correlation of the population size and reporting score on the country-level can be found in Error! Reference source not found.. The Spearman rank correlation coefficient for all countries is 0.28.”: there is any comments on the correlation and it is unclear what we can learn from here. Notice that all references are not reported correctly in the manuscript.

All references were revised according to BMJ guidelines.

7) “Using this count, we created a histogram visualizing that most countries (close to 140) have no (zero) weekdays with a binary reporting rate below 50%. Interestingly, the second most frequent group of countries had a binary reporting rate of less than 50% for all weekdays”: interesting, but what do we learn from here?

Thank you for this important remark. We further clarified why we did this analysis and revised the manuscript accordingly. Briefly, we wanted to show that a majority of the countries (84%) could be described by two extreme groups, i.e., ones that report most of the time on all weekdays and those that hardly ever report on any weekday.

8) Figures are also poor and can be improved (too many lines in the Cluster pictures make these hard to read, despite the red line). It would be nice to see graphically the results of the cluster analysis.

We removed the grey lines to only show the average curve within the cluster. Additionally, we adjusted the y-axis scaling. We believe that comparing the clusters should be easier now. Finally, we added the number of countries that belong to a cluster for even more clarity.

References

Costa Font J., Turati G., Batinti A. (2020), The political economy of health and healthcare, Cambrdige University Press

Odone A., D. Delmonte, T. Scognamiglio, C. Signorelli (2020), COVID-19 deaths in Lombardy, Italy: data in context, The Lancet Public Health, 5(6), e310

Reviewer: 1
Competing interests of Reviewer: No.

Reviewer: 2
Competing interests of Reviewer: No competing interests.

Brier, G. W. (1950). Verification Of Forecasts Expressed In Terms Of Probability. Monthly Weather Review 78, 1, 1-3, available from: < https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2> [Accessed 04 June 2022]

Hyvärinen, A (2005). Estimation of Non-Normalized Statistical Models by Score Matching. Journal of Machine Learning Research 5, 24, available from < https://jmlr.org/papers/v6/hyvarinen05a.html> [Accessed 04 June 2022]

World Health Organization (2022). WHO Coronavirus (COVID-19) Dashboard - Data sources 2022, available from <https://covid19.who.int/data> [Accessed 13 May 2022]

Robert Koch Institute (2022). Information on the designation of international risk area 2022, updated 25/02/2022, available from <https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Transport/Archiv_Risikogebiete/Risikogebiete_2022-02-25_en.pdf?__blob=publicationFile> [Accessed 16 May 2022]

VERSION 2 – REVIEW

REVIEWER	Leng, AnLi Shandong University
REVIEW RETURNED	21-Jul-2022

GENERAL COMMENTS	The study was meaningful. However, I have some questions. This manuscript developed a score to quantify the discrepancy between reporting rate and the epidemiological situation. However, the evidence to prove that the method is scientific and reasonable was not sufficient. For example, has this method been applied and proven in other studies before? Also, as I mentioned before, this study showed a “weekend effect” for most countries. It is important finding. Please discuss the reasons and add the application measures for the political decision-making in the Discussion Section.

REVIEWER	Turati, Gilberto Università Cattolica del Sacro Cuore, of Economics and Finance
REVIEW RETURNED	25-Jul-2022

GENERAL COMMENTS	Review on “Understanding COVID-19 reporting behavior to support political decision-making: A retrospective cross-sectional study of COVID-19 data reported to the World Health Organization” I have now read the revised version R1 of the manuscript. I try to list here what are still open issues in my view. Goal of the paper. The current version is much less ambitious than the original version of this manuscript. The idea in the original version was about using the frequency in reporting behaviour as a way to assess the quality of data related to Covid-19 infections. Why frequency should be related to data quality was never spelled out clearly in the paper. And in my previous report I did report an obvious example that argued in favour of an opposite story: lower frequency can imply a better control and verification of the data, which implies better data; a behaviour which is independent on the number of cases and fatalities and it relates to the checks and effort that authorities in charge of collecting data provide before communicating the data. The goal in this version is a very simple descriptive exercise and an attempt at classifying different countries according to their “reporting behaviour”, accounting for reporting gaps (relative reporting frequency per weekday regardless of the number of reported cases) and reporting lags (share of cases reported each weekdays relative to the number of cases reported each week). My suggestion would be to better

	clarify this in the first rows of the introduction, avoiding any reference to the quality of the data. Unless the authors provide a consistent conceptual explanation, I do not see any way to infer quality of data from analysing reporting behaviour. This is true even in relation to the 7-day incidence rate, used by some countries to implement some policies. I do not see any comparison of the indicators on gaps and lags provided here with the 7-day incidence rate. Motivation. In my previous report, I have suggested the authors to improve the motivation of their paper. Even in the current version, the focus is on the use of Covid-19 WHO data for border closures and travel restrictions. I do believe the authors are missing at least one other important reason to discuss the reporting behaviour of countries: in many countries (all countries?), media and scholars have tried to evaluate the performance of their government in facing the pandemic, and the effectiveness of the harsh measures taken by their government, generating a hot political debate. Suppose I compare the lockdown imposed in Italy with the more liberal stance taken from the Swedish government, and data for Italy and Sweden miss out a number of infections but with different degrees of precision. How can I make any meaningful statement on the effectiveness of government interventions in both countries? This is the reason why it is of paramount important to define standards of collecting data at the world level: to favour cross-country comparisons and understand what works and what is not working. This is clearly not enough to avoid false reporting by countries of this strategic information. In my previous report, I have mentioned two specific issues to improve motivation: first, discuss what should be the appropriate governance for the pandemic. There is no point in making reference to WHO data if different countries do not believe about the benefits of centralizing information (and management of the pandemic) at the world level (and you can expect different types of country to behave differently in this respect). I have mentioned, as a reference to this specific issue of the global governance, the book Costa Font J., Turati G., Batinti A. (2020), The political economy of health and healthcare, Cambrdige University Press. This also to favour an interdisciplinary discussion on issues that should not be bounded in specific disciplines. Second, in stressing that this is not the first paper raising issues on Covid-19 data, I have suggested a brief discussion of the literature. In the motivation, it should be stressed what this paper does in addition to this literature. According to authors, this is the first paper analysing reporting behaviour by countries of Covid-19 cases to the WHO. This brings me to the importance of institutional details. Institutional background and different waves. As I mentioned in my previous report, it would be really important to discuss the data generating process and the standards adopted to count Covid-19 cases (and fatalities). I have posed to the authors some questions: (i) How do countries have to report data to the WHO? (ii) Are there defined standards to be applied? (iii) How does it work within countries, especially those with a federal structure? (iv) Have these standard been changed from the first to the second wave, as knowledge on the pandemic made some progress? The last question is particularly important to study reporting behaviour as it is clear that, in the first wave, the number of cases is severely underestimated for many reasons (different standards in reporting, ability to test, poor testing strategies, …). I do not see any answers to my questions in the revised version of the paper, but for few lines in the “Sensitivity analysis” section toward the end of the

