Open Forum Infectious Diseases

BRIEF REPORT

Aggregating Human Judgment Probabilistic Predictions of Coronavirus Disease 2019 Transmission, Burden, and Preventive Measures

Allison Codi,[1] Damon Luk,[1] David Braun,[2] Juan Cambeiro,[3,4] Tamay Besiroglu,[3,5] Eva Chen,[6] Luis Enrique Urtubey de Cesaris,[6] Paolo Bocchini,[7] and Thomas McAndrew[1]

[1]College of Health, Lehigh University, Bethlehem, Pennsylvania, USA, [2]Department of Psychology, College of Arts and Sciences, Lehigh University, Bethlehem, Pennsylvania, USA, [3]Department of Epidemiology, Mailman School of Public Health, Columbia University, New York, USA, [4]Metaculus, Santa Cruz, California, USA, [5]Computer Science and AI Lab, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA, [6]Good Judgment Inc, New York, New York, USA, and [7]Department of Civil and Environmental Engineering, P.C. Rossin College of Engineering and Applied Science, Lehigh University, Bethlehem, Pennsylvania, USA

Aggregated human judgment forecasts for coronavirus disease 2019 (COVID-19) targets of public health importance are accurate, often outperforming computational models. Our work shows that aggregated human judgment forecasts for infectious agents are timely, accurate, and adaptable, and can be used as a tool to aid public health decision making during outbreaks.

Keywords. forecasting; human judgment; coronavirus disease 2019.

Accurate forecasts of the trajectory of coronavirus disease 2019 (COVID-19) and preventive measures to reduce transmission of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) provide foresight that enables public health officials to mitigate the impact of the pandemic [1]. Mathematical models are the most commonly used tool to improve situational awareness [2]. While some mathematical models for COVID-19 have access to data on community-level dynamics and human behavior, such as models that depend on mobility and contact patterns, most models rely on structured, reported surveillance data to generate forecasts [3, 4]. Many computational models may not have access to the same types of subjective information that are available to humans. Humans are capable of learning from a combination of objective data and subjective data from their surroundings, environment, and community.

Human judgment has produced accurate forecasts of the progression of an infectious agent for seasonal epidemics and pandemic events [5–7]. Past work studying COVID-19 and human judgment has highlighted the potential ability of aggregate human judgment predictions to adapt to changing dynamics faster than mathematical models [7, 8].

When human judgment forecasts have had lower accuracy than mathematical models, previous work has shown that combining the 2 improves performance over the mathematical model alone [9]. Human judgment predictions of an infectious agent have low overhead, are flexible, and supply rapid and adaptable forecasts to public health decision makers [6, 10].

To best prepare for and prevent infectious disease outbreaks, health officials need quick, accurate, and adaptable forecasts [11]. We show evidence supporting that human judgment aggregated probabilistic predictions meet these criteria for COVID-19 targets associated with transmission, burden, and preventive measures.

METHODS

Monthly surveys from 6 January 2021 to 16 June 2021 collected predictions from 2 human judgment forecasting platforms: Metaculus and Good Judgment Open (GJO) [12, 13]. Details surrounding participant solicitation and prediction format for both platforms can be found in Supplementary Appendix A. Subscribers to both platforms were invited to participate via email solicitation. We included monthly forecasts of the pandemic in summary reports to aid real-time public health decision making, which contain a detailed list of human judgment predictions and the exact wording of each question posed to both crowds [14].

Participants provided probabilistic predictions at the US national level for 6 targets of public health importance: (1) weekly incident cases, (2) hospitalizations, (3) deaths, (4) cumulative first and (5) full-dose vaccinations, and (6) prevalence of immunity-evading variants. Prevalence of immunity-evading variants is defined as the percentage of variants of concern and variants of high consequence that are capable of evading antibodies that block infection.

Participants could submit an initial prediction and revise their prediction as many times as they wished within the approximately 12-day survey period. We define a survey period as the time between when forecasters can submit forecasts and when submissions close. Surveys closed on average 10

Received 23 April 2022; editorial decision 12 July 2022; published online 25 July 2022

Correspondence: Allison Codi, Lehigh University College of Health, Health Science and Technology Bldg, 124 East Morton St, Bethlehem, PA 18015, USA (amc522@lehigh.edu).

Open Forum Infectious Diseases®

https://doi.org/10.1093/ofid/ofac354

days before the start date of the week that we asked participants to forecast. The number of weeks between when the survey closed and the end of the week we asked participants to forecast was most often either 1, 2, or 3 weeks. One question asked for a 6-weeks-ahead prediction (Supplementary Appendix B). Participants were only asked to make a prediction about 1 time point per target per survey. We did not ask participants to make predictions over multiple time points for a single target within 1 survey. Participants received feedback about the accuracy of their forecast via email when the ground truth was available.

