

Aggregating human judgment probabilistic predictions of COVID-19 transmission, burden, and preventative measures

Allison Codi,¹ Damon Luk,¹ David Braun,¹ Juan Cambeiro,^{2,3} Tamay Besiroglu,^{2,4} Eva Chen,⁵ Luis Enrique Urtubey de Cèsaris,⁵ Paolo Bocchini,⁶ and Thomas McAndrew^{1,*}

¹*College of Health, Lehigh University, Bethlehem, Pennsylvania, United States of America*

²*Metaculus, Santa Cruz, California, United States of America*

³*Department of Epidemiology, Mailman School of Public Health, Columbia University, New York, United States of America*

⁴*Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America*

⁵*Good Judgment Inc., New York, New York, United States of America*

⁶*Department of Civil and Environmental Engineering, P.C. Rossin College of Engineering and Applied Science, Lehigh University, Bethlehem, Pennsylvania, United States of America*

(Dated: April 15, 2022)

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

* mcandrew@lehigh.edu

I. INTRODUCTION

Accurate forecasts of the trajectory of COVID-19 and preventative measures to reduce transmission of SARS-CoV-2 provide foresight that enables public health officials to mitigate the impact of the pandemic [1]. Mathematical models are the most often used tool to improve situational awareness [2]. However, most mathematical models rely on structured, reported surveillance data and often do not have access to community level transmission dynamics, data related to human behavior, or behavioral responses to policy changes.

Human judgment has produced accurate forecasts of the evolution of an infectious agent for seasonal epidemics and pandemic events [3, 4]. Past work studying COVID-19 and human judgment has highlighted the ability of aggregate human judgment predictions to adapt to changing dynamics faster than mathematical models [5]. When human judgment forecasts have had lower accuracy than mathematical models, previous work has shown that combining the two improves performance over the mathematical model alone [6]. Human judgment predictions of an infectious agent are low-overhead, flexible, and supply rapid and adaptable forecasts to public health decision makers [4].

To best prepare for and prevent infectious disease outbreaks, health officials need quick, accurate, and adaptable forecasts [7]. We show evidence that supports human judgment aggregated probabilistic predictions meet these criteria for COVID-19 targets associated with transmission, burden, and preventative measures.

II. METHODS

Monthly surveys from Jan. 6, 2021 to Jun. 16, 2021 collected predictions from two human judgment forecasting platforms: Metaculus and Good Judgment Open (GJO) [8, 9]. 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. [10].

Participants had approximately twelve days to provide probabilistic predictions one to three weeks ahead of time at the US national level for six 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. Participants could submit an initial prediction and revise their prediction as many times as they wished within the twelve-day period. 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.

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 Jan. 2021 and Jun. 2021 [11]. The date 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 percent error (PE) defined as the absolute error divided by the truth and multiplied by 100.

III. RESULTS

A total of 404 unique participants (71 Metaculus, 333 GJO) submitted probabilistic predictions across the 33 questions for the above six targets for a total of 2,021 unique forecasts (open access data set available here [12]). 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%, 9% for surveys 4 through 6. The largest PE was 73% for a prediction of incident cases that was submitted on survey 5 and smallest PE was 0.1% for a prediction of incident deaths that was submitted on survey 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 2,191 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.

The PE for median consensus predictions of incident deaths was on average 7% (451 mean AE across all six surveys) with a PE less than 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 2,021 individual predictions. When subset to the six incident deaths targets, the consensus prediction was closer to the truth than 75% of individual predictions and in survey five the consensus median prediction of incident deaths was closer to the truth than all of the fifty-nine 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 3/6 predictions of incident cases and 4/6 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.

IV. DISCUSSION

We show that (i) aggregate human judgment forecasts are frequently closer to the truth than individual forecasts, (ii) the accuracy of aggregate forecasts depends on the target, (iii) the accuracy of aggregate forecasts can improve over time, and (iv) 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.