	paper on how the reporting behaviour changed over time. From Appendix 10, it looks like binary reporting reacted swiftly to the onset of the pandemic. Why? When did WHO issued the first definition of a Covid-19 case for its surveillance system? The authors mention that the data are compiled using region-specific WHO dashboards since 22 March 2020 but they never really use this information. Takeaway. I have read conclusions of this paper many times. A fair summary of the takeaway messages is the following: (i) most of the countries report cases when they do have cases to report; (ii) there seems to be a weekend effect, hence counts reported in the middle of the week are more reliable. If there is any other strong message, it does not seem to emerge from the paper.

VERSION 2 – AUTHOR RESPONSE

Reviewer: 1
Dr. AnLi Leng, Shandong University

Comments to the Author:
The study was meaningful. However, I have some questions.

1. This manuscript developed a score to quantify the discrepancy between reporting rate and the epidemiological situation. However, the evidence to prove that the method is scientific and reasonable was not sufficient. For example, has this method been applied and proven in other studies before?

Thank you for this comment. We added references for the method “now-casting” as they are established methods to cope with reporting delays (lines 93ff). However, we do not have data that is usually used for such now-casting (namely infection date and reporting date) and we wanted to use a descriptive method as this is favored for decision makers. Our work presents a novel descriptive approach to which we cannot find comparable prior work.
To improve the scientific character of our work, we developed scenarios to showcase the negative effect of reporting delays on the 7-day incidence. These scenarios were added to the results section (lines 158ff). These observations motivated us to develop our scores that help detecting indicators for such delays. The results of these simulated scenarios are presented in line 252ff.

2. Also, as I mentioned before, this study showed a “weekend effect” for most countries. It is important finding. Please discuss the reasons and add the application measures for the political decision-making in the Discussion Section.

Thank you for this comment. As suggested, we extended the discussion of the weekend effect and added the application measures for epidemiologists advising policy makers (line 377ff).
Further, we added application measures for political decision making in our section “conclusion and implications for epidemiologists advising policy makers” (line 385ff):
“Furthermore, our clustering approach identified a “weekend effect”. Thus, if possible, epidemiologists should prefer using COVID-19 case counts reported by WHO from the middle of the week for advising political ad hoc decisions”.

Reviewer: 2
Gilberto Turati, Università Cattolica del Sacro Cuore

Comments to the Author:
Review on “Understanding COVID-19 reporting behavior to support political decision-making: A retrospective cross-sectional study of COVID-19 data reported to the World Health Organization”
I have now read the revised version R1 of the manuscript. I try to list here what are still open issues in my view.

3. Goal of the paper. The current version is much less ambitious than the original version of this manuscript. The idea in the original version was about using the frequency in reporting behaviour as a way to assess the quality of data related to Covid-19 infections. Why

frequency should be related to data quality was never spelled out clearly in the paper. And in my previous report I did report an obvious example that argued in favour of an opposite story: lower frequency can imply a better control and verification of the data, which implies better data; a behaviour which is independent on the number of cases and fatalities and it relates to the checks and effort that authorities in charge of collecting data provide before communicating the data. The goal in this version is a very simple descriptive exercise and an attempt at classifying different countries according to their "reporting behaviour", accounting for reporting gaps (relative reporting frequency per weekday regardless of the number of reported cases) and reporting lags (share of cases reported each weekdays relative to the number of cases reported each week). My suggestion would be to better clarify this in the first rows of the introduction, avoiding any reference to the quality of the data. Unless the authors provide a consistent conceptual explanation, I do not see any way to infer quality of data from analysing reporting behaviour. This is true even in relation to the 7-day incidence rate, used by some countries to implement some policies. I do not see any comparison of the indicators on gaps and lags provided here with the 7-day incidence rate.

Thank you for this comment. We revised the introduction as we had the feeling that our goal for this work has not been explained well enough. Please find this in line 77ff. Further, we added why now-casting was not applicable for WHO data and our purposes (93ff).

We agree with reviewer 2 that low frequency in reporting behavior can also be an indicator for good data quality as verification and data quality control needs time. Further, some countries may in fact have zero cases. We discussed this already in the previous version of our manuscript, but re-phrased it in the current version (343ff). We added evidence from literature as well as the calculated scenarios to provide a consistent conceptual explanation for our assumptions (line 343ff):
We used this assumption as consequence of our calculated scenarios, existing literature and our experiences as epidemiological advisers for political ad hoc decisions during the COVID-19 pandemic[17]. Even if the pandemic situation was steady and current case counts were low, most countries did report at least some single cases of COVID-19. However, for a few exceptions, this assumption might be incorrect. Less frequent reporting could also be an indicator of high data quality, as counts might be very thoroughly validated before reporting. Based on explanations above we consider our assumption eligible for most countries.".

Further, we attenuated the correlation between frequent reporting and high data quality in our conclusion and revised our conclusion as follows (380ff):
"However, the majority of countries have a high consistency of incidence rates and binary reporting rates ~~suggesting a high quality of reporting for these countries~~."

We hope that it becomes clearer now and that we could dissolve the reviewer's concerns.

4. Motivation. In my previous report, I have suggested the authors to improve the motivation of their paper. Even in the current version, the focus is on the use of Covid-19 WHO data for border closures and travel restrictions. I do believe the authors are missing at least one other important reason to discuss the reporting behaviour of countries: in many countries (all countries?), media and scholars have tried to evaluate the performance of their government in facing the pandemic, and the effectiveness of the harsh measures taken by their government, generating a hot political debate. Suppose I compare the lockdown imposed in Italy with the more liberal stance taken from the Swedish government, and data for Italy and Sweden miss out a number of infections but with different degrees of precision. How can I make any meaningful statement on the effectiveness of government interventions in both countries? This is the reason why it is of paramount important to define standards of collecting data at the world level: to favour cross-country comparisons and understand what works and what is not working. This is clearly not enough to avoid false reporting by countries of this strategic information.