Individual forecasts submitted to Metaculus and GJO forecasting platforms were combined into an equally weighted linear pool called a consensus forecast [15, 16]. The consensus predictive q$^{\text{th}}$ quantile value (F_q) was computed as an equally weighted average of all individual q$^{\text{th}}$ quantile values.

$$F_q = \sum_{c=1}^{C} \pi_c F_{q,c}$$

where C is the number of participants who submitted a forecast, $F_{q,c}$ is the q$^{\text{th}}$ quantile value submitted by participant c, and π_c is a nonnegative weight assigned to participant c such that all weights sum to 1. Weights were chosen to be $\pi_c = 1/C$.

Consensus forecasts of incident cases and deaths were compared to the COVID-19 Forecasthub, an ensemble that combined up to 48 computational models between the months of January 2021 and June 2021 [17]. The dates that forecasts were generated by human judgment and by computational models in the COVID-19 Forecasthub were chosen to be on average within 2 days of one another.

For each target, we report the absolute error (AE), defined as a forecast median prediction minus the truth, and the percentage error (PE), defined as the absolute error divided by the truth and multiplied by 100. Forecasts were also scored using weighted interval scores (WISs), the scoring method adopted by the Centers for Disease Control and Prevention to evaluate forecasts of incident cases, deaths, and hospitalizations submitted as a set of central quantiles [17]. WISs can be found in Supplementary Appendix C.

RESULTS

A total of 404 unique participants (71 Metaculus, 333 GJO) submitted probabilistic predictions across the 33 questions for the above 6 targets for a total of 2021 unique forecasts (open access data set available here [18]). A participant was not required to answer all questions. The median consensus prediction for targets 1–5 had a mean PE of 39% in the first survey, 9% for survey 2, 13% for survey 3, and 11%, 26%, and 9% for surveys 4–6. The largest PE was 73% for a prediction of incident cases that was submitted on survey 5 and the smallest PE was 0.1% for a prediction of incident deaths that was submitted on survey 1 (Figure 1).

PE for the majority of targets decreased over time. The PE of the median consensus prediction was 58% (620 192 AE) for incident cases and 60% (49 201 AE) for incident hospitalizations in the first survey. Both targets reduced their PE to 15% (an AE of 13 803 for cases and an AE of 2191 for hospitalizations) in the last survey. PE decreased from 18% to 2% (9 613 628 AE to 3 821 920 AE) for cumulative first-dose vaccinations and from 6.1% to 5.8% (3 745 157 AE to 9 236 130 AE) for cumulative full vaccinations between the initial and final surveys.

Though PE decreases over time, WISs do not show a clear trend across surveys. The WIS decreases when comparing the first and last surveys for cases (138 160 to 730 WIS), deaths (100 to 21 WIS), hospitalizations (16 158 to 469 WIS), first-dose vaccinations (3 263 485 to 692 935 WIS), and variants (11 to 5 WIS), but rises and falls across the surveys in between. The WIS increases when comparing the first and last survey for cumulative vaccinations (1 084 170 to 3 183 063 WIS) (Supplementary Figure 1 and Supplementary Appendix C).

The PE for median consensus predictions of incident deaths was on average 7% (451 mean AE across all 6 surveys) with a PE <0.5% for survey 1 and survey 4 (27 AE and 13 AE).

The PE for variant prevalence was on average 57% (13 average AE) and the highest PE was 153% (14 AE) in survey 6.

The median consensus prediction was closer to the truth than 62% of the 2021 individual predictions. When subset to the 6 incident deaths targets, the consensus prediction was closer to the truth than 75% of individual predictions and in survey 5 the consensus median prediction of incident deaths was closer to the truth than all of the 59 individual predictions.

Compared to ensemble predictions made by the COVID-19 Forecasthub, the median consensus prediction generated by humans was closer to the truth for 3 of 6 predictions of incident cases and 4 of 6 predictions of incident deaths. For predictions of incident cases, the mean PE was 32.8% for the COVID-19 Forecasthub and 33.5% for aggregate human judgment. For incident deaths, the mean PE was 10% for the COVID-19 Forecasthub vs 7% for human judgment.

DISCUSSION

We show that (1) aggregate human judgment forecasts are frequently closer to the truth than individual forecasts, (2) the accuracy of aggregate forecasts depends on the target, (3) the accuracy of aggregate forecasts can improve over time, and (4) aggregate human judgment can produce forecasts of incident cases and deaths with similar accuracy to an ensemble of computational models.

We are limited by the small number of questions we asked, the short time span over which we surveyed the crowd, and the lack of a controlled environment in which to pose questions.

