

Original Paper

# Crowdsourced Perceptions of Human Behavior to Improve Computational Forecasts of US National Incident Cases of COVID-19: Survey Study

David Braun<sup>1</sup>, PhD; Daniel Ingram<sup>2</sup>, MBA; David Ingram<sup>2</sup>, CERA, FRM, PRM; Bilal Khan<sup>3</sup>, PhD; Jessecae Marsh<sup>1</sup>, PhD; Thomas McAndrew<sup>4</sup>, PhD

<sup>1</sup>Department of Psychology, Lehigh University, Bethlehem, PA, United States

<sup>2</sup>Actuarial Risk Management, Austin, TX, United States

<sup>3</sup>Computer Science and Engineering, Lehigh University, Bethlehem, PA, United States

<sup>4</sup>College of Health, Lehigh University, Bethlehem, PA, United States

**Corresponding Author:**

David Braun, PhD  
Department of Psychology  
Lehigh University  
17 Memorial Dr E  
Bethlehem, PA, 18015  
United States  
Phone: 1 6107583000  
Email: [dab414@lehigh.edu](mailto:dab414@lehigh.edu)

## Abstract

**Background:** Past research has shown that various signals associated with human behavior (eg, social media engagement) can benefit computational forecasts of COVID-19. One behavior that has been shown to reduce the spread of infectious agents is compliance with nonpharmaceutical interventions (NPIs). However, the extent to which the public adheres to NPIs is difficult to measure and consequently difficult to incorporate into computational forecasts of infectious diseases. Soliciting judgments from many individuals (ie, crowdsourcing) can lead to surprisingly accurate estimates of both current and future targets of interest. Therefore, asking a crowd to estimate community-level compliance with NPIs may prove to be an accurate and predictive signal of an infectious disease such as COVID-19.

**Objective:** We aimed to show that crowdsourced perceptions of compliance with NPIs can be a fast and reliable signal that can predict the spread of an infectious agent. We showed this by measuring the correlation between crowdsourced perceptions of NPIs and US incident cases of COVID-19 1-4 weeks ahead, and evaluating whether incorporating crowdsourced perceptions improves the predictive performance of a computational forecast of incident cases.

**Methods:** For 36 weeks from September 2020 to April 2021, we asked 2 crowds 21 questions about their perceptions of community adherence to NPIs and public health guidelines, and collected 10,120 responses. Self-reported state residency was compared to estimates from the US census to determine the representativeness of the crowds. Crowdsourced NPI signals were mapped to 21 mean perceived adherence (MEPA) signals and analyzed descriptively to investigate features, such as how MEPA signals changed over time and whether MEPA time series could be clustered into groups based on response patterns. We investigated whether MEPA signals were associated with incident cases of COVID-19 1-4 weeks ahead by (1) estimating correlations between MEPA and incident cases, and (2) including MEPA into computational forecasts.

**Results:** The crowds were mostly geographically representative of the US population with slight overrepresentation in the Northeast. MEPA signals tended to converge toward moderate levels of compliance throughout the survey period, and an unsupervised analysis revealed signals clustered into 4 groups roughly based on the type of question being asked. Several MEPA signals linearly correlated with incident cases of COVID-19 1-4 weeks ahead at the US national level. Including questions related to social distancing, testing, and limiting large gatherings increased out-of-sample predictive performance for probabilistic forecasts of incident cases of COVID-19 1-3 weeks ahead when compared to a model that was trained on only past incident cases.

**Conclusions:** Crowdsourced perceptions of nonpharmaceutical adherence may be an important signal to improve forecasts of the trajectory of an infectious agent and increase public health situational awareness.

(*JMIR Public Health Surveill* 2022;8(12):e39336) doi: [10.2196/39336](https://doi.org/10.2196/39336)

## KEYWORDS

crowdsourcing; COVID-19; forecasting; human judgment

## Introduction

Forecasting the transmission of infectious agents can support decisions made by public health officials and key decision makers [1,2]. Past forecasts of seasonal influenza, Ebola, dengue, chikungunya, and Zika have helped officials take short-term action to stymie the spread and burden of disease and draft policy decisions [3-8]. The COVID-19 pandemic has further highlighted the importance that forecasts play in support of public health situational awareness [9-11].

The majority of forecasts of an infectious disease are generated by computational models; however, past work has shown that human judgment is also capable of making accurate predictions of a diverse number of phenomena [12,13], including infectious agents [14-17].

Work in human judgment predictions can be categorized into direct and indirect predictions. Direct predictions are collected by asking humans to estimate the probability of a future event of interest. Researchers have used various methods to solicit direct predictions from a lay, expert, or mixed crowd by varying the format humans use to submit predictions and training different algorithms to combine individual forecasts [18-21]. Structured elicitation formalizes how a prediction should be collected to minimize potential biases or undue influences, and a researcher could use several different protocols to rigorously collect predictions [19,20,22].

Past work has found middling performance when asking those with subject matter expertise to make direct predictions [23,24]. As with experts, the performance of predictions made by lay people has been mixed, and the variability in predictive performance is likely due to cues in the environment that are related to the event of interest [25], as well as people's reliance on heuristics to make fast decisions with little information [26-29]. Humans are subject to several cognitive biases that negatively impact our ability to make sound judgments [30,31]. That said, there are many examples where predictions based on mental heuristics outperformed computational models [32].

Work on aggregating direct human judgment predictions has focused on adjusting for correlated predictions between individuals, assessing the number of individual predictions to combine, and determining how to appropriately weight individuals based on past predictive performance [18,21]. Direct predictions take advantage of a human's ability to build a prediction from available structured data and information typically unavailable to a computational model, such as subjective information, intuition, and expertise [33].

Indirect predictions of a future event are collected by (1) extracting human judgment data from a passive source such as

social media [34-37], (2) actively asking a crowd about covariates that may be related to the target of interest, or (3) asking a crowd to take actions in a prediction market, which can be mapped to probabilistic predictions [38,39]. Indirect predictions offer an opportunity to train a statistical model on both measured objective data and subjective data.

Past work that incorporated social media data in a model often mapped behaviors to a set of random variables and included these random variables in a statistical model [34-37,40]. Most studies have framed these human and social media sources as passive signals that can be mined to contribute to more accurate forecasts. For example, a recent study leveraged mobility data gathered from Twitter to improve forecasts of incident COVID-19 cases at multiple geographic levels [41]. Digital interaction and engagement data beyond social media may be useful predictive signals as well, as a recent study found that Google search trends related to COVID-19 symptoms improved both nowcasting and forecasting of COVID-19 incident cases and deaths [42]. Compartmental models have also been proposed that take into account human behavior by estimating the contact network between individuals, and the reproductive and recovery rates, or by building a more complicated function between disease states that takes into account human behavior [43]. Prediction markets are another approach for aggregating human judgment, which ask a pool of participants to place bets on the potential of future events with an incentive for each participant to optimize their total earnings [38,39]. The goal of creating a prediction market is not to link behavior to outcomes of interest but to take advantage of an individual's ability to extract alternative data sources that are not accessible to computational models and respond to the aggregate behaviors of a market. Models that include indirect predictions report improved performance compared to models that do not include indirect predictions; however, performance varies by the infectious agent and type of data collected. Human behavior and perceptions can also be used to predict social media engagement and community behavior that might benefit decision-making of policy makers and community leaders. For example, past work has looked at which types of messages from organizations shared on Twitter foster the strongest public engagement [44], as well as which sources for health-related information are likely to be sought out based on demographics and how these factors contribute to adherence to social distancing guidelines [45].

In this work, we study how crowdsourced questions related to nonpharmaceutical interventions (NPIs) in one's community can contribute to an improved forecast of COVID-19 incident cases at the national level. We posed 21 questions related to NPIs to a representative sample from the United States over a period of 36 weeks. These crowdsourced data were used to estimate the association between perceptions of adherence to

NPIs and incident cases at the US national level 1-4 weeks in advance. In addition, we fit a predictive model and showed that adding crowdsourced data on perceptions of adherence improves forecast accuracy for incident cases when compared to a control that does not include perceptive data.

To the current literature, we contribute a novel data stream of community-scale perceptive information [46] that shows (1) strong associations with incident cases 1-4 weeks ahead at the national level and (2) improved predictive accuracy of out-of-sample predictions 1-3 weeks ahead when included in a computational model.

## Methods

### Ethical Considerations

We obtained retroactive clearance from Lehigh University's institutional review board (IRB) to publish the data (#1808500-1). The IRB determined obtaining informed consent was not necessary because the data were recorded in such a manner that the identity of human subjects cannot be readily ascertained directly or through identifiers linked to the subjects. Data that have been made publicly available are similarly deidentified [46]. Participants completed surveys either (1) on a volunteer basis or (2) in exchange for compensation. Compensated participants earned credits from the survey platform that could be redeemed for gift cards or donated to charity.

### Survey Logistics

#### Participants and Recruitment

There were 10,852 responses to the survey over the course of 36 weeks starting August 30, 2020, and ending April 28, 2021 (281 responses per week on average with an SD of 119). Paid participants were initially recruited through the *SurveyMonkey* platform (4405/10,852, 40.5%) from September 23, 2020, through February 15, 2021. *SurveyMonkey* is a survey platform with access to more than 140 million participants globally. The platform requires a fee per service and comes with assurance that paid participants will be a representative sample from the locale of interest. A survey can be sent to a set of participants who meet specific criteria (called a targeted audience), such as country of origin, age, socioeconomic factors (income, marital status, and employment), etc. Participants in this study were required to reside in the United States and be at least 18 years old. Survey design, distribution, and data collection were managed via *SurveyMonkey* software.

From February 16, 2021, to April 27, 2021, participants were recruited from the *Pollfish* survey platform (3295/10,852, 30.4%). This change was made due to *SurveyMonkey* delivering a highly variable number of responses per week and, in some weeks, failing to deliver the number of responses ordered. *Pollfish* is another fee per response survey platform that allows the researcher to specify a targeted audience and guarantees a representative number of responses. The goals and services of *SurveyMonkey* and *Pollfish* are similar, though *Pollfish* software collects higher resolution spatial data about respondents. The

*Pollfish* platform collected responses from participants who met the same criteria as those for *SurveyMonkey*.

Compensated respondents from *SurveyMonkey* and *Pollfish* accounted for approximately 70% (7700/10,852, 71.0%) of the responses, and the final approximately 30% (3152/10,852, 29.0%) of participants were recruited as volunteers and participated through the *SurveyMonkey* platform from August 30, 2020, to April 28, 2021. These volunteers were mostly recruited via word of mouth and social media.

We removed participant responses from the analysis if (1) more than half of the questions (ie, 11 of the 21 questions) were left blank or had a response of "Don't know" (4.7% [511/10,852] of responses) or (2) a participant gave the same response to every question (2.3% [331/10,852] of responses). All blank and "Don't know" responses were excluded from the analysis (9.7% [20,569/214,200] of total question responses [ie,  $N_{Participants} \times 21$ ]).