Thank you for this comment. We revised the introduction as we had the feeling that we did not explain our motivation for this work well enough (line 77ff). We agree with reviewer 2 that it is very important to define standards of collecting data at the world level. This would improve cross-country comparisons and help to understand which pharmaceutical or non-pharmaceutical interventions (NPI) worked and which did not. We also mentioned this application of international data like WHO data in the introduction of our manuscript (line 64ff). We further agree that the robustness of epidemiological

indicators is not crucial for such retrospective effectiveness analyses that usually use time-dependent regression analyses and analyze long(er) periods of time. Fluctuation by weekdays or variations like the weekend effect are not relevant for this kind of retrospective analyses.
However, the initial motivation of this work was actually the weekly evaluation of WHO data for political ad hoc decisions like the weekly designation of travel risk areas by the European Union and some countries including Germany. For such approaches, fluctuation by weekdays, irregular reporting and / or observations like the weekend effect are relevant and may influence political decisions.
National ad hoc decisions like school closures, curfews or restriction of mass gatherings are most likely based on national data. Most countries use now-casting to support real-time COVID-19 situational awareness. This, however, is currently not possible for WHO data due to lack of infection date. We added the following explanation to the introduction (line 88ff):
"Yet, epidemiological indicators used - including 7-day incidence rates - lose robustness for ad hoc evaluations in case of strongly irregular reporting behavior. Several countries including Germany used 7-day incidence rates as one indicator of implementing travel restrictions[15]. This represents one example for which reporting lags or gaps might have an impact on political decisions. On the regional or national level, real-time COVID-19 decision making may be supported by statistical methods like now-casting[16-18] Such nowcasting approaches mitigate reporting delays but usually require information on infection and reporting date, or use other secondary data[17]. Unfortunately, the infection date does not exist for WHO data. Additionally, inferential techniques may not be reliable enough for strong political measures as the closing of boarders. Thus, an assessment and descriptive classification of the reporting quality of countries to WHO is needed to improve understanding of data quality for better informed political- ad hoc decisions." We hope that our explanation could convince the reviewer of the motivation and goal of our article. "

5. In my previous report, I have mentioned two specific issues to improve motivation: first, discuss what should be the appropriate governance for the pandemic. There is no point in making reference to WHO data if different countries do not believe about the benefits of centralizing information (and management of the pandemic) at the world level (and you can expect different types of country to behave differently in this respect). I have mentioned, as a reference to this specific issue of the global governance, the book Costa Font J., Turati G., Batinti A. (2020), The political economy of health and healthcare, Cambrdige University Press. This also to favour an interdisciplinary discussion on issues that should not be bounded in specific disciplines.

We agree with the reviewer. Willingness to cooperate, share data and knowledge by countries as well as centralizing information and working in interdisciplinary teams are essential for a successful management of the pandemic. We elaborated this in "implications for policy makers" section in the discussion (line 393 ff). Further, as requested, we also cited the book by Costa Font J., Turati G., Batinti A. (2020), The political economy of health and healthcare, Cambridge University Press as suggested by the reviewer.

6. Second, in stressing that this is not the first paper raising issues on Covid-19 data, I have suggested a brief discussion of the literature. In the motivation, it should be stressed what this paper does in addition to this literature.

Thank you for this comment. We discussed existing literature on quality of COVID-19 data in the previous version of our manuscript (line 373ff).

"Difficulties and obstacles in interpreting and comparing COVID-19 case counts within and between countries have been discussed since the early phase of the pandemic[19 20]. Differences may occur due to regional and country-specific variations in testing capabilities, testing policies, case definitions and preparedness[20 21]."
As suggested, we added the following paragraph to the introduction (line 100ff).

"Data quality of COVID-19 case counts has been criticized and discussed since the early phase of the pandemic[19-21]. In addition to the existing literature, this study used a new approach for analyzing reporting behavior of countries based on COVID-19 case count data submitted to the WHO. Our analyses tend to support ad hoc interpretations of WHO data for political decision makers and epidemiologists advising them."

Further, we revised the "strength and limitations" sections in the manuscript (line 45 and 318f): "~~To our knowledge,~~ This ~~is the first~~ study used a new approach for analyzing reporting behavior of countries based on COVID-19 case count data submitted to the World Health Organization (WHO)."

According to authors, this is the first paper analysing reporting behaviour by countries of Covid-19 cases to the WHO. This brings me to the importance of institutional details.
Institutional background and different waves. As I mentioned in my previous report, it would be really important to discuss the data generating process and the standards adopted to count Covid-19 cases (and fatalities). I have posed to the authors some questions:

(i) How do countries have to report data to the WHO?
We added a paragraph describing the reporting pathway in more detail (line 123ff).
(ii) Are there defined standards to be applied?
We addressed this comment in line 124 ff:
"National public health institutes or national ministries of health usually conduct this reporting. In Europe, a system called TESSy is used to report COVID-19 data by national public health institutes to ECDC 29. Subsequently, ECDC reports data to WHO."
(iii) How does it work within countries, especially those with a federal structure?
We explain this question in more detail in line 126ff: "This process differs by country or region. Countries may report data stratified on the federal level. For WHO headquarters, however, this data is available only aggregated to national-level."
(iv) Have these standard been changed from the first to the second wave, as knowledge on the pandemic made some progress?