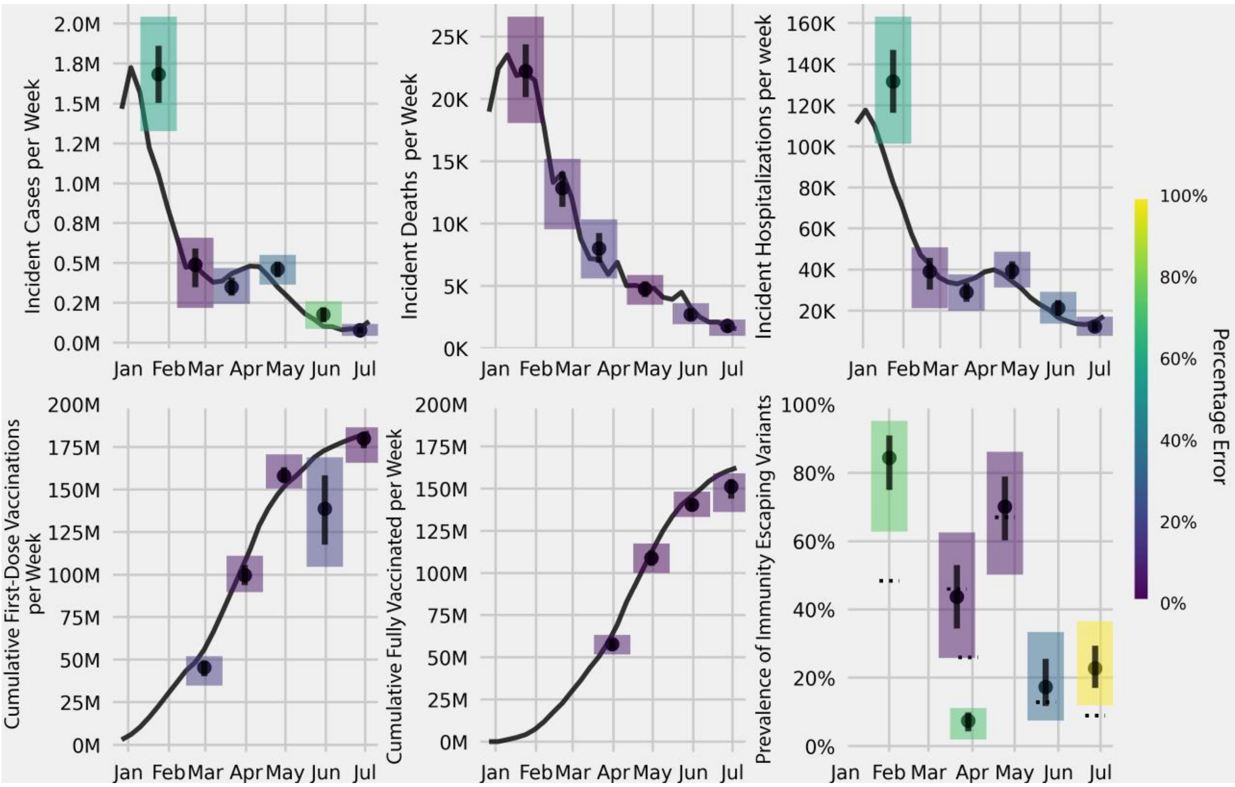

Figure 1. Consensus median (black dot), 25th and 75th percentiles (bottom and top of solid black bar), and the 2.5th and 97.5th percentiles (bottom and top of rectangle) for predictive distributions of aggregate human judgment forecasts of weekly incident cases, hospitalizations, and deaths, cumulative first and full-dose vaccinations, and prevalence of immunity-evading variants at the United States national level. The number of weeks between when consensus predictions were generated to ground truth ranged from 1 to 3 weeks for the majority of predictions (see Supplementary Appendix B for more details). Predictions were submitted between January 2021 and June 2021. Predictions for survey 6 were made for the week starting on 27 June and ending on 3 July. The ground truth is a solid black line or a dashed black line. Rectangles are shaded using the viridis colormap with dark blue rectangles corresponding to low percentage error (PE) and bright yellow rectangles corresponding to high PE. Lighter rectangles correspond to higher PE.

Evidence that the accuracy of the consensus forecast increased over time is limited by the choice of evaluation metric. Our interpretation of our primary evaluation metric, PE, decreasing over time may be influenced by the observed data for cases, hospitalizations, and deaths decreasing over the course of the 6 months of surveys. Our secondary evaluation metric, WIS, does not show as clear of a trend to support PE. Future work should examine the ability of humans to improve the accuracy of their forecasts for a single target over time.

A notable limitation of human judgment forecasting is that humans have a finite amount of cognitive energy, so they can only generate forecasts for a limited number of targets. Computational models do not face this limitation. Future work may address how to map human judgment forecasts for a limited number of targets to additional targets related by location or time horizon. Though out of scope for this work, future work should also explore differences between predictions generated on Metaculus vs GJO platforms.