### Survey Timeline

A total of 36 weekly surveys were sent to participants beginning on September 6, 2020, and ending on April 30, 2021. Surveys were distributed to unique participants each Monday, Wednesday, and Friday, and surveys were closed on Sundays. Surveys were not sent to the same participant more than once in a week.

*SurveyMonkey* surveys were open to participants for compensation from the 4th week of the survey period (September 2020) to the 21st consecutive week of the survey (February 2021), and *SurveyMonkey* surveys were open to volunteers over the entire 36-week survey period. *Pollfish* surveys were open to participants from the 21st week of the survey period (February 2021) until the 36th consecutive week of data collection (the end of the survey period; April 2021).

In July and August 2020, surveys were sent to participants to (1) fill out the survey and (2) solicit feedback about whether the questions asked in the survey were worded clearly. Feedback from these first 2 pilot surveys was used to update and finalize surveys sent between September 2020 and April 2021.

### Survey Content and Questions

Surveys between September 2020 and April 2021 asked participants to answer the same set of 21 "core" questions (see [Textbox 1](#) for a list of core questions). Core questions asked participants about their perceptions of their community members' adherence to NPIs, such as mask wearing, and their adherence to public health guidelines related to testing, quarantine, and large gatherings. Participants gave responses to survey questions on a Likert scale with the following options: "None/not adopted," "Few/20%," "Some/40%," "Many/80%," "All/100%," and "Don't know."

In addition to the 21 core questions, several weeks included topical questions asking participants about their perceptions of behavior during specific events (eg, the size of holiday gatherings). Because these questions were not consistent throughout the duration of the study, we chose not to include them in the analyses. At the end of the survey, participants were also asked for optional thoughts and feedback about how

COVID-19 is being addressed in their community and how the survey may be improved in the future (for summary reports of the data composed in real time, see a previous report [47]).

The order in which questions were presented was randomized across all 21 questions in the Pollfish surveys, and

SurveyMonkey questions were randomized within 5 categories that asked participants about individual NPI behaviors, adherence to guidelines associated with community businesses, testing and quarantine, awareness, and restrictions or policies related to educational institutions (see [Multimedia Appendix 1](#)).

**Textbox 1.** List of the 21 “core” questions that were presented to participants in every survey from September 6, 2020, to April 30, 2021.

Questions
What percent of people in your community do you notice are usually:
1. Wearing a mask in public
2. Maintaining social distance
3. Staying at home
How common is it in your community for:
4. Restaurants to have reduced seating
5. Businesses to be closed – work from home only
6. Hairdressers and barbers to be open with restrictions
7. Visitors to senior living facilities to be restricted
8. Commonly touched surfaces to be sanitized
9. Hospitals to have special protection in areas that treat COVID patients
In your community, how common is it for people to follow recommendations or requirements to:
10. Get tested for active virus
11. Get antibody testing to detect prior infection
12. Quarantine people who have been in close contact with people with positive tests
13. Quarantine people with positive tests
14. Quarantine travelers from higher infection places
15. Limit large gatherings of people
How many people in your community are aware of:
16. Local level of COVID infections
17. Statewide targets for reducing COVID spread
18. Local approach to limiting COVID spread
In your state, what percent of:
19. Colleges are closed or holding only remote classes
20. Schools (K-12) are closed or holding only remote classes
21. Violations of COVID restrictions result in fines or police enforcement

### **Data Acquisition and Availability**

Survey data were acquired retrospectively from a team of actuaries (Daniel Ingram and David Ingram) who were interested in the study of human behavior, crowdsourcing, and how perceptions may be predictive of the spread of SARS-CoV-2. There were several limitations to survey collection: (1) participant identifiers were not collected longitudinally and so we cannot track individuals who contributed to the survey, and (2) the wording of survey instructions was slightly different across the SurveyMonkey and Pollfish platforms, which could bias responses.

Individual respondent data of all 21 questions for all 36 weeks are available in a previous report [46]. The data are in wide

format where each row represents a single survey response, and columns are present for the date the survey was completed and the 21 answers to survey questions.

We obtained approval from Lehigh University’s IRB to publish these data on an open-source platform.

### **Epidemiological Data**

Incident cases per epidemiological week (epidemic week) at the national level were collected from the Johns Hopkins University CSSE GitHub repository [48]. This repository stores cumulative cases per day from January 22, 2020, to the present for all 50 states and a set of 5 territories. To compute incident cases for day  $D$ , we subtracted cumulative cases at day  $D$  from cumulative cases at day  $D+1$ . We computed incident cases for

day  $D$  at the national level by summing incident cases for all 50 states and all 5 territories. Daily incident cases at the national level were summed to arrive at incident cases per epidemic week, where an epidemic week began on Sunday and ended on Saturday.

### Assessing Whether the Crowd was Representative of the US Population

We assessed graphically whether our sample was representative of the US population by plotting for all states ( $s$ ) the pair  $(r_s, e_s)$ , where  $r_s$  is the total number of observed participants for state  $s$  and  $e_s$  is the estimated expected number of responses from state  $s$ .

Our estimate  $e_s$  assumes that  $r_s$  was drawn from a random variable  $R_s \sim \text{Bin}(N, \theta_s)$ , where  $N$  is the total number of participants across all surveys and  $\theta_s$  is the probability of choosing at random a citizen registered in state  $s$ . We estimated  $\theta_s$ ,  $\widehat{\theta}_s$ , as the census estimate for state  $s$  divided by the sum of census estimates for all states. The value  $e_s$  is  $\widehat{E}(R_s) = N\widehat{\theta}_s$ .

We included an estimated correlation coefficient between the observed and expected number of participants sampled across all states. For each state, we also compared the relative difference between the observed and expected proportions of participants ([Multimedia Appendix 2](#)).

### Statistical Setup

We suppose a survey response to question  $q$  from participant  $i$ , at time  $t$ ,  $x_{t,i,q}$ , was generated from a random variable  $X_{t,i,q}$  which has support  $\text{supp}(X_{t,i,q}) = \{0, 1, 2, 3, 4\}$  corresponding to 5 different levels of adherence. The value 0 corresponds to no adherence or adherence not adopted in the community, and the value 4 corresponds to complete adherence (the response “All/100%” on the survey). Random variables at time  $t$  for question  $q$  between 2 participants are considered independent.

Mean perceived adherence (MEPA) is defined for a specific question  $q$  and at a specific time  $t$  as the average of  $x_{t,i,q}$  over participants, or

$$\text{MEPA}_{q,t} = N^{-1} \sum_{i=1}^P x_{t,i,q}$$

where  $N$  is the number of responses for question  $q$  at time  $t$ .  $\text{MEPA}_{q,t}$  is intended to measure an aggregated adherence to a specific type of NPI. Though individual responses are discrete,  $\text{MEPA}_{q,t}$  is a continuous value. If we define the random variable  $\text{MEPA}_{q,t}$  as the average of  $N$  independent random variables with finite variance, then we expect  $\text{MEPA}_{q,t}$  to have a bell-curved distribution that resembles the normal distribution restricted to the closed interval from 0 to 4.

Incident US national COVID-19 cases at epidemiological week  $t$ , ( $c_t$ ), are assumed to be generated from a corresponding random variable  $C_t$ , and we make no additional assumptions about this time series.

### Estimating the Correlation Between MEPA and Incident Cases

For each survey question, we estimated the correlation coefficient between MEPA at epidemiological week  $t$  and US national incident cases at epidemiological week  $t, t+1, t+2, t+3$ , and  $t+4$ . Line lists of the estimated correlation coefficient at each week-ahead time point and 95% CIs are available in [Multimedia Appendix 3](#).

### Clustering Questions

We fit a hierarchical clustering algorithm to all 21 MEPA time series for 2 through 10 clusters. Dissimilarity between 2 time series was computed using the Euclidean distance. The Silhouette coefficient was used to assess the quality of fitting 2 clusters, 3 clusters, and so on (up to 10 clusters) [49]. A dendrogram was plotted to visualize the clustering, and MEPA time series were grouped and plotted over the epidemiological week.

### Forecast Models With and Without Crowdsourced Perceptions

#### SIR Plus Vector Autoregression Moving Average

An SIR (susceptible, infected, and removed) model was fit to the number of US incident cases to produce an estimated number of incident cases  $I_t$ , and residuals ( $e_t = c_t - I_t$ ) were modeled with a vector autoregression moving average (VARMA) model that included one or more MEPA time series.

The SIR model estimates at time  $t$  the number of individuals existing in the susceptible ( $S_t$ ), infected ( $I_t$ ), and removed ( $R_t$ ) compartment according to

$$\begin{aligned} dS_t/dt &= -\beta \frac{S_t I_t}{N} \\ dI_t/dt &= \beta \frac{S_t I_t}{N} - \gamma I_t \\ dR_t/dt &= \gamma I_t \\ N &= S_t + I_t + R_t \end{aligned}$$

with initial values  $S_0$ ,  $I_0$ , and  $R_0$ , and parameters  $\beta > 0$  and  $\gamma > 0$ . We chose  $S_0$  equal to the number of individuals in the United States, according to the most recent census. The initial value  $I_0$  was set equal to the reported number of infections for the first epidemiological week in which survey data were collected (August 30, 2020, to September 05, 2020), and  $R_0$  was set to 0. The initial value problem above was integrated by the Runge-Kutta-Fehlberg method, and parameters  $\beta$  and  $\gamma$  were estimated by minimizing the least squares solution between  $I_t$  and the reported number of incident cases (estimates of the SIR model at 4 different time points can be found in [Multimedia Appendix 4](#)).

Residuals were generated as  $e_t = c_t - I_t$ , and we assumed that these residuals together with one of the MEPA time series can be modeled as a VARMA model. VARMA assumes the residuals, and the MEPA time series  $M_q$  follows

$$\theta(L)Y_t = \psi(L)U_t$$

where  $Y_t = [y_t, m_{q,t}]'$ ,  $U_t$  is a random vector following a white noise process or  $U_t \sim N(0, \Sigma)$ , the operator  $\theta(L) = B_1L + B_2L^2 + \dots$  and  $B_k$  is a matrix of coefficients, the operator  $\psi(L) = A_1L + A_2L^2 + \dots$  and  $A_k$  is a matrix of coefficients, and the operator  $L^j$  is the lag operator or  $L^j Y_t = Y_{t-j}$ . We assumed the covariance between any  $Y_s$  and  $Y_t$  is fixed and equal to  $\Sigma$ .

The optimal number of lags for  $\theta$  and for  $\psi$  was estimated every week through each of the 36 weeks by computing the Akaike information criterion (AIC) for models fit with all combinations of 1 through 3 lags for  $\theta$  and 1 through 3 lags for  $\psi$ . The combination that resulted in the lowest AIC was picked.

