# Reply to the Reviewers

*Re: Manuscript ID 304c3895-aec8-4c97-867a-23905ca7f85d*
*"On the assessment of the ability of measurements, nowcasts, and forecasts to track changes"*
*Jonas Rieger*, Bolin Liu, Bernd Saugel, Oliver Grothe*
*BMC Medical Research Methodology*

We thank the two reviewers for their comments and ideas for improving the manuscript. Please find our answers to the issues raised by the referees below.

---

**Reviewer #1, comment #1**

*The reported ATC Ratio results would be improved if you include the raw number of positive and negative concordance pairs in your tables or table captions. You reference were an interested person could look at the raw data, but it would be helpful to have these summaries in the paper.*

**Our response #1.1**

Thank you for your helpful suggestion. We have added the raw numbers of positive and negative concordance pairs to the table captions, providing readers with direct access to these important summaries within the paper itself.

**Reviewer #1, comment #2**

*I could see this paper being useful for forecasters and public health officials better communicate and understand the limitations of a particular model. I think it would be valuable to expand your discussion on how these metrics can be used in practice, particularly for forecasters and public health officials.*

**Our response #1.2**

Thank you for this valuable insight. We appreciate your perspective on the practical applications of our work. In response, we have expanded the discussion section to include more detailed commentary on how these metrics can be utilized in practice, particularly focusing on their relevance for forecasters and public health officials.

**Reviewer #1, comment #3**

*The literature review seems a little thin on references to concordance metrics for forecasts. It seems that references 21, 34, and 35 are the only one, but I am sure there are more out there.*

**Our response #1.3**

Thank you for bringing this to our attention. We agree that a more comprehensive literature review would ease further review for the readers. In response, we have expanded the number of references pertaining to concordance metrics for forecasts, providing a more thorough overview of the current state of research in this area.

**Reviewer #1, comment #4**

*Related to the previous point, you could more clearly identify the new methods you are proposing versus what has already been explored in the literature. I think just changing some of the wording in Section 2 would help differentiate your work from other people's work in this area.*

**Our response #1.4**

We appreciate your suggestion for improving the clarity of our contribution. We have revised the wording in Section 2 to more distinctly highlight our proposed new methods and differentiate them from existing work in the literature.

**Reviewer #1, comment #5**

*On page 27, there is the line "the BCa is computationally expensive but requires smaller datasets for reasonable confidence intervals." I don't believe this correct. The BCa method does not require smaller sample. It is more robust to small sample size and able to provide reasonable confidence intervals when the sample size is small.*

**Our response #1.5**

Thank you for this important clarification. You are correct, and we appreciate you pointing out this inaccuracy. We have revised the statement to more accurately reflect that the BCa method is more robust to small sample sizes and can provide reasonable confidence intervals even when the sample size is small, rather than requiring smaller datasets. The formulation has been made clearer as you suggested.

---

**Reviewer #2, comment #1**

*Secondly, I suggest the authors elaborate on the importance of the ATC ratio, particularly its relevance to the specific data being analyzed. ATC radio is essentially one specific property of a distribution, and having better prediction accuracy for ATC radio does not automatically make it the superior model. The objects of interest could be mean, variance, quantile, expected shortfall, or even the entire distribution. It would be beneficial to explain why ATC is crucial for the three empirical studies. For instance, with Covid data, one could argue that tail performance (high infection rate or death toll) is more important for decision-makers.*

**Our response #2.1**

Thank you for this insightful comment. We agree that the ATC ratio is indeed one specific property of a distribution and not always the most critical measure. As we stated in the paper, we view it as an additional measure of interest that focuses on a particular aspect of performance. To address your suggestion, we have elaborated on why the ATC ratio adds value to the evaluation in each example's introduction, providing clearer context for its relevance alongside other important metrics such as tail performance.

**Reviewer #2, comment #2**

*Finally, I recommend comparing the ATC-based evaluation methods with more traditional methods based on other statistical properties, like the log score, continuous ranked probability score, or quantile scores. It would be useful to discuss what additional insights the ATC-based approaches can offer compared to these traditional ones.*

**Our response #2.2**

Thank you for this highly interesting suggestion. We agree that comparing ATC-based evaluation methods with traditional statistical approaches would provide valuable context. In response, we have added the log score, continuous ranked probability score, and quantile scores to the tables in the appendix, making these traditional metrics more explicit for comparison. While we have not included a theoretical assessment of the differences between these scores and our methods to maintain the practical focus of this paper, we recognize the importance of this point and plan to expand on it in a future, separate publication dedicated to this comparison.