Initial Review

First submission on May 15, 2020
Reviews received on June 3, 2020
Revision submitted on August 3, 2020

EIC

Thanks again for submitting your article to HDSR, which has been reviewed by three reviewers and an AE. Their reports are attached. As summarized by the AE, the reviewers appreciate the important contributions you made. At the same time, there are some major concerns, one is the data quality, and the other is potential model overfitting, including over-explaining away those cases that did not work well.

What this paper reminded us is that although the data quality is very low (even for deaths, since for example reporting delay and competing risks are all real issues that are not being addressed), there are still three things we can do:

(A) Be as transparent as possible about the sources of the biases

(B) Mitigate those biases that can be modelled/captured (such as time lag handled in the paper)

(C) For biases that we cannot adjust, discuss their likely directions, magnitudes and impact

Regarding the overfitting/over-attributions, I think you need to be more explicit on how you tested the models and made various choices (such as AE's question of how you choose 7 days) to give readers more confidence in your predictions. I believe the vast majority of reviewers recognize the difficult problems and sometimes impossible tasks the authors are dealing with (not just for your paper, but for all other submissions to the special issue), but because the great/grave consequences and impact of the papers on COVID-19, the reviewers still want the authors to exercise their ultimate due diligence.

AE

This paper makes 3 nice contributions: (1) a public data repository, (2) a model averaging approach to prediction, and (3) model-free prediction intervals that performed well in their test cases.

The provision and description of a data repository are major contributions of this paper. Can you say more about how the data in reference [33] complement the data here? Can the two data sources be combined, and if so, why weren't they?

Was the choice of the 7 day maximum for forward prediction driven by the performance of your model? If so, it would be interesting to see how the performance degrades for larger time windows. If not, justification for this choice is needed.

You mention in the introduction that confirmed death counts are likely to be biased downwards. Further discussion of that bias would be helpful. Might it be differential by country? Changing over time? How should stakeholders interpret these predictions, knowing that they likely underestimate true death counts? Should prediction intervals be asymmetric with higher upper bounds to compensate?

Reviewer 1

This paper presents county-level COVID-19 fatality data as a curated dataset as well as producing model-averaged fatality forecasts at the county level. I find this work interesting and timely. And I think the use of model averaging is a good choice. Below I give mainly minor comments as well as some grammatical issues. I'll proceed in the order of the manuscript.

Minor comments:

1. "the confirmed death count is...more reliable". There are a number of issues with the death count as well. I think there are a few options here (potentially beyond the scope, but important). First, there is a real issue of county specific attribution. A local news article from 5/18 described case counts often being incorrectly attributed to the test or hospital county rather than to the address of the patient. Similar issues likely arise with deaths (though admittedly less so). Often, these issues get resolved, but only at a lag. Second, deaths appear significantly undercounted (numerous NYTimes articles). I think it is desirable to discuss the data issues and biases in more detail, as well as the statistical implications of measuring the accuracy of forecasts of reported fatalities as opposed to actual fatalities. Perhaps in Section 2 rather than the intro however.

2. page 3. The other study, reference [33]. It would be nice to have more details of "similar but complementary". What are the positives and negatives of each. Can they be combined?

3. page 7 "using 6 days of data". Similar to the data quality issue above, I wonder the extent to which weekends impact the results. 6 days stuck out to me because it seems that both reported cases and deaths are quite a bit lower on weekends (my current location does not report at all on Sunday, and aggregates the weekend into Monday's release). The models are rolling, so only models with periods longer than 5 would necessarily include weekends in all windows. I think it would be good to discuss this issue somewhere.

4. Equations 2 and 3. I understand the desire to have the "exponential" in your models, but these equations are ugly. Maybe log both sides or else use \exp.

5. Section 3.6. This section seems inordinately long. Maybe just present the specific weights in equation 8 and relegate the generality to the appendix.

6. Section 4.1. In contrast, the description of Conformal Prediction here seems terse. I think the first paragraph could use some more of the details that are in the second paragraph of A.2.

7. p17. I'm pretty sure that the aggregate model reported by the CDC is the average of the ensemble at each quantile. These don't use accuracy to weight the component models. I think the model averaging in this paper is a big improvement and deserves explicit mention here.

Grammar:

1. p2, first sentence "In recent months, the COVID-19 has". I'm not sure "the" is appropriate here.

2. The itemized model list on p.5 has a number of issues. And I'm not sure how useful these generic descriptions are since the detailed ones come later.

3. The sentence that split between the bottom of p.5 and the top of p.7 is hard to understand. I think it's right, but it took me 3 tries to determine what was meant.

4. p16 second paragraph rogue "randomly"

5. Section 5.3. The set notation for the two periods is unnecessarily distracting. Maybe just give the date ranges.

Reviewer 2

I enjoyed reading this paper and am glad to recommend it for publication after a minor revision (see comments below). I appreciate the authors' detailed explanations of methodological decisions and transparency about limitations.

Overall comments:

- There is no mention of weekend data as being often anomalous for covid-19 case and death counts. Given the short term nature of the fitting and prediction windows here, this seems particularly important to mention. It could impact all aspects of your work (model selection, uncertainty calibration, error and coverage.). It should be discussed at least and, if time permits, explored as an added component of the prediction methods, e.g. by having a covariate for weekend day or fitting weekday and weekend time series separately.

- I did not see a script to specifically reproduce the results in the paper (making it fully reproducible). Is this in the covid19-severity-prediction GitHub?

- There are a number of choices made in model-fitting (e.g. number of days to use for fitting, \mu and c in equation (7)) that are progressively selected based on best-performance in the work presented here. Can these be jointly selected as hyperparameters of a given model? How easy would this be to do with the code you shared?

Page-specific comments:

p.1

- Given that the paper methods are not not appropriate for long-term covid forecasting I recommend explicitly putting "short-term" in the paper title.

- "at the county-level in" -> at the county level in

p. 2

- "recent months, the COVID-19 has" -> recent months, COVID-19 has

- "Our goal is to provide access to a large data repository (that combines data collected by a range of different sources) and to provide a predictor to forecast short-term COVID-19 mortality at the county- level in the United States along with uncertainty assessments of our predictors in the form of intervals.” -> (suggested revision) Our goal is to provide access to a large data repository combining data from a range of different sources, and to forecast short-term COVID-19 mortality at the county level in the United States. We also provide uncertainty assessments of our predictors in the form of intervals.

- impose/ease -> impose or ease

- Add citations for this statement “While many other studies focus on predicting the long-term trajectory of COVID-19” and the one below that about national or state-level predictive efforts. (Or maybe just refer earlier to the section on related work.)

- “As a result of researchers across academia and industry collectively refocusing their efforts towards combating this universal viral threat we face, there are now...” -> Researchers across academia and industry collectively refocusing their efforts towards combating this universal viral threat we face. As a result there are now…

p. 3

- I appreciate the visual overview of contributions (Figure 1), and that it recognizes data curation and visualizations as products.

- “It is also updated.“ -> It is updated.

p. 4

- Table 1: it would be good to note which data sources are being regularly updated and which are basically static.

- I was surprised to see that none of the sources on Social Mobility were used in the predictors. Did you consider including them? It seems worth commenting on the choice to leave them out.

p. 5.

- “recent trend” -> recent trends

- Given the predictors that you did include it seems natural to consider a partial pooled predictor. Did you consider this? It is somewhat addressed by including an ensemble of a separate-county and shared-county predictor, but this is selected for its performance. It would be good to discuss explicitly why a balance of separate and shared effects (and potentially spatial dependence) would be useful here and why or why not to include such a model.

p. 7

“over dispersion” -> overdispersion

p. 8

equation (3):

- I found it slightly confusing that this equation is for time t (expectation of deaths_t) as opposed to deaths_{t+k} since it is to predict “number of confirmed deaths k days into the future”, but maybe it’s fine as is.