### SIR Plus Random Forest Plus VARMA

To incorporate all MEPA time series into a model, we first fit an SIR model to the original time series and computed the residuals  $e_t = c_t - I_t$ . Next, we trained a random forest regression  $f$  with 5000 trees, where the desired output is  $y_t$  as a function of  $e_{t-1}$ , and all the MEPA time series values, smoothed using LOWESS, with a lag of 1. The residuals  $\delta_t = e_t - f(e_{t-1}, \hat{M}_{1,t-1}, \hat{M}_{2,t-1}, \dots, \hat{M}_{q,t-1})$ , where  $\hat{M}_{q,t}$  is the LOWESS smoothed MEPA time series value for question  $q$  at time  $t$ , were computed and were assumed to follow an autoregressive integrated moving average (ARIMA) process, or  $\theta(L)\delta_t = \psi(L)u_t$ . Lags were chosen at each week based on the AIC in the same manner as with the above SIR plus VARMA model.

### Control Model

Our control model followed the same SIR “detrending” of the original incident case time series and then fit an ARIMA to the residuals. The ARIMA followed a similar approach as the VARMA model when modeling

$$Y_t \sim e_t$$

$$\Theta(L)Y_t = \psi(L)u_t$$

where  $u_t \sim N(0, \sigma^2)$ . The only addition to this model is that we may “difference”  $Y_t$  by successively subtracting the values of  $Y$  at time  $t-1$  from the values of  $Y$  at time  $t$  for all times. The difference computes  $d_t = \nabla Y_t = Y_t - Y_{t-1}$ , fits the model above, generates forecasts of  $d_{t+1}, d_{t+2}, \dots$ , and then recovers  $Y_{t+l}$  by computing  $Y_{t+l} - 1 + d(t+l)$ .

The ARIMA process is a first attempt model in many time series applications. If models that include MEPA variables cannot improve upon the above SIR plus ARIMA model, then MEPA may not add any predictive value over using lagged values of incident cases alone.

The above VARMA and ARIMA models were fit using the *statsmodels* package in Python [50].

### Predictive Scoring

Forecasts were scored using the weighted interval score (WIS) over  $K$  central quantiles [51].

$$WIS_{\alpha_{(0,K)}}(F, c) = \frac{1}{K+1/2} \left( w_0 \times |c - m| + \sum_{k=1}^K \{w_k \times IS_{\alpha_k}(F, c)\} \right)$$

where the interval score ( $IS_{\alpha_k}$ ) is

$$IS(\alpha)(F, y) = (u - l) + \frac{2}{\alpha}(l - c)\mathbb{1}(c < l) + \frac{2}{\alpha}(c - u)\mathbb{1}(c > u)$$

and where  $F$  is a predictive cumulative distribution function,  $\mathbb{1}(x)$  is an indicator function, the value  $u$  represents the  $(1-\alpha/2)$  quantile of  $F$ ,  $l$  represents the  $\alpha/2$  quantile of  $F$ ,  $m$  represents the median or 0.50 quantile, and  $c$  is the eventually reported truth [52]. Moreover, weight  $w_0$  equals 1/2 and  $w_k = \alpha_k/2$ .

The WIS and interval score are negatively sensed, with larger values indicating worse predictive performance compared to smaller values. The best possible WIS is 0, and the worst possible WIS is positive infinity.

## Results

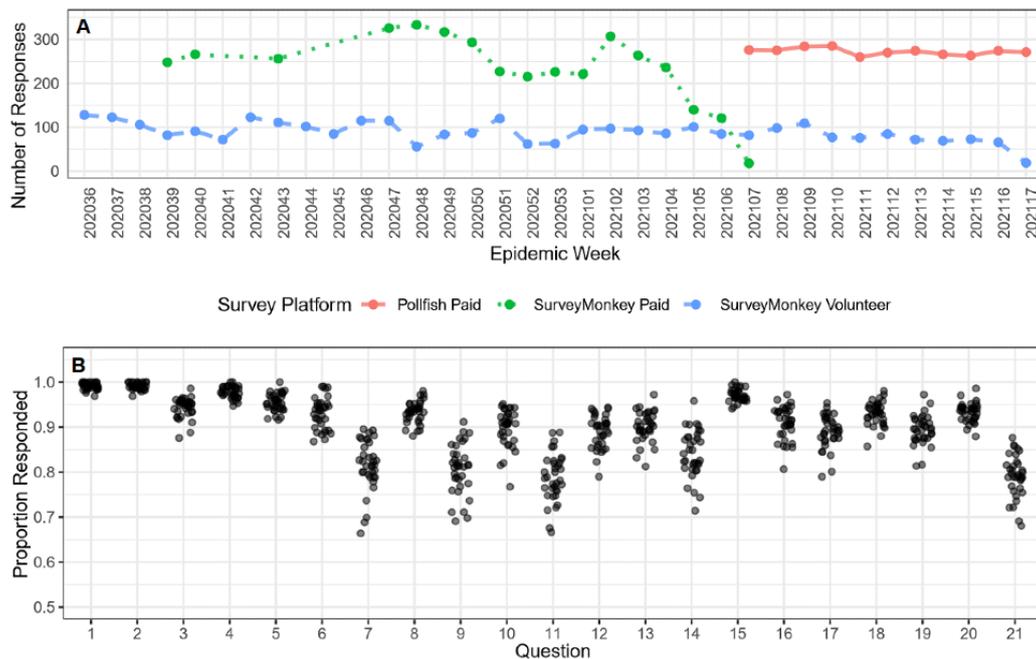
### Overview

Comparison of the response rates across the 2 survey platforms (ie, SurveyMonkey and Pollfish) revealed that sample sizes each week were consistently higher following the switch to Pollfish. The sample was mostly geographically representative of the US population with slight oversampling in the Northeast. MEPA values were more variable at the beginning of the survey period than at the end, suggesting either that responses became more consistent over time or that larger sample sizes throughout the survey period resulted in lower response variability. A clustering analysis revealed that survey questions could be clustered into 4 groups based on question type, suggesting that future surveys might be more efficient by targeting these question types using fewer questions. A correlation analysis revealed reasonably strong correlations between several MEPA time series and incident COVID-19 cases 1-4 weeks ahead. Several MEPA time series also increased the predictive accuracy of a forecasting model of incident COVID-19 cases 1-4 weeks ahead.

### Survey Platform Response Rates

SurveyMonkey surveys received an average of 236.06 (SD 81.14) compensated responses per week and an average of 88.80 (SD 22.68) volunteer responses per week, revealing that response rates for paid surveys were higher but more variable across weeks than volunteer survey responses. Pollfish surveys received an average of 272.55 (SD 7.80) compensated responses per week, and volunteer responses were not collected on the Pollfish system. Overall, sample sizes each week were consistently higher following the switch to Pollfish (Figure 1A).

**Figure 1.** (A) The number of participant responses per epidemic week for the Pollfish platform (red) and for those who submitted responses on SurveyMonkey who were compensated (green) and who were volunteers (blue). (B) The proportion of participants who responded to each question in a given epidemic week. Volunteers made consistent contributions each week as did the Pollfish participants who were compensated, while the number of compensated participant contributions on the SurveyMonkey platform varied. Questions with a lower proportion of responses corresponded to those questions that asked about nonpharmaceutical intervention behaviors that were more difficult to observe, such as visitation rules at senior living facilities (question 7), whether members of the community received antibody testing (question 11), and quarantine of recent travelers (question 14).



### Question Response Rates

The mean percentage of questions that a participant answered was 87.89% (SD 6.15%) (Figure 1B). Questions 1 through 5 and question 15 were answered on average 94.98% (SD 1.47%) of the time, while questions 7, 9, 11, 14, and 21 had the lowest probability of responses, with an average response rate of 78.63% (SD 2.09%).

### Representative Sampling

States from which most responses were collected included California (956/10,120, 9.5%), New York (876/10,120, 8.7%), Pennsylvania (678/10,120, 6.7%), Texas (645/10,120, 6.4%), and Florida (456/10,120, 4.5%).

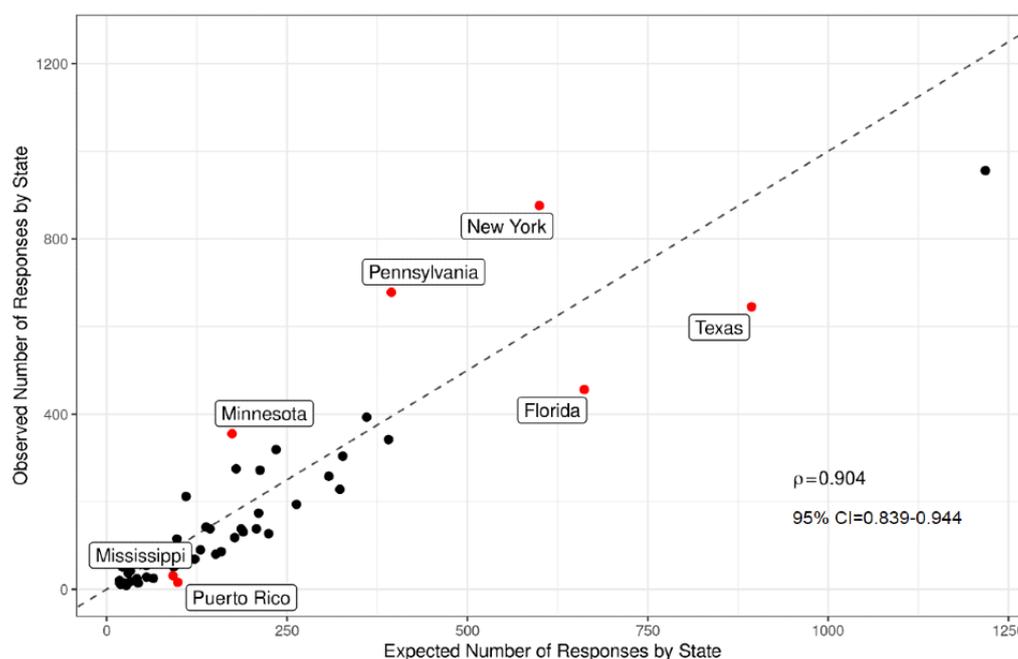
The correlation between the observed frequency of responses and expected frequency was 0.90 (95% CI 0.84-0.94;  $P < .001$ ) and suggested that the response rates were proportional to the population at the state level. We compared for each state the proportion of observed responses to the proportion of individuals in that state according to the census (see Multimedia Appendix 2 for the observed proportion, expected proportion, and relative difference).

Seven states deviated from the expected response rates by more than 9 SDs. Four states were underrepresented (Mississippi, Puerto Rico, Florida, and Texas), and 3 states were overrepresented (Minnesota, Pennsylvania, and New York) (Figure 2). Pennsylvania was the most overrepresented state.

When both compensated and volunteer responses were included, the response frequency in Pennsylvania was 10 SDs above the expectation and when volunteer responses were removed the response frequency decreased to 3.5 SDs below the expectation.