Contrary to recent work that showed a crowd can produce more accurate forecasts for cases than deaths [5], we found 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.

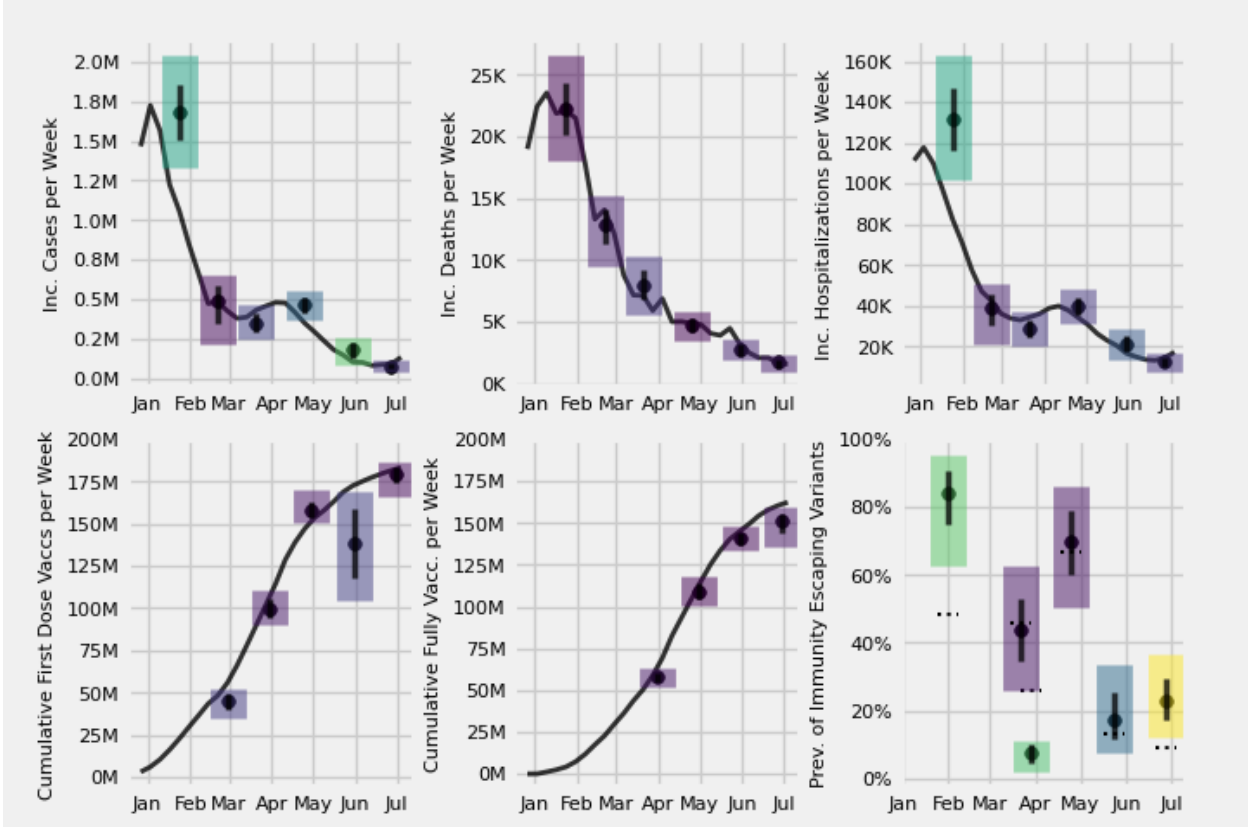


FIG. 1: Consensus median (black dot), 25th and 75th percentiles (bottom and top of solid black bar), and the 2.5th and 97.5th (bottom and top of rectangle) for 1 through 3 week ahead predictive distributions of aggregate human judgment forecasts of weekly incident cases, hospitalizations, and deaths, cumulative first and full-dose vaccinations, and prevalence of immunity evading variants at the US national level.