- Should the deaths _{t_1} term be given the notation of a prediction/expected value, since when predicting death count you iteratively use the daily sequential predictions?

p. 9

equation(4): put \beta_0 first for consistency

p. 12

footnote: “term width and length” -> terms width and length

p. 13

“counties (see Section 5.2 for more discussion on these counties). “ -> counties. (See Section 5.2 for more discussion on these counties.)

P. 14

The CLEP seems to consistently do its worst in the 90th percentile (p90) of MAE. Is there an explanation for this tail?

p. 16

- Is the reason for the uptick in Suffolk County the same as for the other New York counties?

- It seems that many counties in Figure 5, especially in (b), have high uncertainty at the beginning of the time window. Does this reflect the fact that when counts are low the normalized max error is can be high? If so it would be good to say that, or otherwise is there an alternate explanation?

p. 17

- “But to the best of our knowledge, all of these predictions are either made at the country level or the state level, and none of these works have predicted deaths and cases at the county level”. Is this still true?

p. 18

“offer 5-day” -> “of 5-day”

Reviewer 3

The paper combines COVID-19 data across several databases and proposes a predic-tive model for confirmed county-level death counts based on novel predictors proposed. Authors consider separate, shared, expanded shared, demographics shared and linear predictors. Separate and linear predictors build predictions on the past death counts within the county, while shared predictors utilize data across other counties and expanded shared uses data from neighboring counties. The proposed combined predictors model (CLEP) weights different predictors in the model based on past prediction performance. Maximum-absolute-error predictions intervals (MEPI) are proposed for quantifying the uncertainty around the predictions made a week ahead in time.

The authors stress that even though, prediction models have been built for the same outcome before they are not at the county-level, which is the main novelty of the paper. I like the array of predictors considered. Also an effort is made to quantify the uncertainty around the predictions; Section 4.3 nicely explains the intuition behind the building of MEPI.

More specific points are as follows:

1. In the CLEP modeling, why was expanded and linear predictors were the two chosen to be combined? Can't the same weighing framework used to combine other predictions?

2. Regarding Table 3, summarizing model comparison results, the comparison metrics are summarized across all counties in the US. However it is likely that different models may fit different counties better. It seems to me that this is also the reason for the favorable performance of CLEP overall. CLEP allows weights from different predictions to be determined at the county level. Hence it incorporates different predictor/modeling choices at each county (since different models may fit different counties best), leading to more favorable performance when summarized across all counties.

3. However it is not clear to me why the expanded and linear predictors are combined in CLEP. One can imagine for example that based on the growth rate of counts in each county, a linear or a separate predictor may be preferred in modeling for that county's rates. Why can't we combine these predictors in CLEP? Better yet, why can't all the predictors be combined in CLEP with full customization of the predictors at each county (via the weighting of predictions).

4. Even though it may be beyond the scope of this paper to extend the predictors con-sidered to include social distancing practices at the county-level, could the authors elaborate on if such data was available, whether it could have been incorporated into proposed prediction framework. I imagine this variable could cary quite a bit of predictive power.

5. I like the authors explanations on why the confirmed death counts were modeled instead of confirmed cases. In addition, predictions that are only a week into the future also make sense. Finally the motivation behind the MEPI construction is also explained clearly.

Point-by-point response to comments by the referees for the submission “Curating a COVID-19 data repository and forecasting county-level death counts in the United States"

Nick Altieri, Rebecca L Barter, Raaz Dwivedi, James Duncan,
Karl Kumbier, Xiao Li, Robert Netzorg, Briton Park, Chandan Singh
Yan Shuo Tan, Tiffany Tang, Yu Wang, Chao Zhang, Bin Yu

July 16, 2020

The authors would like to thank the three reviewers, the associate editor and the editor-in-chief for their careful read of the manuscript and their elaborate feedback. All of the comments have helped significantly in revising the manuscript. In particular, we appreciate the referees for their detailed major and minor comments, and pointing out the typos (all have been fixed in the revision). In this response, we first summarize our major revisions in Section 1, followed by point-by-point responses to the comments of each referee (in the order EIC, AE, Reviewers 1, 2 and 3) in Sections 2 through 6. We provide response in this order since EIC's comments already summarize the major comments by the reviewers 1, 2 and 3, and the AE in a concise way.

Before we summarize our major changes, we would like to point out two errors made in our first draft, and sincerely apologize for these mistakes. First, we erroneously reported that CLEP was using an absolute loss between the *logarithm* of the counts y_t (Equation 3.7 in the revision). In fact, we had already changed it to the *square-root transform* in our code, to obtain better performance, but forgot to update it in the manuscript. Second the MAEs reported for 5- and 7-day predictions in Table 3 of the first version were inaccurate. While our predictions (pushed to GitHub and the website `https:covidseverity.com`) were correct, we made an error while aggregating them to compute summary statistics. While computing statistics for 5- and 7-day-ahead predictions, we incorrectly pooled 3-day-ahead errors with the 5-day-ahead errors, and 3-day-ahead and 5-day-ahead errors with the 7-day-ahead errors. (We note that the statistics for 3-day-ahead prediction errors were correctly reported in Table 3 of the the first version of the paper.) This mistake in the aggregation code led us to report incorrect statistics of 5-day-ahead and 7-day ahead MAEs, and the reported results in the paper were smaller than the actual MAEs. On correcting the code, we found that the actual raw-scale MAEs for the period March 22-May 10, 2020 got worse as expected; by 16% for 5-day ahead, and 29% for 7-day-ahead (median raw-scale MAEs). We take full responsibilities for these errors and sincerely apologize for them. We have pointed them out in the revision, when describing CLEP weighting scheme (footnote 10 on page 21) and when discussing MAE results (footnote 14 on page 27). The discovery of the errors also made us to take a step back and do a serious vetting of the code and our results. We have now implemented a strict software vetting protocol in Yu-Group for all the ongoing and future research projects.

1 Overview of major changes

Next, we turn to the major and common comments by the referees that can be summarized as follows: (i) add discussion on the possible biases in data, and the likely effects on the predictions of the methods considered in the paper, (ii) analyze effect of social mobility on the predictions, (iii) capture the weekly seasonality observed in the COVID-19 reported case and death counts, (iv) discuss the performance of CLEP and MEPI for more than 7-day-ahead prediction.

Based on the referees's feedback, we significantly revised our discussion on data in Section 2, provided detailed evaluations for CLEP and MEPI for longer horizons (up to 14 days) in Section 5, and discuss the performance with additional features for social distancing and weekday seasonality in Appendix A. Given that the paper has several results, we added a table of contents, and list of tables and figures in the beginning of the paper. The major changes can be summarized broadly as follows:

(i) **Detailed discussion of data sources, biases and potential effects**: We have expanded our discussion of the data sources by adding a new Section (Section 2.1), another table that collects the important features together (potentially useful for other researchers), and providing a detailed comparison with the repository by a John Hopkins group [3].[1] We added Section 2.2 to provide a detailed discussion of the differences between the USA Facts and NY Times datasets, and potential biases in the datasets such as the underestimation of deaths and cases, weekday patterns, and the historical revisions of the case and death counts and how they affect our models.[2]

(ii) **Modeling social mobility:** On receiving these reviews, we experimented with using an indicator features for social mobility: (a) Whether or not social distancing has been implemented in a county on a given day, and (b) whether or not social distancing has been implemented for at least 1 week or at least 2 weeks till a given day. However, none of these improved the aggregate performance of the CLEP method over the days considered in our experiments from 3/22 to 6/21. While we do not discard the possibility that a better featurization may lead to improvements in our predictions, such an investigation remains beyond the scope of the current paper. In the revision, we provide a detailed summary of results for 7-day-ahead predictions, which include the two week indicator feature since implementation social distancing feature in Appendix A.1 and Figure 13.