To assess how switching survey platforms in the midst of data collection may have impacted the results, we analyzed whether the representativeness of the sample changed depending on the survey platform. We computed the average relative difference between expected and observed responses across all states, and compared this measure across survey platforms. This analysis revealed that the state residency of paid participants (ie, not volunteers) from SurveyMonkey was more representative of the US population (mean  $-0.599$ , SE 0.015) compared with the state residency of paid participants from Pollfish (mean  $-0.751$ , SE 0.019;  $t_{51}=7.58$ ;  $P < .001$ ).

**Figure 2.** The number of observed responses to the surveys summed over the survey period (vertical axis) compared to the expected number of total responses according to the census (horizontal axis). The dashed line indicates if the observed and expected numbers of responses equal one another. Some states are oversampled and undersampled.



## MEPA Over Time

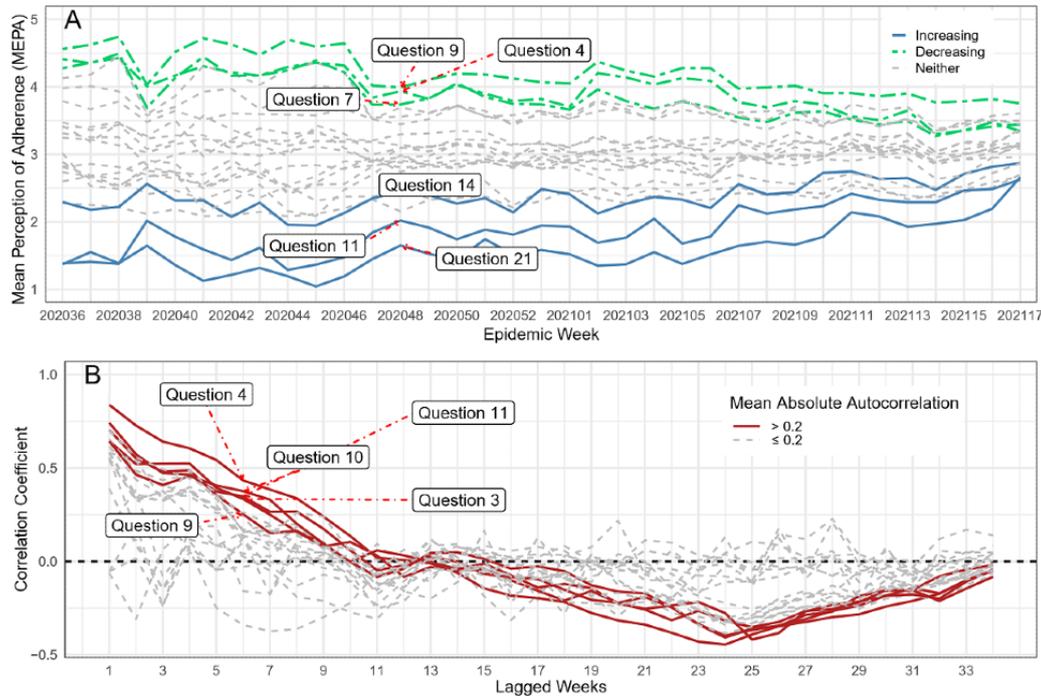
MEPA increased the most from the start to the end of the survey for the following 3 questions: question 21 ( $\Delta\text{mean}_{21} = \text{mean}_{21, \text{week}36} - \text{mean}_{21, \text{week}1} = 1.29$ ) that asked participants about knowledge of their state policies and whether “violations of COVID restrictions result in fines or police enforcement;” question 11 ( $\Delta\text{mean}_{11} = 1.24$ ) that asked how frequently community members follow recommendations to seek “[...] antibody testing to detect prior infection;” and question 14 ( $\Delta\text{mean}_{14} = 0.57$ ) that asked participants how frequently members of their community quarantine after traveling (Figure 3A).

MEPA decreased the most from the start to the end of the survey for the following 3 questions: question 7 ( $\Delta\text{mean}_7 = -1.07$ ) that asked participants how frequently restrictions are placed on visiting senior living facilities; question 4 ( $\Delta\text{mean}_4 = -0.83$ ) that asked how frequently restaurants have reduced seating capacity; and question 9 ( $\Delta\text{mean}_9 = -0.81$ ) that asked about the frequency of special protection in hospitals when treating patients with COVID-19.

The SD between MEPA values at the beginning of the survey period ( $\text{SD}_{\text{beginning}} = 0.89$ ) was larger than the SD between MEPA values at the end of the survey period ( $\text{SD}_{\text{end}} = 0.33$ ) (Figure 3A). The mean MEPA value over all 21 questions remained similar over the course of the survey ( $\text{mean}_{\text{beginning}} = 3.15$ ,  $\text{mean}_{\text{end}} = 3.14$ ). This result could be due to either a convergence in perceptions over time or reduced variability due to increased sample sizes throughout the survey period.

The estimated correlation between MEPA values at time  $t$  and  $t-l$  was greater than 0.35 for lags of up to 4 weeks ( $l=4$ ) for a majority of MEPA time series (Figure 3B) and suggested that many MEPA time series contain more structure than a random walk. Responses to the following 5 survey questions had a mean absolute autocorrelation greater than 0.2: question 3 ([...] staying at home), question 4 ([...] restaurants complying with Centers for Disease Control and Prevention [CDC] recommendations to have reduced seating), question 9 ([...] special protection in hospital areas that treat COVID patients), question 10 ([...] get tested for active virus), and question 11 ([...] get antibody testing to detect prior infection). The mean absolute autocorrelation for these 5 questions across 34 lagged weeks was above 0.2. A more detailed view of autocorrelation for a lag of 1 week has been provided in Multimedia Appendix 5.

**Figure 3.** (A) Mean perception of adherence (MEPA) for 21 questions asked over the survey period. (B) Autocorrelation for all 21 MEPA time series for a lag of 1 to 34 weeks. Perceptions of adherence for questions that asked about state policies (question 21) and antibody testing practices (question 11) show an increase over the survey period, while perceptions of adherence for questions that asked about restrictions placed on senior living facilities (question 7) and restaurants (question 4) show a decrease. Mean absolute autocorrelations for 5 questions across 34 lagged weeks are above 0.2. The estimated correlation between MEPA values at time  $t$  and  $t-1$  is greater than 0.35 for lags of up to 4 weeks ( $l=4$ ) for a majority of MEPA time series. MEPA time series appear to contain more structure than a random walk, suggesting that crowdsourced perceptions may be a useful signal for predicting incident cases.



### Clustering Questions According to Similarities in Responses Over Time

MEPA time series were grouped into the following 4 clusters (Figure 4A and B): (1) cluster of questions with values between 2.5 and 3.5 (ie, low to medium adherence; Figure 4C), (2) cluster with values that decreased over time (Figure 4C), (3) cluster with values near 2.25 at the beginning of the survey and that increased over time (Figure 4C), and (4) cluster with values near 1.25 at the beginning of the survey and that increased over time, ending above 2.50 by the end of the survey (Figure 4C).

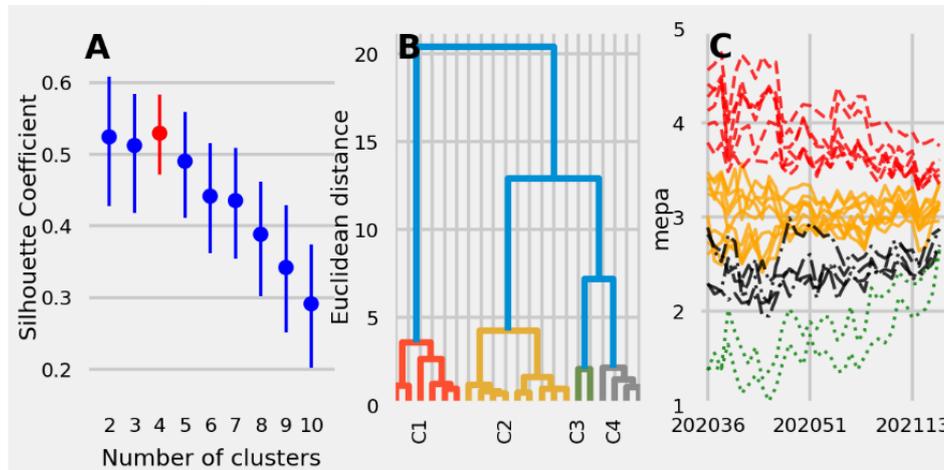
Cluster quality as measured by the silhouette coefficient was the highest when grouping MEPA time series into 4 clusters; however, the silhouette coefficient for 4 clusters was similar to the silhouette coefficient for 2 and 3 clusters (Figure 4A). In the cluster in Figure 4C, there may exist 2 clusters—one with

increasing adherence over time and another with decreasing adherence over time.

MEPA time series within the same cluster asked participants about similar adherence behaviors. Questions corresponding to avoidance behaviors (questions 2, 12, and 15) were more similar to one another than the other questions, as were questions that asked about limitations to businesses (questions 4 and 6), awareness of the high infectivity rate of the virus at a local level (questions 2 and 13), and awareness at the state level (questions 16 and 17). These results suggested that participants might have considered groups of questions in similar ways (eg, those related to avoidance), which suggests that future surveys might benefit from targeting these factors more directly.

For autocorrelations between MEPA responses 1–4 weeks ahead across the different clusters, see Multimedia Appendix 6.

**Figure 4.** Hierarchical clustering of 21 mean perception of adherence (MEPA) time series using Euclidean distance as a measure of dissimilarity between 2 time series. (A) Silhouette coefficients for 2-10 clusters of MEPA time series. (B) Dendrogram that reports questions on the horizontal axis and dissimilarity between individual questions or clusters on the vertical axis. (C) MEPA time series clustered into 4 groups corresponding to the highest silhouette coefficient. Because MEPA time series can be separated into similar groups, a smaller survey may be able to capture the same patterns of the US public’s perceptions of adherence to nonpharmaceutical interventions.