We added the following paragraph to the manuscript (line 131 ff):
<<WHO reports the following changes in the data collection processes: "From the 31 December 2019 to the 21 March 2020, WHO collected the numbers of confirmed COVID-19 cases and deaths through official communications under the International Health Regulations (IHR, 2005), complemented by monitoring the official ministries of health websites and social media accounts. Since 22 March 2020, global data are compiled through WHO region-specific dashboards (see links below), and/or aggregate count data reported to WHO headquarters daily."[14].>>
(v) The last question is particularly important to study reporting behaviour as it is clear that, in the first wave, the number of cases is severely underestimated for many reasons (different standards in reporting, ability to test, poor testing strategies, …). I do not see any answers to my questions in the revised version of the paper, but for few lines in the "Sensitivity analysis" section toward the end of the paper on how the reporting behaviour changed over time. From Appendix 10, it looks like binary reporting reacted swiftly to the onset of the pandemic. Why?
We produced a plot to show that the increase overlaps with the increase of reported cases (here shown using the 7-day incidence, Appenidx 12). We revised the following paragraph in the manuscript (line 302ff):
"Plotting the global binary reporting rate over time shows a low reporting rate that quickly increased around March 2020. The peak was reached in April 2020 suggesting a reliable reporting behavior. A small decline over New Year's Day could be observed (Appendix 11). The great improvement of the reporting behavior overlaps with the worldwide average 7-day incidence that also started to grow noticeable around March 2020 (Appendix 12). However, performing sensitivity analyses by excluding early data (January – March 2020) for calculation of country-specific reporting scores did not change our main findings (Appendix 13)."

7. When did WHO issued the first definition of a Covid-19 case for its surveillance system? The authors mention that the data are compiled using region-specific WHO dashboards since 22 March 2020 but they never really use this information.

We added the following paragraph to the method's section of the manuscript (line 115ff).
The first case definition for "human infection with novel coronavirus (nCoV) was published by WHO as interim guideline in January 2020 [24]. Since then, several updates of WHO COVID-19 case definitions have been released to adapt to the current evidence available [23 25-28]."

8. Takeaway. I have read conclusions of this paper many times. A fair summary of the takeaway messages is the following: (i) most of the countries report cases when they do have cases to report; (ii) there seems to be a weekend effect, hence counts reported in the middle of the

week are more reliable. If there is any other strong message, it does not seem to emerge from the paper.

Thank you for this comment. We revised the take-away messages of this manuscript as it seems that we did not explain it well enough. Our conclusion and implications for epidemiologists / infection control experts advising policy makers are the following (line 377ff):
Global reporting behavior of COVID-19 case counts by WHO member states is diverse. However, the majority of countries have a high consistency of incidence rates and binary reporting rates. Our score system might be a helpful tool for infection control experts and epidemiologists advising policy makers. It may help them to consider country-specific reporting behaviors in political ad hoc decisions based on WHO data including designation of travel risk areas. Furthermore, our clustering approach identified a "weekend effect". Thus, if possible, epidemiologists should prefer using COVID-19 case counts reported by WHO from the middle of the week for advising political ad hoc decisions. Further, spectral clustering identified countries with unusual or irregular reporting that should be evaluated especially carefully. However, our scores and cluster analyses should be applied keeping in mind the limitations. They do not replace thorough analyses of quantitative and qualitative indicators of the COVID-19 situation in each country for an informed decision-making. "

We further revised the section "implication for policy makers" (line 393ff):

Implication for policy makers
"For the current and future pandemics, we need a robust system of epidemic intelligence to timely collect, share and analyze data at the regional, national and international level for better-informed political decisions. In September 2021, the first WHO hub for pandemic and epidemic intelligence was inaugurated in Berlin, Germany to achieve this ambitious goal[39]. The WHO hub aims to create a collaborative, interdisciplinary environment and may become the foundation for "better data, better analytics and better decisions" for its 193 member states[40]. Successful pandemic management requires the willingness of countries to cooperate, share data and knowledge as well as centralizing information. However, healthcare activities including international cooperativeness depend on the design of country's health institutions. These may be affected by institutional influences including federalism, electoral competition, constitutional designs or political ideologies[41]. Finding ways to overcome these barriers will be the big challenges for the future."
Reviewer: 1
Competing interests of Reviewer: No.

Reviewer: 2
Competing interests of Reviewer: No competing interests.

VERSION 3 – REVIEW

REVIEWER	Turati, Gilberto Università Cattolica del Sacro Cuore, of Economics and Finance
REVIEW RETURNED	25-Oct-2022

GENERAL COMMENTS	Review on "Understanding COVID-19 reporting behavior to support political decision-making: A retrospective cross-sectional study of COVID-19 data reported to the World Health Organization" I have now also read the revised version R2 of the manuscript. I list here several remarks/questions that I have marked on the paper when reading it. The list is provided as remarks appear in the current manuscript and does not reflect their importance. These are just raw notes. Line 96: typo: should be borders Line 113: typo: are available at where?

	Lines 137-38: authors claim that confirmed Covid-19 cases data are better reflecting reporting behaviour, but they do not clarify why this should be the case Line 140: would argue that binary reporting rate is not a “dimension” of reporting new cases. Is an indicator related to "binary" reporting meaning whether reporting occurred or not Lines 146-147: “If on Tuesdays, the same country only reported every second week, the country would get a binary reporting rate of 0.5 for Tuesdays and so on”: it is not clear to me whether 0.5 is due to the choice of using 14-days time series. Unfortunately, the reader knows about this choice at line 159 Lines 160-168: I honestly do not understand what the authors are trying to say here Lines 183-187: I do not understand why authors speak about “discrimination” here. Their reporting score varies between -1 and +1. Values close to -1 mean that incidence rate is high but reporting activity does not follow; values close to +1 mean that incidence rate is low but reporting occur. In both cases, there is a misalignment between incidence and reporting. Values around zero should suggest that there is no misalignment between incidence and reporting: countries report when they have to. In this case, the imputation procedure authors speak about does not really make sense: a country reporting less than seven cases per week have an incidence close to zero and coherently should have a reporting close to zero. What is a “perfect reporting score” in authors’ view? Does this imply a value judgement by the authors? Lines 234-236: is this the result of one or few countries? Results like this one should be commented better Lines 243-245: so what are the conclusions from here? What do we learn? Line 249: what are the six clusters? Can we learn something from here? Lines 274-275: authors speak about six clusters but then suggest that countries can be classified only in two groups Lines 302-304: is it fair to say that countries reported when they had to report? the paper could be a note suggesting this. Lines 321-326: authors speak about limitations in the data provided by the WHO. As far as I can understand, the current story is: countries had to take severe decisions (including travel restrictions); we study a way to understand whether a country is reporting well about Covid-19 cases; however, be aware that available data are bad and do not really allow for comparisons across countries. What kind of story is this? Let me now get back to my previous report and to evaluate the current manuscript in the light of the report. Goal of the paper. The goal of the paper comes at lines 104-107 at the end of the introduction. I do not see why, having suggested to start with a clear statement of what the paper is about. There are still some references (e.g., line 73) to the importance of knowing about limitations of the available international data about Covid. These limitations are made clear in lines 321-326 and are not inferred from the authors’ exercise. Motivation. In my previous report, I have suggested the authors to improve the motivation of their paper. Even in the current version, the motivation is weak and one of the two specific issues mentioned to improve motivation has not been used (the appropriate governance for the pandemic). For the second one, it is still unclear what the paper adds to the literature criticizing the quality of Covid-19 data for international comparisons.