(iii) **Modeling weekday patterns:** We did observe a difference in the patterns of recorded death counts between weekdays and weekends, where during weekends the new death counts added were generally lower (Section 2.2, Figure 3(a)). To adjust for this trend, we added an indicator feature that took value 1 if the predicted day was a Sunday or a Monday, the days on which deaths were most underreported, and 0 otherwise. We added this feature to both the expanded shared model and the separate linear model and we evaluated it over the 3-day-ahead predictions for cumulative death counts. However, this adjustment did not lead to any improvements. Once again, we provide details in Appendix A.2, and Figures 14 and 15 of

[1] For details on comparison with this repo, please refer to our response to Comment #1 by AE in Section 3.

[2] For further discussion on the issue of data-bias, please refer to our response to EIC's Comment # 1.

the revised manuscript; and further investigation of what may be a better featurization is left for future work.

(iv) **Prediction for longer horizons**: In our first submission, we reported results for a prediction horizon of up to 7 days partly because of our original goal to provide *reliable* predictions (which can be validated) to help the non-profit for their PPE resource allocation task. In the revision, we show that CLEP and MEPIs both have reasonable performance for up to 14 days, with linear degradation of performance with prediction horizon; the error for 14-day-ahead predictions is generally twice that of 7-day-ahead predictions. We have added several new results for CLEP and MEPI performance in Section 5. Please refer to Table 4, Figure 7 and 8 for details on variation of CLEP MAEs with prediction horizon, and Figures 9, 10 for county-level visualizations, and Figures 11 and 12 for comparing MEPI performance for 7-day-ahead and 14-day-ahead predictions. (Partial results are reported up to 21-day-ahead predictions in Figure 8d-f).

(v) **Comparison with other works:** During the revision of our paper, we became aware of the recent work [2] that appeared on medRxiv on June 8, 2020 that made county-level predictions using Hawkes' process based models. The authors also provided a comparison of their results with that of CLEP. There were several differences to begin with: they used NYTimes data, modelled new counts for selected counties for a fixed period till mid May, and provided results for selected counties stratified in different quantile groups in their paper. Their GitHub Repo did not provide either the predictions for the period reported in the paper or for the more recent periods. Since their codes were in Matlab, we decided to first reproduce their results for CLEP by following their paper as much as possible but were unable to do so due to the lack of a good documentation of all the choices they made. (Our results were generally better than what they reported). Via private email exchanges with the authors, we did get some clarifications, but several choices remained puzzling to us from a practical point of view. To summarize: (i) They provide predictions only in blocks of, say 7 days, meaning that to get the prediction for the entire next week after we will have to wait till the end of the current week. (ii) Results are reported for counties grouped by quantiles based on the end of evaluation period (Table 1) and not for all the counties. (iii) The % errors (Table S1) are reported over the total counts over all counties over the in top quantile groups (a practically non-interpretable group). (iv) They do not report any performance results for their confidence intervals. Nonetheless, they implemented an adaptively tuned CLEP where the hyperparameters c and μ in equation 8 are tuned over a grid over time. This aspect was not mentioned in the paper and became clear to us after the email exchange with the authors. We did find this tuning to be a promising direction to obtain better performance with CLEP even in the context of our work. However, we leave a detailed investigation for adaptive tuning of hyperparameters for our future work. In the revision, we have discussed this paper [2] and our concerns at the end of Section 6 (related work).

(vi) **Monotonicity of predictions:** Finally, during the revision of the manuscript, we realized that we should do a post-hoc adjustment to ensure that our predictions are always greater than the last observed death count (something which we implemented for MEPI but not CLEP in the first version(, and that predictions made on a given day were non-decreasing with horizon. In fact, we found out that for certain counties, on certain days, expanded shared predictor would make predictions that were decreasing with horizon). We have now used maxima calculations for the individual predictors (i.e., replacing $\widehat{y}_{t+k}$ by $\max\{\widehat{y}_{t+k}, \widehat{y}_{t+k-1}, y_t\}$) to implement this, and provided the description in Section 3.7 (page 21) of the revision. We note that such an implementation of *monotonicity* improved the overall results both for predictions and prediction intervals. The median % error for the period March 22 to June 20 reduced from 15.29% to 15.14% for 7-day-ahead predictions, and 28.49% to 26.45% for 14-day-ahead predictions. The coverage of MEPI increased in general—by up to 5% for the cases when the original coverage was in the range 80-90%, and more modestly, by upto 2%, when the coverage was in the range 90-100%.

2 Comments by the EIC

Besides the summary of comments, EIC also suggested a few formatting guidelines, namely, (i) providing upto 6 keywords, (ii) media summary, and (iii) following APA citation style, all of which have been accounted for in the revised manuscript (thanks to the overleaf project link shared by the EIC).

Comment #1. *"There are still three things we can do: (A) Be as transparent as possible about the sources of the biases (B) Mitigate those biases that can be modelled/captured (such as time lag handled in the paper) (C) For biases that we cannot adjust, discuss their likely directions, magnitudes and impact"*

Response: We thank the reviewers, AE and EIC for highlighting the aspect of data biases. To address these concerns, we have added a new Section 2.2, where we discuss in detail the biases that we observed in the data. For the biases that have potential to be dealt with in the modeling stage such as weekday/weekend patterns, we did subsequent analysis trying to mitigate those biases and report our findings. For other biases that cannot be mitigated in the modeling, we added more exploratory data analysis and analyzed their scale as well as the possible directions. An important potential bias is the under-reporting of death counts. In a nutshell, while there are definitely under-reported death counts, starting from April 15, the data sources we use started to include the probable deaths due to COVID-19 along with the reported deaths though these numbers are reported as a sum and not separately. So while the problem of under-reporting does exist, there are already efforts on the data provider's side that try to mitigate this. A natural effect of these biases on our predictions are that our predictions are likely to under-predict the true (unknown and unobserved) death counts.

We remark that sometimes the adjustment of probable and unreported COVID-19 related deaths from the past leads to a sudden uptick in recorded death counts on a given day. Since our models get updated on a daily basis, such an uptick leads to a (not unexpected) sudden over-shoot in our predictions for the next few days. These observations were already reported in the previous version of manuscript, and the discussion has been further elaborated in the revision. Moreover, as a proof of concept, we also checked that if we manually remove such an uptick in the death counts (consider a slightly modified adjusted training data), our predictions do not over-shoot and continue to perform well. We discuss these aspects in Sections 5.1 (page 28), Section 5.2 (page 31), and Appendix B.1 (Figure 16).