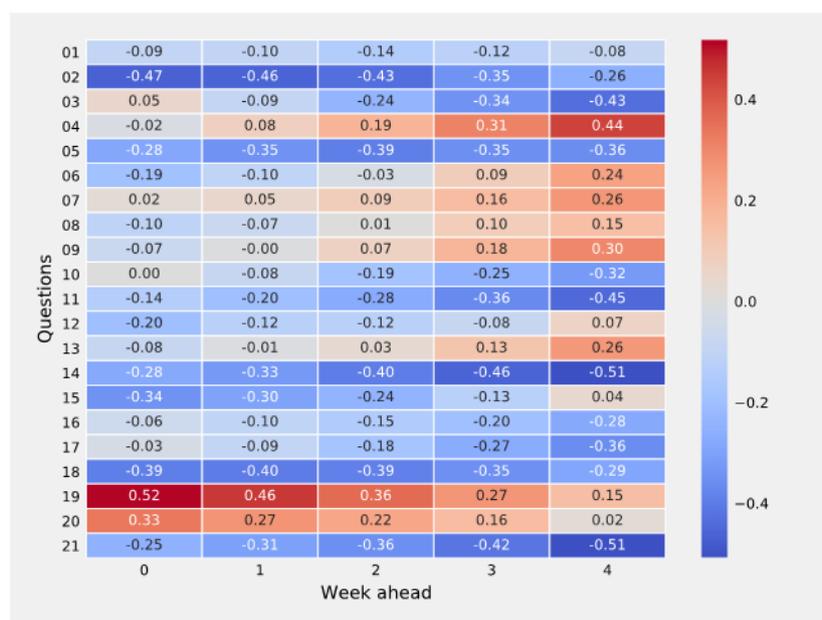


**Correlation Between Perceptions of Adherence and Reported Incident Cases**

The estimated correlation ( $\rho$ ) between the MEPA time series representing responses to the question “What percent of people in your community do you notice are usually maintaining social distance?” and incident cases 1 week ahead was  $-0.46$  (95% CI  $-0.69$  to  $-0.15$ ). Moreover, the correlation ( $\rho$ ) was  $-0.3$  (95% CI  $-0.67$  to  $-0.12$ ) for incident cases 2 weeks ahead,  $-0.35$  (95% CI  $-0.61$  to  $-0.02$ ) for those 3 weeks ahead, and  $-0.26$  (95% CI  $-0.55$  to  $0.08$ ) for those 4 weeks ahead (Figure 5). The MEPA time series for the question “In your state, what percent

of colleges are closed or holding only remote classes?” had an estimated correlation ( $\rho$ ) of  $0.46$  (95% CI  $0.15$  to  $0.69$ ) for cases 1 week ahead. Moreover, the correlations ( $\rho$ ) were  $0.36$  (95% CI  $0.04$  to  $0.62$ ),  $0.27$  (95% CI  $-0.07$  to  $0.55$ ), and  $0.15$  (95% CI  $-0.19$  to  $0.46$ ) for reported incident cases 2 weeks, 3 weeks, and 4 weeks ahead, respectively, at the US national level (Figure 5, row 19). Correlation coefficients and 95% CIs for each question are available in Multimedia Appendix 3. Taken together, these results show that changes in the perceptions of NPI compliance (ie, MEPA time series) are associated with changes in COVID-19 incident cases.

**Figure 5.** Linear correlation between 21 mean perception of adherence (MEPA) time series associated with questions about the perception of adherence and incident cases 1-4 weeks ahead at the US national level. The correlation between question 2 that asked “What percent of people in your community do you notice are usually wearing a mask in public?” and incident cases 1-4 weeks ahead was  $-0.26$  or lower, and the correlation between question 19 that asked “In your state, what percent of colleges are closed or holding only remote classes?” and cases 1-3 weeks ahead was  $0.27$  or higher. Select crowdsourced perceptions of adherence to nonpharmaceutical interventions correlated with short-range and long-range reported incident cases at the national level.



## Out-of-Sample Improvement in Forecasting With the Crowdsourced MEPA

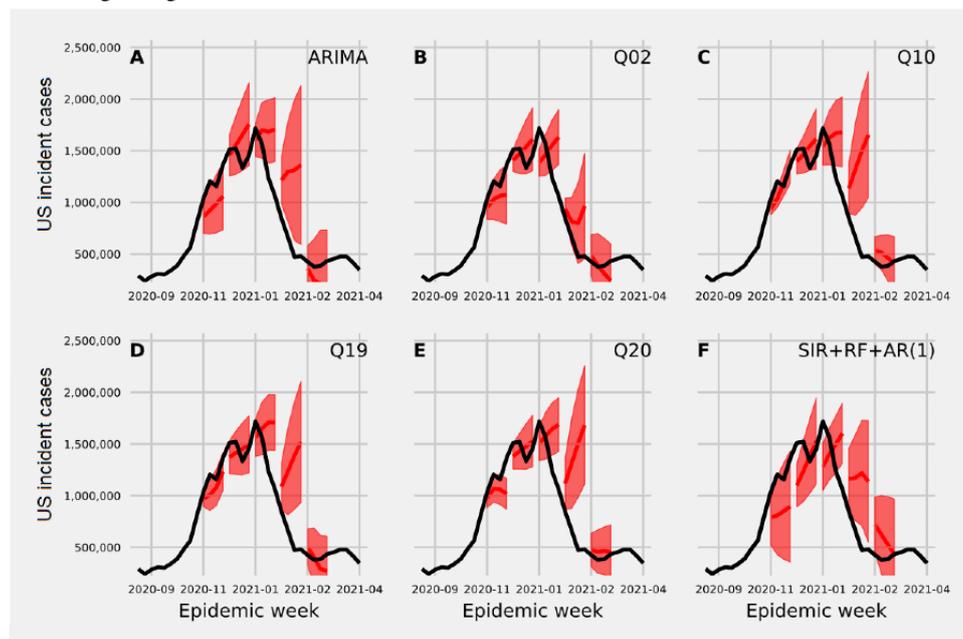
Models that included both historical counts of US national incident cases and MEPA data changed the forecast trajectory and the width of prediction intervals compared to a model that only took into account the past time series of incident US national cases (Figure 6). The model that included historical counts and a random forecast regression incorporating all MEPA data proposed a similar trajectory to the ARIMA (control) model that included only case data, had wider prediction intervals before the peak of reported cases, and had a smaller prediction interval just after the peak of reported cases (Figure 6A and F).

The proportion of times a forecast that included a single MEPA time series generated a smaller (improved) WIS compared to a model that did not use MEPA, was above 50% for the majority of adherence questions for forecast horizons of 1-3 weeks ahead (Figure 7). MEPA most improved forecasts 2 weeks ahead. The MEPA time series corresponding to the questions “What percent of people usually stay home?” “How common do people follow recommendations to receive antibody testing?” and “How

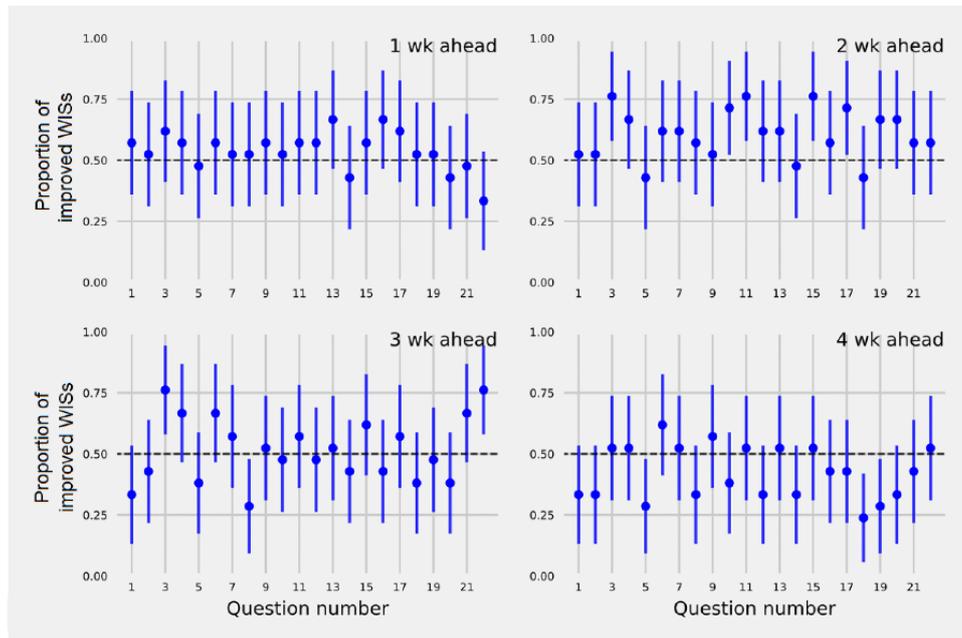
common do people in your community follow guidelines to limit large gatherings?” improved 76% (95% CI 58%-94%) of forecasts 2 weeks ahead. For 3 weeks ahead, the question “What percent of people usually stay home?” improved 76% (95% CI 58%-94%) of forecasts and the machine learning model that incorporated all adherence questions improved 76% (95% CI 58%-94%) of forecasts. Including MEPA data improved forecasts 4 weeks ahead minimally and for only a small set of questions.

Compared with the control model, including MEPA data improved forecast accuracy 1-4 weeks ahead (ie, reduced WIS) at and after the peak reported number of incident cases (Figure 8). Forecasts 1 week ahead showed consistent small gains in forecast accuracy over time (Figure 8A). Forecasts 2 and 3 weeks ahead showed large gains in forecast accuracy at and just after the peak number of incident cases (Figure 8B and C), and improvements in forecast accuracy 4 weeks ahead appeared near the peak number of cases (Figure 8D). Overall, these results revealed that certain perceptions of NPI compliance can be useful signals in a model predicting COVID-19 incident cases.

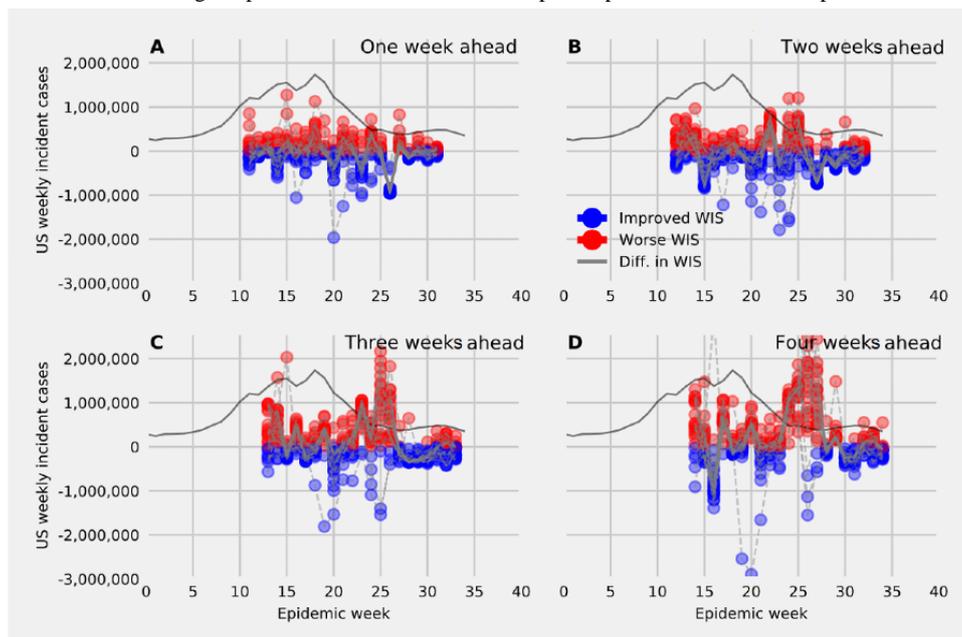
**Figure 6.** Forecasts of US national incident cases 1-4 week ahead at 6 time points throughout the survey period by first fitting an SIR (susceptible, infected, and removed) model and then modeling the residuals by (A) fitting an autoregressive model with 1 lag, (B-E) fitting a vector autoregression moving average that includes the residual time series and mean perception of adherence (MEPA) values for select questions, and (F) fitting a random forecast to residuals including MEPA values for all questions asked of participants plus an AR(1) model. AR(1): autoregression with lag of 1; ARIMA: autoregressive integrated moving average; RF: random forest.