	Institutional background and different waves. This is the area of the paper where I can see improvements with respect to my previous report (lines 109-141). Takeaway. I have read conclusions of this paper many times, and the paper itself many times. I continue to think that a fair summary of the takeaway message is the following: (i) most of the countries report cases when they do have cases to report; (ii) there seems to be a weekend effect, hence counts reported in the middle of the week are more reliable. This paper lacked and continue to lack a coherent story to tell since the beginning. Authors should lay down their main message and build a coherent structure around this message.

VERSION 3 – AUTHOR RESPONSE

Reviewer: 2
Prof. Gilberto Turati, Università Cattolica del Sacro Cuore

Comments to the Author:
Review on "Understanding COVID-19 reporting behavior to support political decision-making: A retrospective cross-sectional study of COVID-19 data reported to the World Health Organization"
I have now also read the revised version R2 of the manuscript. I list here several remarks/questions that I have marked on the paper when reading it. The list is provided as remarks appear in the current manuscript and does not reflect their importance. These are just raw notes.

1. Line 96: typo: should be borders

Thank you for this comment. We revised this typo as suggested (line 126).
"Additionally, inferential techniques may not be reliable enough for strong political measures as the closing of borders".

2. Line 113: typo: are available at where?

The dataset is available at our github repository (https://github.com/aauss/reporting_behavior). We revised the sentence and the respective reference accordingly. (line 158 f.):
"Data sets generated for this study and raw data used are available at our github repository [1]."

3. Lines 137-38: authors claim that confirmed Covid-19 cases data are better reflecting reporting behaviour, but they do not clarify why this should be the case

We added explanations for our assumption. Please find the revised section in line 183 ff.:
"We expected COVID-19-case counts to provide a more robust data base for analyses of reporting behaviors due to the following reasons. COVID-19 related deaths occur less frequently (compared with COVID-19 case counts), they occur with a delay of days to weeks after infection (and consequently lag behind the current epidemic situation) and do face challenges in the attribution of cause of death [2]."

4. Line 140: would argue that binary reporting rate is not a "dimension" of reporting new cases. Is an indicator related to "binary" reporting meaning whether reporting occurred or not

Thank you for this comment. We used the term "indicator" instead of "dimension" and revised the section as follows (line 189):
Data analyses were stratified by time (day of the week) and location (global, regional, and country-specific). Reporting behavior was assessed by investigating two ~~dimensions~~ indicators ~~of reporting new cases~~: binary reporting rate and relative reporting behavior.

5. Lines 146-147: “If on Tuesdays, the same country only reported every second week, the country would get a binary reporting rate of 0.5 for Tuesdays and so on”: it is not clear to me whether 0.5 is due to the choice of using 14-days time series. Unfortunately, the reader knows about this choice at line 159

Thank you for this remark. We revised the respective paragraph and elaborated the explanations of our examples in the method’s section as follows (line 191 ff.):

First, binary reporting rate was defined as the relative reporting frequency per weekday regardless of the number of reported cases, i.e., whether reporting occurred or not. The binary reporting rate was calculated over the whole time series if not indicated otherwise. As an example, if a country reported case numbers larger 0 for all Mondays in the period under review, it would get a binary reporting rate of 1 for Mondays. Explicitly, we divided the number of Mondays with reporting in the dataset – in this case all Mondays – with the number of all Mondays in the dataset. If on Tuesdays, the same country only reported every second week, the country would get a binary reporting rate of 0.5 for Tuesdays. For this example, we assumed an even number of Tuesdays in the time series. In this case, reporting every second week is identical with reporting for half of all Tuesdays. Again, being explicit, we divided the number of Tuesdays – in this case half of all Tuesdays - by the number of all Tuesdays. In this example, we obtained the value 0.5. This process was repeated for all weekdays.

6. Lines 160-168: I honestly do not understand what the authors are trying to say here

We apologize for not explaining our approach well enough. We considerably revised this section to provide more details and to improve the structure. Further, we included a subheading and removed this paragraph to the end of the method’s section. Details were moved to the appendix. We hope that this could improve the readability (line 296 ff.):

Simulations of reporting delays to illustrate adverse effects

To ~~illustrate~~ motivate our scores and show adverse ~~the~~ effects of delayed reporting, we produced several scenarios ~~in which the~~ with ~~impact on the reported cases due to~~ simulated reporting delays using the WHO COVID-19 dataset of this work [1]. ~~was shown.~~ Delayed reporting decreases timeliness that is an important data quality property and particularly important for COVID-19 [3 4].

The impact of the delay in our scenarios is measured by the relative difference between the reported case counts from the actual, non-delayed case counts. The higher their difference, the worse the impact of the delay. We picked three 14-day long time series from the data set with low, medium and high number of reported cases. Details of this method are described in the Methods Appendix.

Methods Appendix

To select three adequate time series, ~~fFirst~~, we constructed 14-day long time series of reported COVID-19 cases for all countries and over the entire time period. 14 days capture meaningful dynamics of infections while not being too long. Subsequently, we kept only time series in which more than zero cases were reported each day. We performed a linear regression on them and only kept time series with ~~showed~~ a positive slope since an increase of cases will require an *ad hoc* evaluation for mitigation measures opposed to a decrease. To select time series with low, medium and high number, we divided the remaining time series by quantile. Using the sum over the reported cases per time series, we calculated the 25%, 50%, and 75% quantile ~~for all time series~~ and picked the ones closest to those quantiles.