Comment #2. *..I think you need to be more explicit on how you tested the models and made various choices (such as AE's question of how you choose 7 days) to give readers more confidence in your predictions...*

Response: In the revision, we have added new results for up to 14 days-ahead predictions and for the longer evaluation period March 22 to June 20. Our results illustrate the good performance of CLEP and MEPI for this period. The performance gets worse with longer horizon in a linear manner (14-day-ahead performance is generally twice as worse as 7-day-ahead performance). Our first submission was done May 15, 2020 where we had presented results for the period March 22 to May 10. We also report MEPI performance separately for the period May 11 to June 20 (in Figures 11 and 12) as another set of evidence for good out-of-sample performance for our methods (since our first submission, we made minor changes to code to ensure monotonicity besides vetting it). With these extensive evaluations and indirect out-of-sample validation for the period May 11 to June 20, we hope that readers will be confident about the predictions made by our methods. Please also refer to the overview bullet (iv) in Section 1. Moreover, at our website `https://covidseverity.com`, we provide daily 7-day predictions with histories of CLEP's performance for all the counties in the US.

3 Comments by the AE

We thank the AE for their comments. They asked us to discuss the relevance of the paper by Angelopoulos *et al* [1]: the goals/contributions of that paper are not directly related to our work; but as both the AE and EIC pointed out, we found their discussion on data biases helpful. We cite them accordingly in beginning of our discussion of data biases in Section 2.2 (page 9). Next, we provide detailed responses to other major comments by the AE.

Comment #1. *"Can you say more about how the data in reference [3] complement the data here? Can the two data sources be combined, and if so, why weren't they?"*

Response: At the county-level, both our data repository and the repository by the John Hopkins group [3] generally provide similar types of information (i.e., demographics, socioeconomic information, social mobility, daily COVID-19 cases and death counts). Some of our data sources overlap (e.g., ICU beds data from Kaiser Health News, poverty data from USDA) while others (e.g., the demographics, socioeconomic, and COVID-19 cases/deaths data) come from slightly different sources. The differences, in most cases, are generally minor. A couple of main differences between the two repositories are as follows. First, in addition to the county-level data set, our repository features a hospital-level data set while that of [3] provides only a secondary state-level data set. Second, our data repository contains county-level data on COVID-19 health risk factors such as the prevalence of various underlying health conditions, which is not present in [3] county-level data set. On the other hand, [3] manually curated a data set with the dates of government orders and interventions, which we merged into our repository (see, the dataset JHU date of intervention data in Table 3 of the revised manuscript). We have added a more detailed explanation of the similarities and differences between our repository and that in [3] in the last paragraph in the (new) Section 2.1 (page 7).

Our main reservation for not fully merging the two data repositories however is the lack of details about data cleaning and processing steps taken by [3]. Ideally, we would like our repository to be as transparent as possible. Consequently, in many cases, we directly pulled the data from the original source (e.g., CDC), and not the processed one from [3] due to lack of data-processing documentation for the latter.

Comment #2. *"Was the choice of the 7 day maximum for forward prediction driven by the performance of your model? If so, it would be interesting to see how the performance degrades for larger time windows. If not, justification for this choice is needed. "*
Response: The choice of 7-day horizon was made earlier to provide *accurate* predictions to the nonprofit Response4Life which have *very good out-of-sample validation error*. In the beginning, there was not enough data to really build/validate longer term predictions. However, in the revision, we have added new results with longer prediction horizons without any tuning (other than imposing monotonicity) with CLEP and MEPIs—complete results up to 14 days and partial results up to 21 days in Table 4, Figures 6-12 in Section 5. Overall, the performance of our model remains reasonable up to 14-days-horizon; and the degradation of MAE is roughly linear with the prediction-horizon. Please also refer to our response to the EIC's comment #2.

Comment #3. *"You mention in the introduction that confirmed death counts are likely to be biased downwards. Further discussion of that bias would be helpful. Might it be differential by country? Changing over time? How should stakeholders interpret these predictions, knowing that they likely underestimate true death counts? Should prediction intervals be asymmetric with higher upper bounds to compensate?"*
Response: You raise a valid point, the bias may indeed vary by county and appropriate adjustments would be needed. The amount of underestimation is a problem currently under investigation, and some of our data sources try to account for the same by revising data (an aspect we discuss in Section 2.2 of our revised manuscript). However, given the (i) the lack of a consensus on how much underestimation exists, and that (ii) a thorough investigation of concurrent news articles or research pre-prints is out of the scope of our current work, it is not clear to us how much increase in the upper bounds of the MEPIs (and county-wise variation of the adjustment) would be enough. Please also refer to our response to the EIC's comment #1.

4 Comments by the Reviewer 1

We thank the reviewer for their comments, in particular the one asking to incorporate the weekend effect. We appreciate the reviewer's note that our ensembling is based on performance and that we should mention it explicitly. We also thank the reviewer for pointing grammatical errors and typos, all of which have been fixed in the revision. As we summarize below, the major comments of the reviewer have already been discussed in the prior sections.

Comment #1. *"...I think it is desirable to discuss the data issues and biases in more detail, as well as the statistical implications of measuring the accuracy of forecasts of reported fatalities as opposed to actual fatalities. Perhaps in Section 2 rather than the intro however..."*
Response: We agree and we have now added informations on data issues and biases in Section 2.1 and 2.2. For revision details, please refer to our response to the EIC's Comment #1.

Comment #2. *"The other study, reference [3]. It would be nice to have more details of "similar but complementary". What are the positives and negatives of each. Can they be combined? "*
Response: We have now added information on comparisons to data repository [3] in Section 2.1 last paragraph on page 7. For details, please refer to our response to the AE's Comment #1.

Comment #3. *".. I wonder the extent to which weekends impact the results..I think it would be good to discuss this issue somewhere.."*
Response: Great point. We have investigated the weekend effects. We added an indicator feature but did not obtain any improvements (please see Appendix A.2, Figures 14 and 15 in the revision). For a detailed summary, please refer to the overview bullet (iii) in Section 1.

5 Comments by the Reviewer 2

We thank the reviewer for their several insightful comments. We want to extend special thanks for the careful list of typos and very useful suggestions for rephrasing sentences for increasing readability. We appreciate the reviewer's note that the paper was an enjoyable read and that they found our detailed explanations and transparency useful. The reviewer asked for the link to the script to reproduce the results; `https://github.com/Yu-Group/covid19-severity-prediction/tree/master/modeling` (also mentioned in the beginning of Section 5 in the revised paper). We now turn to our responses to other major comments made by the reviewer.

Comment #1. *"...There is no mention of weekend data as being often anomalous for covid-19 case and death counts...It should be discussed at least and, if time permits, explored as an added component of the prediction methods, e.g. by having a covariate for weekend day or fitting weekday and weekend time series separately.."*
Response: Great point. Other referees also asked for this effect. We have investigated them in detail in Appendix A.2, Figures 14 and 15 of the revision. For a detailed summary, please refer to the overview bullet (iii) in Section 1.

Comment #2. *"There are a number of choices made in model-fitting (e.g. number of days to use for fitting, μ and c in equation (7)) that are progressively selected based on best-performance in the work presented here. Can these be jointly selected as hyperparameters of a given model? How easy would this be to do with the code you shared? "*

Response: We agree that the hyperparameters in the CLEP can be tuned jointly and such a tuning if done right (using train, validation and test set) is likely to improve the results. In our current work, we took a step-by-step tuning approach and provided reasonable justifications for our choices. For the revision, in the interest of time and for consistency, we decided not to re-tune the hyperparameters—except the change of transform from logarithm to square-root in the weight definition. For a user, minor changes to the code would enable the joint tuning of the hyperparameters: (a) the decaying exponent μ (we chose 0.5 in equation 6), (b) the lag used for predictions (we found 3-day ahead predictions worked pretty well), (c) number of days for computing the error (we took 7 day window in equation (6)), (d) the transformation of the predictions and death counts (we tried linear, square-root and logarithm and stuck with square-root in the revision, other choices like α-th root are also possible), and (e) number of predictors (we tried combinations of all 5 and top 2 models). To implement this tuning, a few (easy to do) minor changes would be needed in our publicly available code. We have highlighted that these aspects more clearly in the last paragraph of Section 3.6 of page 19.