**Figure 7.** The proportion and 95% CIs of weighted interval scores (WISs) that were improved (smaller) for an SIR (susceptible, infected, and removed) plus vector autoregression moving average (VARMA) model that included mean perceived adherence (MEPA) time series 1-21 compared to the control SIR model without using a MEPA time series for forecasts 1-4 weeks ahead. An additional model, to the right of model number 21, is an SIR model plus a random forecast that includes all 21 MEPA time series and an ARIMA to model residuals. The majority of MEPA time series improved forecasts of incident cases 1 and 2 weeks ahead. A smaller number of MEPA time series improved forecasts 3 weeks ahead, and forecasts 4 weeks ahead were improved only modestly.



**Figure 8.** The differences in weighted interval scores (WISs) for forecasts of US national incident cases (A) 1 week ahead, (B) 2 weeks ahead, (C) 3 weeks ahead, and (D) 4 weeks ahead between models that included 1 mean perceived adherence time series and the control model that used only past incident case data to produce a forecast. The differences in WISs correspond to the forecasted epidemic weeks, not when the forecast was generated. The reported number of incident cases at the US national level is provided in grey. A point represents the difference in the WIS at the specific epidemic week and is colored red when a model weakens predictive performance and blue when this forecast improves upon the control model. Including perceptions of human behavior surrounding nonpharmaceutical interventions improves predictions at and after peak incident cases.



**Figure 15.** Inline graphic 7.

## Discussion

We found that crowdsourced perceptions of adherence to NPIs correlated with incident cases 1-4 weeks ahead at the US national level and that including perceptual data into a computational model improved forecast accuracy 1-3 weeks ahead. Because responses from a crowd can be collected quickly (ie, within hours of distributing an online survey), these responses can be included into a computational model that could provide real-time weekly forecasts of epidemiological targets to organizations such as the CDC.

Since forecasts based on public perceptions are rapid and informative, these forecasts would be highly effective at times following the issuing of new NPI guidelines from state or federal agencies to assess the effectiveness of these new guidelines. Our models could reveal the extent to which people perceive public compliance with these guidelines and how changes in compliance impact the trajectory of an infectious agent, thereby informing public health officials about which interventions are able to curtail risk-seeking behaviors. These forecasts may also be valuable for policy makers and community leaders as they decide, for example, whether college classes should be held in person or remotely.

This work supports the hypothesis that a crowd may be able to assign realistic probabilities to outcomes about community adherence to NPIs in line with recent work, which has shown that lay people can elicit accurate probabilistic predictions of diverse real-world phenomena such as box-office income of a new movie or the impact of an infectious agent [13,53]; however, much more work needs to be completed to assess to what degree including human judgment perceptions improves the predictive accuracy of an infectious disease model (Multimedia Appendix 7). Past literature about lay people's ability to make accurate probabilistic predictions is mixed. Some past work suggests people may not be able to map environmental cues to accurate probabilities of outcomes [54], while other work has shown people's statistical intuitions may overlap with the statistics of their environment [53].

Evidence from this study suggests that participants were able to gauge what activities they were able to observe and predict, and at what spatial level they could make predictions. For instance, participants were given the option to reply "Don't know" or to leave questions blank. Participants responded more often to questions that were related to their environment, such as the proportion of people wearing masks, and responded less

often to questions that were not related to their environment, such as restrictions on visitation to senior living centers. Survey questions during the initial pilot stage of the study asked participants to make predictions at the state level rather than community level, and many participants during this pilot stage protested that they could not make reasonable predictions at this level, suggesting that participants have some sense of how far a local community-level prediction could be extrapolated. Lastly, strong correlations between weekly responses to specific NPI questions indicated that the judgment of participants in this study was consistent (see Multimedia Appendix 5). Our results may support the idea that human judgment is predictive of incident cases because people can accurately perceive and make inferences about their surroundings.

However, relying on human judgment presents challenges that are absent when using computational models for prediction. Human judgment is susceptible to a wide array of biases often triggered by subtle changes in how a judgment prompt is presented [55]. Seemingly irrelevant information can have large impacts on judgment. For example, when asked to complete an irrelevant task, such as writing down the last 2 digits of their social security number before bidding on common items (such as a bottle of wine), people with higher social security numbers bid more money on wine than those with lower numbers [56]. Such findings underscore the importance of carefully crafting judgment questions to avoid activating judgment biases. Human judgment data must also be inspected for quality, as participants in this study often left one or more questions blank in a single survey and approximately 2% of participants gave the same response for every question, suggesting that they were not reading the survey items closely. Lastly, recruiting human participants demands time, effort, and money. Recruiting volunteers saves money but demands effort and implies an uncertain number of responses, which can be challenging when collecting data in response to a time-sensitive event such as an epidemic or pandemic. Participation rates in this study tended to increase throughout the data collection period, which created difficulties in assessing whether changes in MEPA over time were driven more by changes in perceived adherence or by changes in participation rates.

There are several limitations to address in future work. One limitation that we wish to overcome is that participants were not traced longitudinally, and so, we could not analyze how responses from individuals changed over time. Another limitation is that emails used to solicit volunteer participants

contained a link to a summary of the findings from previous months of data collection. While this may have added value to a participant's experience in the study, it may have biased their subsequent responses by anchoring their judgments to those summary values [56]. Another limitation arose from switching survey platforms (from SurveyMonkey to Pollfish) in the midst of data collection. The need for this switch was driven by a sudden decrease in the ability of SurveyMonkey to provide the requested number of paid responses each week (see Figure 1A). This switch seemed to have an impact on the geographical representativeness of the sample, as Pollfish provided a less representative sample than SurveyMonkey. Because switching survey platforms was confounded with both number of responses and epidemic week, the impact that switching survey platforms may have had on responses is largely unclear. Additionally, variable sampling rates across states created difficulties in estimating predictions at the state level. Oversampling from states with lower populations would ensure that a predictive model has sufficient data for estimating reliable predictions. No other demographic information was consistently collected throughout the surveys, and so, we were not able to assess whether the sample was representative for other demographic dimensions. Finally, there is evidence to suggest that self-expression may vary by geographic location [57]. Future research should consider how location and surrounding demographics may impact perceptions by, for example, leading to an overestimation of the prevalence of mask wearing in more densely populated areas.

Future research should explore whether more accurate and calibrated predictions of incident cases from human judgments

can be made by matching the spatial scale of the questions posed to the crowd with the epidemiological target of interest. Instead of predicting incident cases at the national level, much stronger connections may be observed between state- or community-level judgments and state- or community-level incident cases. For example, one could investigate whether the accuracy of forecasts depends on factors such as the geographical size of the state (eg, Texas vs Delaware) or ethnic diversity (eg, California vs West Virginia). Additionally, respondents could be asked to judge compliance specifically at the level of their county, and then, these judgments could be added to a model that produces county-level predictions. Strong predictions at this local level would be valuable for community leaders when deciding, for example, whether a town hall meeting should be in person or remote. A significant challenge to estimating these local predictions is collecting enough responses from a given community over time, which, as mentioned above, can be remedied by targeting and oversampling from areas of interest to make local predictions. Future research should also explore whether perceptions of NPI compliance can predict other epidemiological targets. While we focused on incident cases in this study, our current methods should scale to other prediction outcomes of interest, such as COVID-19 hospitalizations and deaths.

Crowdsourced perceptions of human behavior, such as nonpharmaceutical adherence, may be a fast and informative signal that can improve probabilistic forecasts of the trajectory of an infectious agent and may have important implications for policy around infectious diseases.

---

## Acknowledgments

We thank the following members of the COVID Mitigation Monitoring Project (CMMP): Dale Hall, FSA, CERA, MAAA, CFA; Robert Wolf, FSA, CFA, CERA; Kailan Shang, FSA, CFA, PRM, SCJP; and John Stark, FSA, CERA. We thank the Society of Actuaries for support in data collection. We also wish to thank all participants who offered their perceptions over the survey period.

---

## Conflicts of Interest

None declared.

---

## Multimedia Appendix 1

Survey provided to participants to capture perceptions of adherence to nonpharmaceutical interventions.

[\[DOCX File , 942 KB-Multimedia Appendix 1\]](#)

---

## Multimedia Appendix 2

Observed and expected proportions of participants for each state.

[\[DOCX File , 26 KB-Multimedia Appendix 2\]](#)

---

## Multimedia Appendix 3

Correlation between mean perceived adherence and US national incident cases.

[\[DOCX File , 39 KB-Multimedia Appendix 3\]](#)

---

## Multimedia Appendix 4

SIR (susceptible, infected, and removed) model fit to US national incident cases.

[\[DOCX File , 68 KB-Multimedia Appendix 4\]](#)

## Multimedia Appendix 5

Autocorrelation of 1 week for mean perceived adherence time series.

[\[DOCX File , 447 KB-Multimedia Appendix 5\]](#)

## Multimedia Appendix 6

Bivariate relationships between question clusters and incident cases.

[\[DOCX File , 287 KB-Multimedia Appendix 6\]](#)

## Multimedia Appendix 7

Hypothesis testing across models in forecasting with crowdsourced mean perceived adherence.