Using the three selected time series ~~Afterwards~~, we simulated delayed reporting. A delay was simulated by removing all reported cases from the delay period and adding them to the next successful reporting day. We looked at reporting delays from one to six days. ~~during the delay period (1-6 days) and added them to the next successful reporting day.~~ For example, a country reported the following case numbers [10,14,22,15,7] from day one to day five. ~~G~~iven Assuming a two-day delay, the following reported cases [10,14,22,15,7] would turn into [0,0,24,0,0,~~44~~37]. Cases from the first two-day delay period are reported on the third day (10+14=24). Cases from the third and fourth day are added to the fifth (22+15=37). The cases of the fifth day (7 cases) will appear on day seven. Finally, we calculated the relative difference of the real case counts and the ones caused by the delay per day and for each time series.

7. Lines 183-187: I do not understand why authors speak about "discrimination" here. Their reporting score varies between -1 and +1. Values close to -1 mean that incidence rate is high but reporting activity does not follow; values close to +1 mean that incidence rate is low but reporting occur. In both cases, there is a misalignment between incidence and reporting. Values around zero should suggest that there is no misalignment between incidence and reporting: countries report when they have to. In this case, the imputation procedure authors speak about does not really make sense: a country reporting less than seven cases per week have an incidence close to zero and coherently should have a reporting close to zero. What is a "perfect reporting score" in authors' view? Does this imply a value judgement by the authors?

We thank you for your comment and questions. We understand that it is not clear what a preferable and unpreferable reporting score is. While it is relatively obvious for negative values – as you also write – a reporting score >= 0 is less intuitive to classify. You are right that there is a numerical misalignment with positive scores as well. From a reporting quality perspective, however, we assume more reliability of data if a country is frequently reporting even the smallest number of cases. In the methods section, we added examples for more clarity (line 249 ff.):

In consequence, scores close to 1 indicate a high reporting rate and a comparably small incidence. This could be observed in countries that reported very frequently even if the number of new cases was small. We assumed that most countries with values close or equal to 1 have a successful COVID-19 response and a "high probability of a high reporting diligence". Values below 0 indicated insufficient reporting frequencies given relatively high incidences. These cases might indicate a strong reporting delay or other difficulties in reporting and represented countries with a "low probability of a high reporting diligence". A reporting score equal or close to zero can be observed, i.e., when the binary reporting score matches the scaled incidence of a country. Such scores may be interpreted as "medium probability of a high reporting diligence" that need closer examination with medium priority. If incidences were generally high among countries, a higher binary reporting rate would be needed to achieve a score of zero and *vice versa*. We call this measure the *reporting score*. Thus, the reporting score was defined as country-specific measure within the range of -1 and +1 that is based on the normalized means of 7-day incidence rates and binary reporting rates. Values close or equal to 1 are interpreted as the optimum while low reporting rates might require closer examinations with medium (reporting scores equal or close to zero) or high priority (negative reporting scores). However, very high reporting rates might also be an indicator of false-reporting (please see limitations for more details).

Further, we would like to comment on the term "discrimination". Discrimination is not meant as a moral or value judgement but as an undesired property of the score that we wanted to balance out. We deleted the term "discrimination" to avoid any judgmental wording from our side (line 267).

Due to our main problem that we have not information on the reporting date of a country, as described in the Introduction (lines 124 f.), we cannot distinguish between a country not reporting or a country faithfully reporting no cases. We assume overall faithful reporting and therefore focus on the latter (line 250 ff.). A country with no COVID-19 cases should therefore not receive a high reporting score by default (lines 267 ff.). We also clarified why we impute from case numbers lower than 7 (271 ff.) which is the number of cases for which a perfect binary reporting score is not possible anymore causing an undesirable decrease in the reporting score (273 ff.):

~~"Unfortunately~~, ~~t~~The reporting score ~~discriminates against~~ in countries with no/few COVID-19 cases ~~since this leads to a~~ will be close to 0 as the binary reporting by definition ~~rate of or~~ will be close to 0 in this scenario. When excluding false reporting of no cases, countries reporting no or small numbers of cases should also be able to receive a perfect reporting score as there are no problems with reporting. For such countries, we applied an imputation on the reporting rate. If a country reported less than seven cases per week, which is the minimum number of cases to theoretically achieve a perfect binary reporting score, it could never obtain a perfect reporting score although it might be reporting reliably. In this case, we imputed one single new case on days where actually no new case was reported."

8. Lines 234-236: is this the result of one or few countries? Results like this one should be commented better
   "The comparison of the mean incidence rates and binary reporting rates showed that a decreasing incidence rate does not always match with low reporting rates and vice versa. The WHO region Americas, for instance, had the second lowest reporting rate while reporting the second highest incidence rates at the same time."

We apologize for not being clear enough. We revised this section as follows to make ourselves clear that we are still referring to Table 2 and that our statements refer to all countries available per WHO region (line 350 ff.):

The comparison of the mean incidence rates and binary reporting rates within one WHO region showed that a ~~decreasing~~ low incidence rate does not always match with low reporting rates and vice versa (Table 2). The WHO region Americas, for instance, had the second lowest reporting rate while reporting the second highest incidence rates at the same time.

9. Lines 243-245: so what are the conclusions from here? What do we learn? "When excluding states with less than 500,000 inhabitants and an area less than 1000 km² which usually are islands with an advantage in mitigating the import of COVID-19 cases, we identified the highest scores for China (0.98), Tajikistan (0.97) and Egypt (0.97), while the lowest scores were found in Montenegro (0.11), Czechia (0.13), and Slovenia (0.29)."

High scores (close to 1) indicate that countries reported very frequently even if the number of new cases was small. We assumed that most countries with values close or equal to 1 have a successful COVID-19 response and a "high probability of high reporting diligence". Positive scores close to zero observed for some European countries can be interpreted as "medium probability of high reporting diligence" that require any further qualitative analyses or careful interpretation with medium priority.

We rephrased the respective section in results and added the following interpretation (line 361 ff.):
Even the lowest scores ~~were~~ found in Montenegro (0.11), Czechia (0.13), and Slovenia (0.29) were in the positive range. These low positive scores might be interpreted as "medium probability of high reporting diligence" that required closer examination with only medium priority.

However, high scores can also be achieved by consciously false reporting that we do not detect by our approach. We consider those as exceptional cases that need careful interpretation in the political context. We extended the limitations in the discussion section accordingly (line 491 f.).

"High scores can also be achieved by consciously false reporting that we do not detect by our approach. However, we consider those as exceptional cases that need careful interpretation in the political context."