It is worth noting that one may also consider an adaptive tuning of all hyper-parameters over time; this adaptation would need some more work and a few more careful considerations. While we originally did not tune our hyper-parameters in this manner, the authors of the recent work [2] (in private email exchange) mentioned that they had implemented CLEP with an adaptive tuning of c and μ (hyperparameters in the weighting scheme given by equation 8). In our own future work, we have started investigating the adaptive tuning over time of the several hyper-parameters for improving CLEP performance. We highlight this new possibility at the end of Section 6 and at the end of the paper.

Comment #3. *".. I was surprised to see that none of the sources on Social Mobility were used in the predictors. Did you consider including them? It seems worth commenting on the choice to leave them out..."*
Response: We tried using social distance in the revision[3], but unfortunately it did not help in the way we used it (please see Appendix A.1 and Figure 13). For more details, please refer to the overview bullet (ii) in Section 1.

Comment #4. *".. But to the best of our knowledge all of these predictions are either made at the country level or the state level, and none of these works have predicted deaths and cases at the county level. Is this still true?t..."*
Response: Thank you for asking this question. During the revision of our paper, we became aware of the recent work [2] based on Hawkes processes that provided predictions at county-level. But we found their results to be generally not reproducible besides the several puzzling choices made in that paper for reporting performance (even after private email exchanges). For further discussion about our concerns, please refer to the overview bullet (v) in Section 1. We have also added a detailed discussion in Section 6 of the revised paper.

Comment #5. We now collect several other helpful comments made by the reviewer:

(i) *"Table 1: it would be good to note which data sources are being regularly updated and which are basically static."*

(ii) *"Given the predictors that you did include it seems natural to consider a partial pooled predictor. Did you consider this?..would be useful here and why or why not to include such a model."*

(iii) *"The CLEP seems to consistently do its worst in the 90th percentile (p90) of MAE. Is there an explanation for this tail?"*

[3] In the earlier stage of this project social distancing was beginning to be implemented and thus those aspects were not investigated too much.

(iv) "Is the reason for the uptick in Suffolk County the same as for the other New York counties?"

(v) "It seems that many counties in Figure 5, especially in (b), have high uncertainty at the beginning of the time window. Does this reflect the fact that when counts are low the normalized max error is can be high? If so it would be good to say that, or otherwise is there an alternate explanation?"

Response:

(i) We have added a flag (†) to denote the data sets in Table 2 which are updated daily. All other data sets are basically static.

(ii) We have considered pooling our shared models over subset of counties. For example, instead of having a national model, having models pooled over states, pooled over neighboring counties, or pooled over counties that are similar in some way. We think this direction is very promising for future work. We have not addressed it in the current work, due to the careful considerations needed to determine the best ways to pool the counties together.

(iii) The 90th percentile of MAE reflects the predictive accuracy of the predictors on its worst days. Figure 6 (in the revision) shows that the worst days for the 7-day-ahead CLEP were around 4/19. Indeed, around those days, the extended shared predictor has particularly large MAEs due to the data correction on 4/14, and the CLEP also has large MAEs because the extended shared predictor is one of its component. The linear predictor has smaller MAE on those days because it fits a linear curve to deaths counts and its predictions are thus more robust to sudden uptick in reported deaths. We have elaborated to this end in Section 5.1.

(iv) Indeed. The uptick of Suffolk county happened on May 7, when 227 fatalities were added. According to news report (e.g., https://riverheadlocal.com/2020/05/07/bellone-spike-in-suffolk-fatalities-attributed-to-presumed-covid-deaths/), most of them were previously unreported "presumed" deaths.

(v) This intuition is exactly correct. When the counts are low, the normalized error can be high. We mention this intuition in our discussion on coverage in Section 5.3 (pages 34-35).

6 Comments by the Reviewer 3

We appreciate the reviewer's positive comments on the manuscript and the remark that the intuition behind MEPI was well explained. Please find below answers to their other major comments.

Comment #1. *"In the CLEP modeling, why was expanded and linear predictors were the two chosen to be combined? Can't the same weighing framework used to combine other predictions?...However it is not clear to me why the expanded and linear predictors are combined in CLEP. One can imagine for example that based on the growth rate of counts in each county, a linear or a separate predictor may be preferred in modeling for that county's rates. Why can't we combine these predictors in CLEP? Better yet, why can't all the predictors be combined in CLEP with full customization of the predictors at each county (via the weighting of predictions)."*

Response: Indeed, we attempted to combine all models. However, this combination performed worse than using our two best models. It is also important to note that poor models will receive some weight and many correlated poor models will negatively impact performance. As a result, even with our adaptive combination (CLEP) scheme, some tuning of which models to pick can be useful. We explain this aspect in the first paragraph of Section 5 on page 25.

Comment #2. *"Regarding Table 3.... However it is likely that different models may fit different counties better. It seems to me that this is also the reason for the favorable performance of CLEP overall. CLEP allows weights from different predictions to be determined at the county level. Hence it incorporates different predictor/modeling choices at each county (since different models may fit different counties best), leading to more favorable performance when summarized across all counties."*
Response: The reviewer's note matches our evaluation (Figure 6 and Table 4), and the reasoning behind county-wise ensembling. One may even consider fitting the parameters of shared model at state or some super-county (pool of several counties in a region) to improve performance. But given that such a finer fitting would require a fair amount of tuning and further justification, we leave this direction as an interesting future direction. Please also refer to our response to your previous comment.

Comment #3. *"Even though it may be beyond the scope of this paper to extend the predictors con-sidered to include social distancing practices at the county-level, could the authors elaborate on if such data was available, whether it could have been incorporated into proposed prediction framework. I imagine this variable could carry quite a bit of predictive power."*
Response: We thought so too and did try using social distance info, but unfortunately it didn't help with prediction in the way we used the feature (Appendix A.1). For a detailed response, please refer to the overview bullet (ii) in Section 1.

References

[1] A. N. Angelopoulos, R. Pathak, R. Varma, and M. I. Jordan. On identifying and mitigating bias in the estimation of the COVID-19 case fatality rate. *Harvard Data Science Review*, 6 2020. https://hdsr.mitpress.mit.edu/pub/y9vc2u36.

[2] W.-H. Chiang, X. Liu, and G. Mohler. Hawkes process modeling of COVID-19 with mobility leading indicators and spatial covariates. *medRxiv preprint 2020.06.06.20124149*, 2020.

[3] B. D. Killeen, J. Y. Wu, K. Shah, A. Zapaishchykova, P. Nikutta, A. Tamhane, S. Chakraborty, J. Wei, T. Gao, M. Thies, and M. Unberath. A county-level dataset for informing the United States' response to COVID-19. *arXiv preprint arXiv:2004.00756*, 2020.

Second Review

First revision submitted on August 3, 2020
Second reviews received on August 25, 2020
Second revision submitted on September 12, 2020

AE

I recommend accepting the paper now and strongly encouraging the authors to address Reviewer 2's suggestions [on the revised manuscript] ahead of publication. I recommend accepting the paper now and strongly encouraging the authors to address Reviewer 2's suggestions ahead of publication.

Reviewer 2’s comments on the revised manuscript

I thank the authors for their detailed revision and response letter. I also appreciate the additions to the text that encourage reproducibility, e.g. the link to the reproduction code and footnote 8. I again recommend this paper for publication pending some minor revisions listed below. Note: a few of the comments are not specific to this version, but only occurred to me on second reading.