[\[DOCX File , 20 KB-Multimedia Appendix 7\]](#)

## References

1. Lutz CS, Huynh MP, Schroeder M, Anyatonwu S, Dahlgren FS, Danyluk G, et al. Applying infectious disease forecasting to public health: a path forward using influenza forecasting examples. *BMC Public Health* 2019 Dec 10;19(1):1659 [FREE Full text] [doi: [10.1186/s12889-019-7966-8](https://doi.org/10.1186/s12889-019-7966-8)] [Medline: [31823751](https://pubmed.ncbi.nlm.nih.gov/31823751/)]
2. Pollett S, Johansson MA, Reich NG, Brett-Major D, Del Valle SY, Venkatramanan S, et al. Recommended reporting items for epidemic forecasting and prediction research: The EPIFORGE 2020 guidelines. *PLoS Med* 2021 Oct 19;18(10):e1003793 [FREE Full text] [doi: [10.1371/journal.pmed.1003793](https://doi.org/10.1371/journal.pmed.1003793)] [Medline: [34665805](https://pubmed.ncbi.nlm.nih.gov/34665805/)]
3. Del Valle SY, McMahon BH, Asher J, Hatchett R, Lega JC, Brown HE, et al. Summary results of the 2014-2015 DARPA Chikungunya challenge. *BMC Infect Dis* 2018 May 30;18(1):245 [FREE Full text] [doi: [10.1186/s12879-018-3124-7](https://doi.org/10.1186/s12879-018-3124-7)] [Medline: [29843621](https://pubmed.ncbi.nlm.nih.gov/29843621/)]
4. Evans MV, Dallas TA, Han BA, Murdock CC, Drake JM. Data-driven identification of potential Zika virus vectors. *Elife* 2017 Feb 28;6:e22053 [FREE Full text] [doi: [10.7554/eLife.22053](https://doi.org/10.7554/eLife.22053)] [Medline: [28244371](https://pubmed.ncbi.nlm.nih.gov/28244371/)]
5. Johansson M, Apfeldorf K, Dobson S, Devita J, Buczak A, Baugher B, et al. An open challenge to advance probabilistic forecasting for dengue epidemics. *Proc Natl Acad Sci U S A* 2019 Nov 26;116(48):24268-24274 [FREE Full text] [doi: [10.1073/pnas.1909865116](https://doi.org/10.1073/pnas.1909865116)] [Medline: [31712420](https://pubmed.ncbi.nlm.nih.gov/31712420/)]
6. Kobres P, Chretien J, Johansson MA, Morgan JJ, Whung P, Mukundan H, et al. A systematic review and evaluation of Zika virus forecasting and prediction research during a public health emergency of international concern. *PLoS Negl Trop Dis* 2019 Oct 4;13(10):e0007451 [FREE Full text] [doi: [10.1371/journal.pntd.0007451](https://doi.org/10.1371/journal.pntd.0007451)] [Medline: [31584946](https://pubmed.ncbi.nlm.nih.gov/31584946/)]
7. Reich NG, McGowan CJ, Yamana TK, Tushar A, Ray EL, Osthus D, et al. Accuracy of real-time multi-model ensemble forecasts for seasonal influenza in the U.S. *PLoS Comput Biol* 2019 Nov 22;15(11):e1007486 [FREE Full text] [doi: [10.1371/journal.pcbi.1007486](https://doi.org/10.1371/journal.pcbi.1007486)] [Medline: [31756193](https://pubmed.ncbi.nlm.nih.gov/31756193/)]
8. Viboud C, Sun K, Gaffey R, Ajelli M, Fumanelli L, Merler S, RAPIDD Ebola Forecasting Challenge group. The RAPIDD ebola forecasting challenge: Synthesis and lessons learnt. *Epidemics* 2018 Mar;22:13-21 [FREE Full text] [doi: [10.1016/j.epidem.2017.08.002](https://doi.org/10.1016/j.epidem.2017.08.002)] [Medline: [28958414](https://pubmed.ncbi.nlm.nih.gov/28958414/)]
9. Biggerstaff M, Slayton RB, Johansson MA, Butler JC. Improving Pandemic Response: Employing Mathematical Modeling to Confront Coronavirus Disease 2019. *Clin Infect Dis* 2022 Mar 09;74(5):913-917 [FREE Full text] [doi: [10.1093/cid/ciab673](https://doi.org/10.1093/cid/ciab673)] [Medline: [34343282](https://pubmed.ncbi.nlm.nih.gov/34343282/)]
10. Cramer E, Ray E, Lopez V, Bracher J, Brennen A, Castro Rivadeneira A, et al. Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States. *Proc Natl Acad Sci U S A* 2022 Apr 12;119(15):e2113561119 [FREE Full text] [doi: [10.1073/pnas.2113561119](https://doi.org/10.1073/pnas.2113561119)] [Medline: [35394862](https://pubmed.ncbi.nlm.nih.gov/35394862/)]
11. Evan LR, Nutchu W, Jarad N, Abdul HK, Katie H, Estee YC, et al. Ensemble Forecasts of Coronavirus Disease 2019 (COVID-19) in the U.S. *MedRxiv*. 2020. URL: <https://www.medrxiv.org/content/10.1101/2020.08.19.20177493v1> [accessed 2022-11-16]
12. Hanea AM, Hemming V, Nane GF. Uncertainty Quantification with Experts: Present Status and Research Needs. *Risk Anal* 2022 Feb;42(2):254-263. [doi: [10.1111/risa.13718](https://doi.org/10.1111/risa.13718)] [Medline: [33629402](https://pubmed.ncbi.nlm.nih.gov/33629402/)]
13. McAndrew T, Wattanachit N, Gibson GC, Reich NG. Aggregating predictions from experts: a review of statistical methods, experiments, and applications. *Wiley Interdiscip Rev Comput Stat* 2021 Jun 16;13(2):e1514 [FREE Full text] [doi: [10.1002/wics.1514](https://doi.org/10.1002/wics.1514)] [Medline: [33777310](https://pubmed.ncbi.nlm.nih.gov/33777310/)]
14. Bosse NI, Abbott S, Bracher J, Hain H, Quilty BJ, Jit M, Centre for the Mathematical Modelling of Infectious Diseases COVID-19 Working Group, et al. Comparing human and model-based forecasts of COVID-19 in Germany and Poland. *PLoS Comput Biol* 2022 Sep;18(9):e1010405 [FREE Full text] [doi: [10.1371/journal.pcbi.1010405](https://doi.org/10.1371/journal.pcbi.1010405)] [Medline: [36121848](https://pubmed.ncbi.nlm.nih.gov/36121848/)]
15. Farrow DC, Brooks LC, Hyun S, Tibshirani RJ, Burke DS, Rosenfeld R. A human judgment approach to epidemiological forecasting. *PLoS Comput Biol* 2017 Mar 10;13(3):e1005248 [FREE Full text] [doi: [10.1371/journal.pcbi.1005248](https://doi.org/10.1371/journal.pcbi.1005248)] [Medline: [28282375](https://pubmed.ncbi.nlm.nih.gov/28282375/)]

16. McAndrew T, Reich NG. An expert judgment model to predict early stages of the COVID-19 pandemic in the United States. *PLoS Comput Biol* 2022 Sep;18(9):e1010485 [FREE Full text] [doi: [10.1371/journal.pcbi.1010485](https://doi.org/10.1371/journal.pcbi.1010485)] [Medline: [36149916](https://pubmed.ncbi.nlm.nih.gov/36149916/)]
17. Recchia G, Freeman ALJ, Spiegelhalter D. How well did experts and laypeople forecast the size of the COVID-19 pandemic? *PLoS One* 2021 May 5;16(5):e0250935 [FREE Full text] [doi: [10.1371/journal.pone.0250935](https://doi.org/10.1371/journal.pone.0250935)] [Medline: [33951092](https://pubmed.ncbi.nlm.nih.gov/33951092/)]
18. Clemen RT, Winkler RL. Combining Probability Distributions From Experts in Risk Analysis. *Risk Analysis* 2006 May 29;19(2):187-203. [doi: [10.1111/j.1539-6924.1999.tb00399.x](https://doi.org/10.1111/j.1539-6924.1999.tb00399.x)]
19. Grigore B, Peters J, Hyde C, Stein K. EXPLICIT: a feasibility study of remote expert elicitation in health technology assessment. *BMC Med Inform Decis Mak* 2017 Sep 04;17(1):131 [FREE Full text] [doi: [10.1186/s12911-017-0527-0](https://doi.org/10.1186/s12911-017-0527-0)] [Medline: [28870196](https://pubmed.ncbi.nlm.nih.gov/28870196/)]
20. Hemming V, Burgman MA, Hanea AM, McBride MF, Wintle BC. A practical guide to structured expert elicitation using the IDEA protocol. *Methods Ecol Evol* 2017 Sep 05;9(1):169-180. [doi: [10.1111/2041-210x.12857](https://doi.org/10.1111/2041-210x.12857)]
21. Winkler RL, Grushka-Cockayne Y, Lichtendahl KC, Jose VRR. Probability Forecasts and Their Combination: A Research Perspective. *Decision Analysis* 2019 Dec;16(4):239-260. [doi: [10.1287/deca.2019.0391](https://doi.org/10.1287/deca.2019.0391)]
22. Cooke RM. *Experts in Uncertainty: Opinion and Subjective Probability in Science*. Oxford, United Kingdom: Oxford University Press; 1991.
23. Burgman MA, McBride M, Ashton R, Speirs-Bridge A, Flander L, Wintle B, et al. Expert status and performance. *PLoS One* 2011 Jul 29;6(7):e22998 [FREE Full text] [doi: [10.1371/journal.pone.0022998](https://doi.org/10.1371/journal.pone.0022998)] [Medline: [21829574](https://pubmed.ncbi.nlm.nih.gov/21829574/)]
24. Spence MT, Brucks M. The Moderating Effects of Problem Characteristics on Experts' and Novices' Judgments. *Journal of Marketing Research* 2018 Dec 20;34(2):233-247. [doi: [10.1177/002224379703400204](https://doi.org/10.1177/002224379703400204)]
25. Stewart TR, Roebber PJ, Bosart LF. The Importance of the Task in Analyzing Expert Judgment. *Organizational Behavior and Human Decision Processes* 1997 Mar;69(3):205-219. [doi: [10.1006/obhd.1997.2682](https://doi.org/10.1006/obhd.1997.2682)]
26. Kahneman D, Tversky A. Subjective probability: A judgment of representativeness. *Cognitive Psychology* 1972 Jul;3(3):430-454. [doi: [10.1016/0010-0285\(72\)90016-3](https://doi.org/10.1016/0010-0285(72)90016-3)]
27. Kahneman D, Tversky A. On the psychology of prediction. *Psychological Review* 1973 Jul;80(4):237-251. [doi: [10.1037/h0034747](https://doi.org/10.1037/h0034747)]
28. Tversky A, Kahneman D. Availability: A heuristic for judging frequency and probability. *Cognitive Psychology* 1973 Sep;5(2):207-232. [doi: [10.1016/0010-0285\(73\)90033-9](https://doi.org/10.1016/0010-0285(73)90033-9)]
29. Tversky A, Kahneman D. Evidential impact of base rates. DTIC. 1981. URL: <https://apps.dtic.mil/sti/pdfs/ADA099501.pdf> [accessed 2022-11-16]
30. Gigerenzer G, Gaissmaier W. Heuristic decision making. *Annu Rev Psychol* 2011 Jan 10;62(1):451-482. [doi: [10.1146/annurev-psych-120709-145346](https://doi.org/10.1146/annurev-psych-120709-145346)] [Medline: [21126183](https://pubmed.ncbi.nlm.nih.gov/21126183/)]
31. Shah AK, Oppenheimer DM. Heuristics made easy: an effort-reduction framework. *Psychol Bull* 2008 Mar;134(2):207-222. [doi: [10.1037/0033-2909.134.2.207](https://doi.org/10.1037/0033-2909.134.2.207)] [Medline: [18298269](https://pubmed.ncbi.nlm.nih.gov/18298269/)]
32. Wübben M, Wangenheim FV. Instant Customer Base Analysis: Managerial Heuristics Often “Get It Right”. *Journal of Marketing* 2008 May;72(3):82-93. [doi: [10.1509/jmkg.72.3.82](https://doi.org/10.1509/jmkg.72.3.82)]
33. Ibrahim R, Kim S, Tong J. Eliciting Human Judgment for Prediction Algorithms. *Management Science* 2021 Apr;67(4):2314-2325. [doi: [10.1287/mnsc.2020.3856](https://doi.org/10.1287/mnsc.2020.3856)]
34. Bae S, (Christine) Sung E, Kwon O. Accounting for social media effects to improve the accuracy of infection models: combatting the COVID-19 pandemic and infodemic. *European Journal of Information Systems* 2021 Feb 25;30(3):342-355. [doi: [10.1080/0960085x.2021.1890530](https://doi.org/10.1080/0960085x.2021.1890530)]
35. Hirose H, Wang L. Prediction of Infectious Disease Spread Using Twitter: A Case of Influenza. 2012 Presented at: Fifth International Symposium on Parallel Architectures, Algorithms and Programming; December 17-20, 2012; Taipei, Taiwan. [doi: [10.1109/PAAP.2012.23](https://doi.org/10.1109/PAAP.2012.23)]
36. McDonald DJ, Bien J, Green A, Hu AJ, DeFries N, Hyun S, et al. Can auxiliary indicators improve COVID-19 forecasting and hotspot prediction? *Proc Natl Acad Sci U S A* 2021 Dec 21;118(51):e2111453118 [FREE Full text] [doi: [10.1073/pnas.2111453118](https://doi.org/10.1073/pnas.2111453118)] [Medline: [34903655](https://pubmed.ncbi.nlm.nih.gov/34903655/)]
37. Shen C, Chen A, Luo C, Zhang J, Feng B, Liao W. Using Reports of Symptoms and Diagnoses on Social Media to Predict COVID-19 Case Counts in Mainland China: Observational Infoveillance Study. *J Med Internet Res* 2020 May 28;22(5):e19421 [FREE Full text] [doi: [10.2196/19421](https://doi.org/10.2196/19421)] [Medline: [32452804](https://pubmed.ncbi.nlm.nih.gov/32452804/)]
38. Polgreen PM, Nelson FD, Neumann GR. Use of prediction markets to forecast infectious disease activity. *Clin Infect Dis* 2007 Jan 15;44(2):272-279. [doi: [10.1086/510427](https://doi.org/10.1086/510427)] [Medline: [17173231](https://pubmed.ncbi.nlm.nih.gov/17173231/)]
39. Sell TK, Warmbrod KL, Watson C, Trotochaud M, Martin E, Ravi SJ, et al. Using prediction polling to harness collective intelligence for disease forecasting. *BMC Public Health* 2021 Nov 20;21(1):2132 [FREE Full text] [doi: [10.1186/s12889-021-12083-y](https://doi.org/10.1186/s12889-021-12083-y)] [Medline: [34801014](https://pubmed.ncbi.nlm.nih.gov/34801014/)]
40. Yang S, Santillana M, Kou SC. Accurate estimation of influenza epidemics using Google search data via ARGO. *Proc Natl Acad Sci U S A* 2015 Nov 24;112(47):14473-14478 [FREE Full text] [doi: [10.1073/pnas.1515373112](https://doi.org/10.1073/pnas.1515373112)] [Medline: [26553980](https://pubmed.ncbi.nlm.nih.gov/26553980/)]