10. Line 249: what are the six clusters? Can we learn something from here? "To identify outstanding patterns in a country's reporting behavior, we applied spectral clustering to all 222 country-specific different binary reporting rates and mapped them to 6 clusters."

The 6 clusters were generated by our spectral clustering approach and can be seen in Figure 2. We now refer to Figure 2 as can be seen in line 370 f.:

"To identify outstanding patterns in a country's reporting behavior, we applied spectral clustering to all 222 country/area-specific different binary reporting rates and mapped them to 6 clusters (Figure 2)."

Further, we removed the paragraph on our scenarios (former 371 ff.) to the end of the section (now 401 ff.). With this, we do not interrupt the read flow on our cluster interpretations (368 ff.).

11. Lines 274-275: authors speak about six clusters but then suggest that countries can be classified only in two groups Lines 302-304: is it fair to say that countries reported when they had to report? the paper could be a note suggesting this.

By our spectral clustering, we identified six clusters. For further simplification, we applied another country level analysis that counted the number of weekdays with a binary reporting rate below 50%. Thereby, we identified two main groups that included more than 80% of our countries. We agree with the reviewer' conclusion. Most countries (n = 137, 62%) reported COVID-19 cases when they had to report those. We revised our conclusion as suggested by the reviewer in line 39 (abstract) and 514 ff. (please see point 16 for more details):

~~"Many~~ The majority of countries ~~have~~ reported COVID-19 cases when they did have cases to report. ~~had to report- high consistencies of incidence rates and binary reporting rates. Some, however, show highly irregular reporting.~~ The identification of a slight "weekend effect" suggests that COVID-19 case counts reported in the middle of the week may represent the best data basis for political *ad hoc* decisions. A few countries, however, showed unusual or highly irregular reporting that might require careful interpretation. Our score system and cluster analyses might be applied by epidemiologists advising policy makers to consider country-specific reporting behaviors in political *ad hoc* decisions. "

As reported, the second largest group (n = 49 countries) had a binary reporting rate below 0.5 for each workday. This was also found by our country-level analysis counting the number of weekdays with a binary reporting rate below 50%. Thus, irregular reporting was observed for 22% of countries, e.g. by reporting cases only every second week. We consider this as relevant finding as our clustering approach can be used in addition to the reporting score to quickly identify countries with irregular reporting. By this approach, time-consuming manual review of WHO data time series and additional information could be restricted to countries that belong to certain clusters and / or had a "low probability of high reporting diligence. We highlighted this issue in the conclusion (line 513 ff.):

Conclusion ~~and implication for epidemiologists advising policy makers~~

Global reporting behavior of COVID-19 case counts by WHO member states ~~is~~ was diverse, but ~~However,~~ the majority of countries reported COVID-19 cases when they did have cases to report. ~~have a high consistency of incidence rates and binary reporting rates.~~ Furthermore, our clustering approach identified a "weekend effect" suggesting COVID-19 case counts reported by WHO from the middle of the week being more reliable for advising political *ad hoc* decisions. Spectral clustering identified a few countries with unusual or irregular reporting that should be interpreted especially carefully. In consequence, elaborative manual review of WHO data time series and additional information could be

restricted to countries with a “low probability of high reporting diligence” and / or affiliation to certain clusters. However, our scores and cluster analyses should be applied keeping in mind its limitations. They do not replace thorough analyses of quantitative and qualitative indicators of the COVID-19 situation in each country for an informed decision-making.

12. Lines 321-326: authors speak about limitations in the data provided by the WHO. As far as I can understand, the current story is: countries had to take severe decisions (including travel restrictions); we study a way to understand whether a country is reporting well about Covid-19 cases; however, be aware that available data are bad and do not really allow for comparisons across countries. What kind of story is this?

We appreciate the scientific discussion and your efforts to improve our manuscript. We think that this is an interesting work on the interface between science and politics that will gain even higher significance in the future. We need to find pragmatic solutions that are as precise as possible to timely support political decisions. At the same time, we need awareness on data limitations. It is correct that our work does not have an effect on the general data limitations as reported by WHO (line 456 ff.). However, we think it is relevant to the scientific community due to the following facts:

- limitations around data quality concerns or incomplete data (in our case the omission of the reporting date mentioned in lines 124 f.) are ubiquitous,
- limitations are hard to mitigate, and often ignored in practice.

Thus, we tried to support severe decisions in times of emergency by providing reproducible and comprehensible metrics describing and analyzing reporting characteristics of countries. The goal of this work was handing experts a tool for decision making, but does not solve deep rooted problems with data quality.

We clarified in the discussion that general data limitations are not due to our metrics but due to the setting and are thoroughly and already reported by WHO (line 456 ff.). We revised this section as follows to avoid confusion on where those limitations originate from and who identified those (line 454 ff.).

“However, ~~there are~~ some limitations generally apply to analyses of public health data across countries with different resources and policies. These limitations were not identified by this work and continue to exist even with the results of the metrics introduced here. WHO reports these limitations in detail in the data source explanation[16].”

13. Let me now get back to my previous report and to evaluate the current manuscript in the light of the report. Goal of the paper. The goal of the paper comes at lines 104-107 at the end of the introduction. I do not see why, having suggested to start with a clear statement of what the paper is about.

According to the STROBE guidelines for reporting observational studies in epidemiology, we report our hypotheses (research question) at the end of the introduction (https://www.equator-network.org/reporting-guidelines/strobe/). However, as suggested by the issues 15 and 18 by the reviewer, we considerably revised the introduction. The motivation and goal of this work is now presented earlier in the manuscript (line 97 ff.). We revised the last sentence of our introduction as follows (line 148 ff.): “

This work, however, focused on a methodological approach to improve understanding of data sources, reporting behavior, and obstacles to consider for better informed political *ad hoc* decisions. We did an observational cross-sectional study of COVID-19 case counts reported by the WHO headquarter ~~were analyzed~~ from January 2020 until June 2021. We used this data to characterize, describe and analyze

COVID-19 reporting properties and patterns by developing reporting scores and performing spectral clustering analyses.

14. There are still some references (e.g., line 73) to the importance of knowing about limitations of the available international data about Covid. These limitations are made clear in lines 321-326 and are not inferred from the authors' exercise.