Overall comments:

- Model notation is clearer now, especially with respect to time. However, the procedure around k-day ahead prediction is still complicated. I made one sentence suggestion (see comment below for p. 18) to hopefully clarify further. I also wonder if overall it would be simpler to fit k-day ahead predictions in the same way you have fit 1-day ahead predictions to avoid recursive calculations and the mismatch of using predicted death and (earlier) observed covariants like neighboring-county case numbers. This is not a necessary addition to this paper at this stage, but something to comment on or for future work.

- You write that the max confidence intervals approximate 95% given that it they are constructed using only five data points. Does it really approximate 95% or would it be more accurate to say it approximates 100%? Would it be appropriate to approximate a 95% interval (or other %) by an interpolation between the max and second-highest error? Although the authors are clear that the assumption of exchangeability does not hold here, the amount of precision that can be expected is left a bit vague.

- Regarding my previous comment about a partially pooled predictor: The directions you mentioned in your response (e.g. pooling over states, neighboring counties or similar counties) are possible, but I actually had in mind a hierarchical modeling approach, possibly with spatial dependence. I am still wondering if you considered this and think it is worth mentioning in the paper. I also reiterate part of that earlier comment that it would be good to mention that your ensembling approach settles on a balance of separate and shared effects models, which makes intuitive sense (given the presence of county and country-wide patterns in the data). I still think it would be good to mention this explicitly, as it may guide future work by you or your readers.

- It was surprising to see that the attempt to adjust for the clear weekday bias did not improve results and I would like to see a little it more investigation (but it can stay brief and in the Appendix). A couple thoughts: 1) Having added the Sunday/Monday feature, it seems appropriate to also add a Tuesday/Wednesday features since these days tend to compensate in the opposite direction. The Tu/We feature could possibly have negative \beta^c_2 as its coefficient to reinforce the balance in bias and prevent overfitting. 2) Should the weekday feature only be applied to some counties?

- Ideally the paper would not be so much longer than the first version. Below I’ve mentioned a couple places where I think content can be moved to the repository. The authors are encouraged to move additional details to the repository where possible. I also don't think the added table of contents and list of figures and tables is necessary, but it's fine to keep them.

Page-specific comments:

p.6. I associate "social mobility" more with class mobility than human mobility and would probably just refer to this as "mobility," but I leave it to the authors.

p. 7 Given the length of the paper, I recommend moving the "Other potential use-cases for our repository" paragraph to the repository.

p. 11 Also to reduce length, I suggest keeping tables 1 and 2 and moving 3 and 4 to the data section of the repo (where not superfluous). You could then mark in tables 1 and 2 the features used in your predictors.

p. 11 "Although the probable death counts address imperfect reporting and attribution, it is unclear to what extent the problem is mitigated." Explain — do you think they're still biased in a specific direction?

p.12 (figures)

- Figure 2 seems to have significant overplotting. Is there a better way to visualize this data? Maybe just reduce the alpha?

- Add a line to 3a (and similar figures later) to indicate a hypothetical uniform distribution (to make it easier to see deviation from it)

- Would 3b be better as a table (or in text) listing key quantiles of days to revision?

p. 17 re: overdispersion/negative binomial dist. Why did you only consider values of the inverse-scale parameter in the given list? (Also remove the hyphen from "over-dispersion".)

p. 18 Re: "To predict k-days-ahead cumulative death count…" It seems that you could add a recursive formula here which would help describe the shape of the +k day prediction.

p. 18 Add a clause for clarity: "However, we only use the new features (cases in the current county, cases in neighboring counties, and deaths in neighboring counties) up to the end of day t − k + 1." -> However, we only use the new features (cases in the current county, cases in neighboring counties, and deaths in neighboring counties) up to the end of day t − k + 1, since when predicting deaths^c_t+k this covariate information will only be available up to k days before.

p. 26 (Possibly) Add confidence bands to the rank distribution plot in Figure 5 (and Figures 17 and 18) assuming complete exchangeability. Did you avoid this because complete exchangeability is not met?

p. 27. It is surprising to me that the demographics shared predictors has such poor performance, and seems so overfit, as you say. You mention experimenting with l1 and l2 regularization for this predictor but not ultimately using it. Given the results though, I wonder why it wasn't used/clearly beneficial. Please comment on this. Do you think this would change when joint and/or adaptive tuning of hyperparameters is implemented?

p. 28. Figure 6 raised the question for me of how the linear and expanded shared predictor weight relatively in the CLEP (knowing this changes over time), since it seems the CLEP is much more similar to the linear predictor curve. This seems worth noting at some point.

p. 31 I wouldn't refer to this as "additional out-of-sample validation" since the predictors are fit to the new data. Rather just "additional validation""

p. 34-5, re:normalized length: Based on your experience using your predictions with Response4Life, is there a normalized width that the MEPIs need to stay under to be useful in practice? If so please mention this to add to the interpretability of the length results.

p. 35 "(1) the predictors mostly make strong assumptions and typically do not involve data-fitting." Does this really mean that the predictors aren't fit to covid-19 data? That's very surprising so I wanted to check if this needs clarification.

p. 39 You could also note here (or earlier) that large errors in your predictions could be a tool for finding data anomalies, reporting errors, etc. which are a growing concern.

p. 41-2. Figures 14 and 15 would be easier to interpret if you combined each (a) and (b) into a single grouped bar chart, with different bar coloring for with and without weekday.

A few typos, grammar etc. (not a complete list — the paper should go through another round of proofreading)

p. 2 rephrase: "are accurate for 7-day into the future and decent for 14-day into the future."

p. 4-5
- Style note: there are a lot of parenthetical comments that don't always need to be in parens, e.g. "such as whether or not to impose or ease lock-downs (or whether to reopen)."
- "e.g., due to policy changes, or behavioral changes in the society." -> e.g., due to policy change or behavioral changes in society."

- "and at website"-> and on our website

p. 12 Introduce the acronym EDA when first used.

p. 15 "Although, eventually the predictors (3)-(5) reported in this paper did not use any explicit l_1 or l_2 regularization, " -> Although the predictors (3)-(5) reported in this paper did not ultimately use any…

p. 16 "the recorded count deaths" -> the recorded death count

p. 17 Appendix 14?

p. 19 rephrase: "finally obtain the compute"

p. 31 "MEPIs are pretty wide" -> More precise term than "pretty wide"

p. 48 Move the last two figures above the references

Final response to comments by the referees for the submission “Curating a COVID-19 data repository and forecasting county-level death counts in the United States"

Nick Altieri, Rebecca L Barter, Raaz Dwivedi, James Duncan,
Karl Kumbier, Xiao Li, Robert Netzorg, Briton Park, Chandan Singh
Yan Shuo Tan, Tiffany Tang, Yu Wang, Chao Zhang, Bin Yu

August 31, 2020

The authors would like to thank the AE, and the reviewers for their comments on the revised manuscript. Since only reviewer 2 provided detailed comments, we provide detailed responses to their comments below. We have taken special care to fix all the typos and grammatical errors in the final revision. Besides their comments on some figures have also been addressed in the revision. The reviewer also suggested to move certain tables, and discussion from Section 2 to the appendix—we decided to keep them in the main paper to emphasize the role of datasets and curation part in a real-world project, and to remain consistent with the title of the paper.