Contrary to recent work that showed a crowd can produce more accurate forecasts for cases than deaths [8], we found that aggregate median predictions of incident deaths were more accurate than predictions of incident cases.

This may be because humans have the innate capacity to learn relationships between a set of evolving signals, such as incident cases, hospitalizations, and vaccinations, that are correlated with the target they aim to predict. The lack of signals and environmental cues related to questions about the prevalence of specific variants may be why these aggregate forecasts were inaccurate. The availability of environmental cues related to cases, deaths, and hospitalizations may explain why participants were able to learn over time; however, more experimental work related to how humans incorporate data to make predictions should be explored.

Supplementary Data

Supplementary materials are available at *Open Forum Infectious Diseases* online. Consisting of data provided by the authors to benefit the reader, the posted materials are not copyedited and are the sole responsibility of the authors, so questions or comments should be addressed to the corresponding author.

Notes

Acknowledgments. We thank Phillip Rescober for data science support from Good Judgment Inc, and all of the individual forecasters who contributed their time and energy to generate predictions about the trajectory of COVID-19.

Patient consent. This work was approved by the Lehigh University Institutional Review Board.

Financial support. This research was supported by the National Institute of General Medical Sciences (3U24GM132013-02S2) through supplemental funding to the MIDAS Coordination Center (MIDASNI2020-1).

Potential conflicts of interest. T. B. has been employed by Metaculus and owns stock in this company. All other authors report no potential conflicts.

All authors have submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest. Conflicts that the editors consider relevant to the content of the manuscript have been disclosed.

References

1. Pollett S, Johansson MA, Reich NG, et al. Recommended reporting items for epidemic forecasting and prediction research: the EPIFORGE 2020 guidelines. PLoS Med **2021**; 18:e1003793.
2. Biggerstaff M, Slayton RB, Johansson MA, Butler JC. Improving pandemic response: employing mathematical modeling to confront coronavirus disease 2019. Clin Infect Dis **2022**; 74:913–7.
3. Balcan D, Vespignani A. Phase transitions in contagion processes mediated by recurrent mobility patterns. Nat Phys **2011**; 7:581–6.
4. Zhang J, Litvinova M, Liang Y, et al. Changes in contact patterns shape the dynamics of the COVID-19 outbreak in China. Science **2020**; 368:1481–86.
5. Farrow DC, Brooks LC, Hyun S, Tibshirani RJ, Burke DS, Rosenfeld R. A human judgment approach to epidemiological forecasting. PLoS Comput Biol **2017**; 13: e1005248.
6. McAndrew TC, Reich NG. An expert judgment model to predict early stages of the COVID-19 outbreak in the United States. medRxiv[Preprint]. Posted online 23 September **2020**. https://doi.org/10.1101/2020.09.21.20196725
7. McAndrew T, Majumder MS, Lover AA, et al. Early human judgment forecasts of human monkeypox, May 2022. The Lancet Digital Health **2022**; 4(8):e569–e571.
8. Bosse NI, Abbott S, Bracher J, et al. Comparing human and model-based forecasts of COVID-19 in Germany and Poland. medRxiv [Preprint]. Posted online 5 September **2021**. https://doi.org/10.1101/2021.12.01.21266598
9. Ibrahim R, Kim S-H, Tong J. Eliciting human judgment for prediction algorithms. Manag Sci **2021**; 67:2314–25.
10. McAndrew T, Cambeiro J, Besiroglu T. Aggregating human judgment probabilistic predictions of the safety, efficacy, and timing of a COVID-19 vaccine. Vaccine **2022**; 40:2331–41.
11. Lutz CS, Huynh MP, Schroeder M, et al. Applying infectious disease forecasting to public health: a path forward using influenza forecasting examples. BMC Public Health **2019**; 19:1659.
12. Metaculus. About. https://www.metaculus.com/questions/. Accessed 22 June 2021.
13. Cultivate Labs. Good Judgment Open. https://www.gjopen.com/. Accessed 29 June 2022.
14. GitHub. aggStatModelsAndHumanJudgment_PUBL/summaryreports at main· computationalUncertaintyLab/aggStatModelsAndHumanJudgment_PUBL. https://github.com/computationalUncertaintyLab/aggStatModelsAndHumanJudgment_PUBL. Accessed 29 June 2022.
15. Winkler RL. The consensus of subjective probability distributions. Manag Sci **1968**; 15:B61–75.
16. Clemen RT, Winkler RL. Combining probability distributions from experts in risk analysis. Risk Anal **1999**; 19:187–203.
17. Cramer EY, Huang Y, Wang Y, et al. The United States COVID-19 Forecast Hub dataset. medRxiv [Preprint]. Posted online 4 November **2021**. https://doi.org/10.1101/2021.11.04.21265886
18. Zoltar. Project: aggregating statistical models and human judgment. https://zoltardata.com/project/239. Accessed 29 June 2022.