Thank you for this comment. During the considerable revision of our introduction based on your comments, we deleted the section on limitations from the introduction. The references cited in the introduction refer to sources of global COVID-19 data (e.g. ECDC, WHO, OWD). Further, for clarification we added the following section to the discussion (line 456 ff.):

"However, ~~there are~~ some limitations generally apply to analyses of public health data across countries with different resources and policies. These limitations were not identified by this work and continue to exist even with the results of the metrics introduced here. WHO reports these limitations in detail in the data source explanation[16]."

15. Motivation. In my previous report, I have suggested the authors to improve the motivation of their paper. Even in the current version, the motivation is weak and one of the two specific issues mentioned to improve motivation has not been used (the appropriate governance for the pandemic). For the second one, it is still unclear what the paper adds to the literature criticizing the quality of Covid-19 data for international comparisons.

Thank you for this important comment. We considerably revised the introduction to improve the presentation of our motivations. In the previous version of our manuscript, we included the topic "appropriate governance for the pandemic" in the section of the discussion "Implication for policy makers". As suggested by the reviewer, we included this motivation to our introduction (fact 3). Thus, this work was mainly motivated by three facts:

1) WHO data provided an important basis for political decisions, but analyses on characteristics of this data were scarce. We aimed to analyze and understand COVID-19 case count data submitted to the WHO best possible to optimize ad hoc interpretations of WHO data for political decision makers. In addition to the existing literature, this work provides an assessment and descriptive classification of WHO data including analyses on the reporting behavior / characteristics.

2) Timeliness of data is crucial for political ad hoc decisions like travel restrictions. Unfortunately, the infection date that is required for "now-casting" or any secondary data were not available for WHO data. Additionally, inferential techniques may not be reliable enough for strong political measures as the closing of borders boarders. Thus, we wanted to provide an easy tool that allows timely evaluations of COVID-19 reporting behavior by country without applying prediction models or now-casting. We developed a reporting score to support interpretation of reporting behavior for better informed ad hoc decision-making including implementation of travel restrictions.

3) This work should highlight the importance of an appropriate pandemic governance by all countries including publicly and timely shared data. Successful pandemic management requires appropriate governance within countries as well as internationally coordinated actions. Such efforts are highly dependent on the willingness of countries to cooperate, share data and knowledge as well as centralizing information (25).

Please see line 99 ff. for our revised introduction. We hope that the motivation of this work is clear and sound now.

16. Institutional background and different waves. This is the area of the paper where I can see improvements with respect to my previous report (lines 109-141).
Takeaway. I have read conclusions of this paper many times, and the paper itself many times. I continue to think that a fair summary of the takeaway message is the following: (i) most of the

countries report cases when they do have cases to report; (ii) there seems to be a weekend effect, hence counts reported in the middle of the week are more reliable.

Thank you for this comment. We revised the conclusion as suggested (line 39 and 513 ff.) and added one more take-away message. We sincerely hope that you agree with the following conclusion (abstract):

~~Many~~ The majority of countries ~~have~~ reported COVID-19 cases when they did have cases to report. ~~had to report high consistencies of incidence rates and binary reporting rates. Some, however, show highly irregular reporting.~~ The identification of a slight "weekend effect" suggests that COVID-19 case counts reported in the middle of the week may represent the best data basis for political *ad hoc* decisions. A few countries, however, showed unusual or highly irregular reporting that might require more careful interpretation. Our score system and cluster analyses might be applied by epidemiologists advising policy makers to consider country-specific reporting behaviors in political *ad hoc* decisions.

And conclusion (line 513):

"Conclusion ~~and implication for epidemiologists advising policy makers~~

Global reporting behavior of COVID-19 case counts by WHO member states ~~is~~ was diverse, but ~~However,~~ the majority of countries reported COVID-19 cases when they did have cases to report. ~~have a high consistency of incidence rates and binary reporting rates.~~ Furthermore, our clustering approach identified a "weekend effect" suggesting COVID-19 case counts reported by WHO from the middle of the week being more reliable for advising political *ad hoc* decisions. Spectral clustering identified a few countries with unusual or irregular reporting that should be interpreted especially carefully. In consequence, elaborative manual review of WHO data time series and additional information could be restricted to countries with a "low probability of high reporting diligence" and / or affiliation to certain clusters. However, our scores and cluster analyses should be applied keeping in mind its limitations. They do not replace thorough analyses of quantitative and qualitative indicators of the COVID-19 situation in each country for an informed decision-making. "

17. This paper lacked and continue to lack a coherent story to tell since the beginning. Authors should lay down their main message and build a coherent structure around this message.

Thank you for this important comment. We considerably revised the introduction accordingly (please also see point 13 and 15 for details). The revised introduction can be found in line 65 ff.

References

1. Abbood A UA, Denkel LA. Understanding COVID-19 reporting behavior to support political decision-making: A retrospective cross-sectional study of COVID-19 data reported to the World Health Organization, 2021.
2. Battegay M, Kuehl R, Tschudin-Sutter S, et al. 2019-novel Coronavirus (2019-nCoV): estimating the case fatality rate–a word of caution. *Swiss medical weekly* 2020(5)
3. Pipino Leo L LYW, Wang Richard Y. Data quality assessment. *Commun ACM* 2002;45(4):211–18. doi: 10.1145/505248.506010

4. Tucker Catherine WYC. The Role of Delayed Data in the COVID-19 Pandemic (Preprint). *SSM* 2021

[i] Brier, G. W. (1950). Verification Of Forecasts Expressed In Terms Of Probability. Monthly Weather Review 78, 1, 1-3, available from: < https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2> [Accessed 04 June 2022]

[ii] Hyvärinen, A (2005). Estimation of Non-Normalized Statistical Models by Score Matching. Journal of Machine Learning Research 5, 24, available from < https://jmlr.org/papers/v6/hyvarinen05a.html> [Accessed 04 June 2022]

[iii] World Health Organization (2022). WHO Coronavirus (COVID-19) Dashboard - Data sources 2022, available from <https://covid19.who.int/data> [Accessed 13 May 2022]

[iv] Robert Koch Institute (2022). Information on the designation of international risk area 2022, updated 25/02/2022, available from <https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Transport/Archiv_Risikogebiete/Risikogebiete_2022-02-25_en.pdf?__blob=publicationFile> [Accessed 16 May 2022]