Comment #1. *“...Model notation is clearer now, especially with respect to time. However, the procedure around k-day ahead prediction is still complicated. I made one sentence suggestion (see comment below for p. 18) to hopefully clarify further. I also wonder if overall it would be simpler to fit k-day ahead predictions in the same way you have fit 1-day ahead predictions to avoid recursive calculations and the mismatch of using predicted death and (earlier) observed covariants like neighboring-county case numbers. This is not a necessary addition to this paper at this stage, but something to comment on or for future work. ..”*
Response: We indeed tried to directly predict the k-day-ahead cumulative death counts instead of doing the recursive computation, but we found this performed worse, perhaps due to the greater variability of outcomes from inputs.

Comment #2. *“ You write that the max confidence intervals approximate 95% given that it they are constructed using only five data points. Does it really approximate 95% or would it be more accurate to say it approximates 100%? Would it be appropriate to approximate a 95% interval (or other %) by an interpolation between the max and second-highest error? Although the authors are clear that the assumption of exchangeability does not hold here, the amount of precision that can be expected is left a bit vague. ”*

Response: We agree that our phrasing was heuristic-based and may have been confusing, thanks for pointing it out. We have rephrased it to as follows: “Since we only use five time points, to construct the interval, we opt for the more conservative choice of simply taking the maximum—or the 100th percentile—of the five errors. Moreover, note that 95th percentile is not well defined for five data points."

Comment #3. *".. Regarding my previous comment about a partially pooled predictor: The directions you mentioned in your response (e.g. pooling over states, neighboring counties or similar counties) are possible, but I actually had in mind a hierarchical modeling approach, possibly with spatial dependence. I am still wondering if you considered this and think it is worth mentioning in the paper. I also reiterate part of that earlier comment that it would be good to mention that your ensembling approach settles on a balance of separate and shared effects models, which makes intuitive sense (given the presence of county and country-wide patterns in the data). I still think it would be good to mention this explicitly, as it may guide future work by you or your readers. ..."*

Response: Thank you for the suggestion. Indeed, this provides an alternative approach, possibly computationally more involved and more expensive, to take into account the spatial dependence among counties (it is worth noting that we have taken into account some spatial dependence in using the neighbor-county total case and death counts). However, since such a modeling strategy is significantly different from what we pursued, given our limited time and energy on the project right now, we are afraid that we are not able to do a different analysis at this point. Hopefully colleagues who are more experienced with such modelling approaches will pursue this direction and compare with the CLEP approach in our paper.

Comment #4. *".. It was surprising to see that the attempt to adjust for the clear weekday bias did not improve results and I would like to see a little it more investigation (but it can stay brief and in the Appendix). A couple thoughts: 1) Having added the Sunday/Monday feature, it seems appropriate to also add a Tuesday/Wednesday features since these days tend to compensate in the opposite direction. The Tu/We feature could possibly have negative β_2^c as its coefficient to reinforce the balance in bias and prevent overfitting. 2) Should the weekday feature only be applied to some counties?..."*

Response: Thank you for asking this question. We investigated adding in addition to a Sunday/Monday feature, a Tuesday/Wednesday feature. While the coefficients we saw were intuitive, a negative coefficient for the Sunday/Monday feature and a positive one for the Tuesday/Wednesday feature, we did not see an improvement in performance. It is possible that some counties may have a relatively worse over or underreporting of deaths. We include in appendix A.2 a brief comment regarding this "In addition to this Sunday/Monday feature, we incorporated a feature to account for the overcounting of deaths on Tuesday and Wednesday. However, we did not see any improvement in results from initial experiments, and thereby omit further details here."

Comment #5. We now collect several other helpful comments made by the reviewer:

(i) *"p. 17 re: overdispersion/negative binomial dist. Why did you only consider values of the inverse-scale parameter in the given list?"*

(ii) *"p. 26 (Possibly) Add confidence bands to the rank distribution plot in Figure 5 (and Figures 17 and 18) assuming complete exchangeability. Did you avoid this because complete exchangeability is not met?"*

(iii) *"p. 27 It is surprising to me that the demographics shared predictors has such poor performance, and seems so overfit, as you say. You mention experimenting with l1 and l2 regularization for this predictor but not ultimately using it. Given the results though, I*

wonder why it wasn't used/clearly beneficial. Please comment on this. Do you think this would change when joint and/or adaptive tuning of hyperparameters is implemented?"

(iv) *"p. 28. Figure 6 raised the question for me of how the linear and expanded shared predictor weight relatively in the CLEP...."*

(v) *"p. 34-5, re:normalized length: Based on your experience using your predictions with Response4Life, is there a normalized width that the MEPIs need to stay under to be useful in practice? If so please mention this to add to the interpretability of the length results."*

(vi) *"p. 35 "(1) the predictors mostly make strong assumptions and typically do not involve data-fitting." Does this really mean that the predictors aren't fit to covid-19 data? That's very surprising so I wanted to check if this needs clarification."*

(vii) *"p. 39 You could also note here (or earlier) that large errors in your predictions could be a tool for finding data anomalies, reporting errors, etc. which are a growing concern."*

Response:

(i) The documentation for negative Binomial regression [1] says that "Permissible values are usually assumed to be between .01 and 2.", and hence we tried the values 0.05, 0.15, 1 to cover a wide range of this permissible range. We have added footnote 10 to clarify this point. Since we did not observe any improvements over the Poisson model, we did not explore further in this direction.

(ii) Thank you for the suggestion. We have replaced the plots of average rank of the errors with a set of heatmaps that show the number of days for which each of the six errors are ranked 1 through 6. See the revised figures 5, 19 and 20 for details.

(iii) Thank you for the suggestion. We indeed found that regularization improves performance slightly, sometimes upto 15% (for 7-day median MAE) in comparison to this predictor without any regularization. However the qualitative conclusions do not change as overall the errors are still larger than the expanded shared and linear predictors (and on certain days there are huge errors). Nonetheless, we have included our improved results in Table 6 and added the following brief comment in section 3.4 "We found that regularization was helpful in addressing over-fitting in this predictor and found that an ℓ_1-penalized Poisson regression with a penalty of 0.5 performed the best."

(iv) We have added Figure 7, which shows the weight of the linear predictor in the CLEP over time for select counties. We also added the following comment to Section 5.1: "We found that for counties with large number of cumulative deaths, the prediction of the CLEP has become much more similar to the prediction of the linear predictor in late May and June. For example, for the six worst affected counties on June 20 (panel (a)), the average weight of the linear predictor in the CLEP is larger than 0.91 from May 17 to June 20. In contrast, the average weight of linear predictor of these six counties is less than 0.5 from March 23 to March 31." We also added another Figure 8 that shows a US wide visualization of weights for the linear predictor for two timepoints: April 1 and June 10.

[1] https://www.statsmodels.org/stable/generated/statsmodels.genmod.families.family.NegativeBinomial.html

(v) The intervals were developed to provide a ball-park estimate for the uncertainty. Our estimates were used in a qualitative sense for the R4L collaboration via the severity index (See Section 7 for details) and we did not have a concrete baseline for MEPI lengths at hand. We did not make any changes in the paper towards this end.

(vi) This statement was valid for very early projects on COVID forecasts where the forecasts were made on a few weeks of data for several months in future under strong modeling assumptions without much validation check, e.g., the first couple of predictions released by the IHME group, or the agent-based simulations provided by the Cambridge group. However, this comment is probably outdated given the surge in number of COVID-19 works since then, and thereby has been removed from the final version of the paper.

(vii) Thanks for the recommendation, we have mentioned the following as the footnote 17 at the end of Section 5.2 in the paper:“These observations also suggest that one may possibly use large errors from our predictors as a warning flag for anomaly or reporting error/sudden revision in the data. We leave a detailed investigation on this aspect as an interesting future direction."