Predictions were submitted between Jan. 2021 and Jun. 2021. Predictions for survey 6 were made for the week starting on Jun. 27 and ending on Jul. 3. The ground truth is a solid black line or a dashed black line. Lighter rectangles correspond to higher percent error.

V. ACKNOWLEDGEMENTS

This research was supported through the MIDAS Coordination Center (MIDASNI2020-1) by a grant from the National Institute of General Medical Science (3U24GM132013-02S2). We wish to thank Phillip Rescober for data science support from Good Judgment Inc. We wish to thank all of the individual forecasters who

contributed their time and energy to generate predictions about the trajectory of COVID-19.

-
- [1] Simon Pollett, Michael A Johansson, Nicholas G Reich, David Brett-Major, Sara Y Del Valle, Srinivasan Venkatramanan, Rachel Lowe, Travis Porco, Irina Maljkovic Berry, Alina Deshpande, et al. Recommended reporting items for epidemic forecasting and prediction research: The epiforge 2020 guidelines. *PLoS medicine*, 18(10):e1003793, 2021.
 - [2] Matthew Biggerstaff, Rachel B Slayton, Michael A Johansson, and Jay C Butler. Improving pandemic response: Employing mathematical modeling to confront coronavirus disease 2019. *Clinical Infectious Diseases*, 2021.
 - [3] David C Farrow, Logan C Brooks, Sangwon Hyun, Ryan J Tibshirani, Donald S Burke, and Roni Rosenfeld. A human judgment approach to epidemiological forecasting. *PLoS computational biology*, 13(3):e1005248, 2017.
 - [4] Thomas McAndrew and Nicholas G Reich. An expert judgment model to predict early stages of the covid-19 outbreak in the united states. *Medrxiv*, 2020.
 - [5] Nikos I Bosse, Sam Abbott, Johannes Bracher, Habakuk Hain, Billy J Quilty, Mark Jit, Edwin van Leeuwen, Anne Cori, Sebastian Funk, et al. Comparing human and model-based forecasts of covid-19 in germany and poland. *medRxiv*, 2021.
 - [6] Rouba Ibrahim, Kim Song-Hee, and Jordan Tong. Eliciting human judgment for prediction algorithms. *Management Science*, 67(4):2314–2325, 2021.
 - [7] Chelsea S Lutz, Mimi P Huynh, Monica Schroeder, Sophia Anyatonwu, F Scott Dahlgren, Gregory Danyluk, Danielle Fernandez, Sharon K Greene, Nodar Kipshidze, Leann Liu, et al. Applying infectious disease forecasting to public health: a path forward using influenza forecasting examples. *BMC Public Health*, 19(1):1–12, 2019.
 - [8] Metaculus Home Page. <https://www.metaculus.com/questions/>, 2021. [Online; accessed 29-03-2021].
 - [9] Good Judgement Open Home Page. <https://www.gjopen.com/>, 2021. [Online; accessed 29-03-2021].
 - [10] Thomas McAndrew. Summary reports. <https://github.com/computationalUncertaintyLab/aggStatModelsAndHumanJudgment.PUBL/tree/main/summaryreports>, 2021.
 - [11] Estee Y Cramer, Yuxin Huang, Yijin Wang, Evan L Ray, Matthew Cornell, Johannes Bracher, Andrea Brennen, Alvaro J Castro Rivadeneira, Aaron Gerding, Katie House, Dasuni Jayawardena, Abdul H Kanji, Ayush Khandelwal, Khoa Le, Jarad Niemi, Ariane Stark, Apurv Shah, Nutch Wattanachit, Martha W Zorn, Nicholas G Reich, and US COVID-19 Forecast Hub Consortium. The united states covid-19 forecast hub dataset. *medRxiv*, 2021.
 - [12] Thomas McAndrew and Allison Codi. Archive of human judgment forecasts. <https://zoltardata.com/project/239>, 2021.