41. Zeng C, Zhang J, Li Z, Sun X, Olatosi B, Weissman S, et al. Spatial-Temporal Relationship Between Population Mobility and COVID-19 Outbreaks in South Carolina: Time Series Forecasting Analysis. *J Med Internet Res* 2021 Apr 13;23(4):e27045 [[FREE Full text](#)] [doi: [10.2196/27045](https://doi.org/10.2196/27045)] [Medline: [33784239](https://pubmed.ncbi.nlm.nih.gov/33784239/)]
42. Rabiolo A, Alladio E, Morales E, McNaught AI, Bandello F, Afifi AA, et al. Forecasting the COVID-19 Epidemic by Integrating Symptom Search Behavior Into Predictive Models: Infoveillance Study. *J Med Internet Res* 2021 Aug 11;23(8):e28876 [[FREE Full text](#)] [doi: [10.2196/28876](https://doi.org/10.2196/28876)] [Medline: [34156966](https://pubmed.ncbi.nlm.nih.gov/34156966/)]
43. Funk S, Salathé M, Jansen VAA. Modelling the influence of human behaviour on the spread of infectious diseases: a review. *J R Soc Interface* 2010 Sep 06;7(50):1247-1256 [[FREE Full text](#)] [doi: [10.1098/rsif.2010.0142](https://doi.org/10.1098/rsif.2010.0142)] [Medline: [20504800](https://pubmed.ncbi.nlm.nih.gov/20504800/)]
44. Singhal A, Baxi MK, Mago V. Synergy Between Public and Private Health Care Organizations During COVID-19 on Twitter: Sentiment and Engagement Analysis Using Forecasting Models. *JMIR Med Inform* 2022 Aug 18;10(8):e37829 [[FREE Full text](#)] [doi: [10.2196/37829](https://doi.org/10.2196/37829)] [Medline: [35849795](https://pubmed.ncbi.nlm.nih.gov/35849795/)]
45. Fridman I, Lucas N, Henke D, Zigler CK. Association Between Public Knowledge About COVID-19, Trust in Information Sources, and Adherence to Social Distancing: Cross-Sectional Survey. *JMIR Public Health Surveill* 2020 Sep 15;6(3):e22060 [[FREE Full text](#)] [doi: [10.2196/22060](https://doi.org/10.2196/22060)] [Medline: [32930670](https://pubmed.ncbi.nlm.nih.gov/32930670/)]
46. Braun D, Ingram D, Ingram D, Khan B, Marsh J, McAndrew T. Incorporating crowdsourced perceptions of human behavior into computational forecasts of US national incident cases of COVID-19. *OSF Preprints*. 2022. URL: <https://osf.io/7vrmy> [accessed 2022-11-16]
47. Ingram D, Ingram D. COVID-19 Mitigations in the U.S. September 2020 to April 2021. Society of Actuaries. URL: <https://www.soa.org/globalassets/assets/files/resources/research-report/2021/covid-eight-month-report.pdf> [accessed 2022-11-16]
48. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases* 2020 May;20(5):533-534. [doi: [10.1016/s1473-3099\(20\)30120-1](https://doi.org/10.1016/s1473-3099(20)30120-1)]
49. Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* 1987 Nov;20:53-65. [doi: [10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7)]
50. Seabold S, Perktold J. Statsmodels: Econometric and Statistical Modeling with Python. *SciPy*. URL: <https://conference.scipy.org/proceedings/scipy2010/pdfs/seabold.pdf> [accessed 2022-11-16]
51. Bracher J, Ray EL, Gneiting T, Reich NG. Evaluating epidemic forecasts in an interval format. *PLoS Comput Biol* 2021 Feb;17(2):e1008618 [[FREE Full text](#)] [doi: [10.1371/journal.pcbi.1008618](https://doi.org/10.1371/journal.pcbi.1008618)] [Medline: [33577550](https://pubmed.ncbi.nlm.nih.gov/33577550/)]
52. Gneiting T, Balabdaoui F, Raftery AE. Probabilistic forecasts, calibration and sharpness. *J Royal Statistical Soc B* 2007 Apr;69(2):243-268. [doi: [10.1111/j.1467-9868.2007.00587.x](https://doi.org/10.1111/j.1467-9868.2007.00587.x)]
53. Griffiths TL, Tenenbaum JB. Optimal predictions in everyday cognition. *Psychol Sci* 2006 Sep 17;17(9):767-773. [doi: [10.1111/j.1467-9280.2006.01780.x](https://doi.org/10.1111/j.1467-9280.2006.01780.x)] [Medline: [16984293](https://pubmed.ncbi.nlm.nih.gov/16984293/)]
54. McDowell M, Jacobs P. Meta-analysis of the effect of natural frequencies on Bayesian reasoning. *Psychol Bull* 2017 Dec;143(12):1273-1312. [doi: [10.1037/bul0000126](https://doi.org/10.1037/bul0000126)] [Medline: [29048176](https://pubmed.ncbi.nlm.nih.gov/29048176/)]
55. Tversky A, Kahneman D. Rational Choice and the Framing of Decisions. In: Karpak B, Zionts S, editors. *Multiple Criteria Decision Making and Risk Analysis Using Microcomputers*. NATO ASI Series, vol 56. Berlin, Heidelberg: Springer; 1989:81-126.
56. Ariely D, Loewenstein G, Prelec D. "Coherent Arbitrariness": Stable Demand Curves Without Stable Preferences. *The Quarterly Journal of Economics* 2003 Feb 01;118(1):73-106. [doi: [10.1162/00335530360535153](https://doi.org/10.1162/00335530360535153)]
57. Mitchell L, Frank MR, Harris KD, Dodds PS, Danforth CM. The geography of happiness: connecting twitter sentiment and expression, demographics, and objective characteristics of place. *PLoS One* 2013 May 29;8(5):e64417 [[FREE Full text](#)] [doi: [10.1371/journal.pone.0064417](https://doi.org/10.1371/journal.pone.0064417)] [Medline: [23734200](https://pubmed.ncbi.nlm.nih.gov/23734200/)]

## Abbreviations

- AIC:** Akaike information criterion
- ARIMA:** autoregressive integrated moving average
- CDC:** Centers for Disease Control and Prevention
- IRB:** institutional review board
- MEPA:** mean perceived adherence
- NPI:** nonpharmaceutical intervention
- SIR:** susceptible, infected, and removed
- VARMA:** vector autoregression moving average
- WIS:** weighted interval score

*Edited by A Mavragani, G Eysenbach; submitted 06.05.22; peer-reviewed by R Gore, S Yang; comments to author 12.09.22; revised version received 08.10.22; accepted 09.10.22; published 30.12.22*

*Please cite as:*

*Braun D, Ingram D, Ingram D, Khan B, Marsh J, McAndrew T*

*Crowdsourced Perceptions of Human Behavior to Improve Computational Forecasts of US National Incident Cases of COVID-19: Survey Study*

*JMIR Public Health Surveill 2022;8(12):e39336*

*URL: <https://publichealth.jmir.org/2022/12/e39336>*

*doi: [10.2196/39336](https://doi.org/10.2196/39336)*

*PMID: [36219845](https://pubmed.ncbi.nlm.nih.gov/36219845/)*

©David Braun, Daniel Ingram, David Ingram, Bilal Khan, Jessecae Marsh, Thomas McAndrew. Originally published in JMIR Public Health and Surveillance (<https://publichealth.jmir.org>), 30.12.2022. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Public Health and Surveillance, is properly cited. The complete bibliographic information, a link to the original publication on <https://publichealth.jmir.org>, as well as this copyright and license information must be included.