Dataviz Review

Second revision submitted on September 12, 2020
Dataviz review received on September 22, 2020
Final draft submitted on October 21, 2020

EIC

Here are the comments from the dataviz editor for your final revision. The AE has signed it [the revised paper] off.

Dataviz Editor Comments

This is a long and involved paper that deserves careful study. My comments are based on a quick reading and on concentrating on the graphics.

Figure 2 intends to show absolute differences in death counts between sources. It is a curious collection of displays that provides little of the information you would like to know. How many county/day combinations are possible? How many differences are zero? Are there patterns in the differences? Do differences balance out over time? How big does a difference for a county with a big population have to be to matter? Is it always the same counties? Are USAFacts figures more likely to be higher or lower? What have they done about overplotting?

Figure 4 As is commonplace in Covid articles, maps are used without any discussion of the base population sizes. xkcd has an educational example on this (https://xkcd.com/1138/).
Using the same colours for different plots of the same type on the same page with different colour scalings is potentially misleading.
The authors write:
p14 "Overall, Figure 4 clearly shows that the COVID-19 outbreak in the United States is incredibly dynamic both in time and across different regions."
Does it?

Figure 5 takes a bit of work. One thing that is apparent, but not mentioned, is that the four NY counties amongst the six worst-affected have similar heatmaps.
Given that the scaling is the same for all 12 heatmaps, it is unnecessary to repeat it 12 times. There is a lot of repitition in Figure 5.

Figure 6 says it plots all errors from display 5.1, do they mean Table 6?
Figures 6(a) and (b) show a combination predictor that is very close to one of the predictors, but far away from the other sometimes. How can this be? The weighting? Then they should say so in the caption.

Figure 7 is interesting, but would you put faith in a system in which the weights can change so dramatically? Look at the time series for Suffolk, NY in Figure 7(b).

Figure 8 would need a lot more study than there is time for if I am to respond quickly, but I am not convinced by the authors' comments. The same comment on population sizes arises here as with Figure 4.

Figures 11 and 12 Predicting cumulative death counts gets easier as time goes on, shouldn't they be predicting additional death counts?

Response to comments by the dataviz editor for the submission “Curating a COVID-19 data repository and forecasting county-level death counts in the United States"

Nick Altieri, Rebecca L Barter, Raaz Dwivedi, James Duncan,
Karl Kumbier, Xiao Li, Robert Netzorg, Briton Park, Chandan Singh
Yan Shuo Tan, Tiffany Tang, Yu Wang, Chao Zhang, Bin Yu

August 31, 2020

The authors would like to thank the dataviz editor (DE), and the EIC for their comments on the revised manuscript. We now provide our responses and summarize the changes made in the final submission.

Comment #1. *“...Figure 2 intends to show absolute differences in death counts between sources. It is a curious collection of displays that provides little of the information you would like to know. How many county/day combinations are possible? How many differences are zero? Are there patterns in the differences? Do differences balance out over time? How big does a difference for a county with a big population have to be to matter? Is it always the same counties? Are USAFacts figures more likely to be higher or lower? What have they done about overplotting? ..”*
Response: The three plots in Fig. 2 were a later addition to our paper after receiving the first round of reviews. In hindsight, we agree with DE’s comments that this collection of plots in Fig. 2 could be clearer. Consequently, we have replaced this figure with just two histograms (one is the truncated version of the other to denote the scale). These histograms make the direction and magnitude of the differences clearer and removes issues associated with overplotting. In the text, we provide further details that address your other questions related to the differences between USAFacts and NY Times datasets (on page 8 last paragraph, while discussing Figure 2).

Comment #2. *“...Figure 4 As is commonplace in Covid articles, maps are used without any discussion of the base population sizes. xkcd has an educational example on this (https://xkcd.com/1138/). Using the same colours for different plots of the same type on the same page with different colour scalings is potentially misleading. The authors write: p14 "Overall, Figure 4 clearly shows that the COVID-19 outbreak in the United States is incredibly dynamic both in time and across different regions.” Does it?...”*
Response: Thanks a lot for this comment. We agree and have revised the map to include a visualization of per-capita death counts in panel (b) of Figure 4 (using a different color). We also refer the reader to additional visualizations that were already available at `https://covidseverity.com` that were not included in the paper. Moreover, we agree that our conclusion about the map stated in the text was not accurate, and we have thus replaced it with a more detailed discussion at the beginning of Section 3. For instance, one of the lines in the revised description reads “Large population centers were particularly affected in terms of total deaths, and the north- and southeast in terms of deaths per

capita. While Los Angeles County recorded 3,110 deaths up to June 20—the 5th highest in the country—it appears relatively unaffected in panel (b) due to the large population size (over 10 million)."

Comment #3. *".. Figure 5 takes a bit of work. One thing that is apparent, but not mentioned, is that the four NY counties amongst the six worst-affected have similar heatmaps. Given that the scaling is the same for all 12 heatmaps, it is unnecessary to repeat it 12 times. There is a lot of repitition in Figure 5. ..."*

Response: Thank you for this great comment. We included the heatmaps after the second round of reviews to display variability in the rank distribution. However, we agree the figures were not as informative as we intended. As a result, we have reversed our figures to those in the first submission, namely the plots that only show the mean rank across days. However, we still provide the full heatmap in the Appendix to provide a clear picture of variability in the ranks across days. We made this choice in place of showing confidence bands since the ranks are discrete random variables and confidence bands in our case are non-informative. We have also edited the corresponding discussion in the main text.

Comment #4. We now collect several other helpful comments made by the reviewer:

(i) *"Figure 6 says it plots all errors from display 5.1, do they mean Table 6? Figures 6(a) and (b) show a combination predictor that is very close to one of the predictors, but far away from the other sometimes. How can this be? The weighting? Then they should say so in the caption."*

(ii) *"Figure 7 is interesting, but would you put faith in a system in which the weights can change so dramatically? Look at the time series for Suffolk, NY in Figure 7(b)."*

(iii) *"Figure 8 would need a lot more study than there is time for if I am to respond quickly, but I am not convinced by the authors' comments. The same comment on population sizes arises here as with Figure 4."*

(iv) *"Figures 11 and 12 Predicting cumulative death counts gets easier as time goes on, shouldn't they be predicting additional death counts?"*

Response:

(i) The errors are indeed from display 5.1. In Table 6, we had provided a summary statistics of the errors. To remind the reader we added a sentence in that paragraph. And yes CLEP used time-varying weights, which we clarified in the caption (as well as text) by adding this line: "Note that the weights (3.6) used in CLEP are adaptive over time."

(ii) The CLEP weights are unstable in the early stages due to lack of enough data for modeling (and the fact that it was harder to reliably predict deaths earlier in the outbreak due to low death counts). However, in some cases like Suffolk county, the COVID-19 data itself has sudden changes (sharp upticks due to revisions on a single day) which cause the instability in our predictors. We discussed this aspect a few times in various parts of the paper and have also edited the discussion in the context o fthis figure.

(iii) We have removed this figure as we believe, in hindsight, that it was a poor judgment call to add this figure in the paper and it did not provide any insights as we intended.

(iv) We agree that with time the improvement in MAE error for CLEP, linear, and expanded shared predictors is due to the task of predicting deaths getting easier as the cumulative death trend becomes closer to linear growth. We note that we can find new deaths by subtracting our estimate of cumulative deaths from the current number of cumulative deaths. At our website covidseverity.com, we do provide such new death count predictions. Unfortunately, developing a new model to predict new deaths and replicating our analyses are out of the scope of this paper.