Incorporating crowdsourced perceptions of human behavior into computational forecasts of US national incident cases of COVID-19

David Braun,[1] Daniel Ingram,[2] David Ingram,[2] Bilal Khan,[3] Jessecae Marsh,[1] and Thomas McAndrew[3,*]

[1] *Department of Psychology, Lehigh University, Bethlehem, Pennsylvania, United States of America*

[2] *Actuarial Risk Management, LTD, Austin, Texas, United States of America*

[3] *College of Health, Lehigh University, Bethlehem, Pennsylvania, United States of America*

(Dated: May 5, 2022)

Computational forecasts of COVID-19 targets may benefit from temporal signals associated with human behavior. Non-pharmaceutical interventions (NPI) have been shown to reduce the spread of an infectious agent, but accurate information about how the general public interprets and acts upon guidelines developed by public health officials is difficult to collect. For 36 weeks from September, 2020 to April, 2021, we asked two crowds twenty one questions about their perceptions of their communities adherence to NPI and public health guidelines and collected 10,120 responses. Crowdsourced NPI signals were mapped to a mean perception of adherence—or MEPA—and included in computational forecasts. Several MEPA signals linearly correlated with one through four week ahead incident cases of COVID-19 at the US national level. Including questions related to masking, testing, and limiting large gatherings increased out of sample predictive performance for 1-3 week ahead probabilistic forecasts of incident cases of COVID-19 when compared to a model that was trained on only past incident cases. In addition, we found that MEPA signals could be clustered which suggests a more focused survey may have sufficed and provided similar performance. Crowdsourced perceptions of non-pharmaceutical adherence may be an important signal to improve forecasts of the trajectory of an infectious agent and increase public health situational awareness.

* mcandrew@lehigh.edu

I. INTRODUCTION

Forecasting the transmission of infectious agents can support decisions made by public health officials and key decision makers [29, 35]. 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 [11, 13, 25, 28, 38, 49]. The COVID-19 pandemic has further highlighted the importance that forecasts play in support of public health situational awareness [3, 10, 36].

The majority of forecasts of an infectious disease are generated by computational models, however past work has shown human judgement is also capable of making accurate predictions of a diverse number of phenomena [20, 31], including infectious agents [4, 14, 30, 37].

Work in human judgement 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 [8, 19, 21, 50]. 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 [9, 19, 21].

Past work has found middling performance when asking those with subject matter expertise to make direct predictions [7, 44]. 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 [45], as well as people's reliance on heuristics to make fast decisions with little information [26, 27, 46, 47]. Humans are subject to several cognitive biases that negatively impact our ability to make sound judgements [16, 42]. That said, there are many examples where predictions based on mental heuristics outperformed computational models [51].

Work on aggregating direct human judgement predictions has focused on adjusting for correlated predictions between individuals, assessing the number of individual predictions to combine, and how to appropriately weight individuals based on past predictive performance [8, 50]. Direct predictions take advantage of a human's ability to build a prediction from available, structured data and information that is typically unavailable to a computational model such as subjective information, intuition, and expertise [23].

Indirect predictions of a future event are collected by either (i) extracting human judgement data from a passive source such as social media [2, 22, 32, 43], (ii) actively asking a crowd about covariates that may be related to the target of interest, or (iii) asking a crowd to take actions in a prediction market which can be mapped to probabilistic predictions [34, 41]. Indirect predictions offer an opportunity to train a statistical model on both measured, objective data and subjective data.

Past work that incorporates social media data into a model often maps behaviors to a set of random variables and includes these random variables into a statistical model [2, 22, 32, 43, 52]. Most work frames these human and social media sources as passive signals that can be mined to contribute to more accurate forecasts. Compartmental models have also been proposed that take into account human behavior by estimating the contact network between individuals, reproductive and recovery rates, or by building a more complicated function between disease states that takes into account human behavior [15]. Prediction markets are another approach to aggregating human judgement that 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 [34, 41]. 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.

In this work we study how crowdsourced questions related to non-pharmaceutical interventions (NPI) in one's community can contribute to an improved forecast of COVID-19 incident cases at the national level. We posed twenty one questions related to NPI to a representative sample of the US over 36 weeks. This crowdsourced data was used to estimate the association between perceptions of adherence to NPI and incident cases at the US national level one to four weeks in advance. In addition, we fit a predictive model and show that adding crowdsourced data on perceptions of adherence, when compared to a control that does not include perceptive data, improves forecast accuracy for incident cases.

To the current literature we contribute a novel data stream of community-scale perceptive information [6] that shows (i) strong associations with 1-4 week ahead incident cases at the national level and (ii) improved predictive accuracy of 1-3 week ahead out of sample predictions when included in a computational model.

II. METHODS

II.1. Survey logistics

II.1.1. Participants and recruitment

There were a total of $10,852$ 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 a standard deviation of 119). Paid participants were initially recruited through the *SurveyMonkey* platform (40.50%, $4,405/10,852$) 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, employment), etc. Participants in this study were required to reside in the U.S. and be at least 18 years old. Survey design, distribution, and data collection was managed via SurveyMonkey software.

From February 16, 2021 to April 27, 2021, participants were recruited from the *Pollfish* survey platform (30.36% $3,295/10,852$). 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 that was 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 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 SurveyMonkey.

Compensated respondents from Survey Monkey and Pollfish accounted for approximately 70% of responses, and the final 30% of participants ($3,152/10,852$) 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 (i) more than half of questions (i.e., 11 of the 21)

were left blank or had a response of "Don't know", which occurred among 4.71% (511/10,852) of responses, or (ii) a participant gave the same response to every question, which occurred among 2.3% (331/10,852) of responses. All blank and "Don't know" responses were excluded from the analysis, which occurred among 9.68% (20,569/214,200) of total question responses (i.e., $N_{Participants} \cdot 21$).

II.1.2. 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 one week.

SurveyMonkey surveys were open to participants for compensation from the fourth 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 for participants from the 21st week of the survey period (Feb 2021) until the 36th consecutive week of data collection, the end of the survey period (April 2021).

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

II.1.3. Survey content and questions

Surveys between Sept. 2020 and April 2021 asked participants to answer the same set of 21 "core" questions (see supplemental table I for a list of core questions). Core questions asked participants about their perceptions of their community members adherence to non-pharmaceutical interventions 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%, or the answer "Don't know".

In addition to the core 21 questions, several weeks included topical questions asking participants about their perceptions of behavior during specific events (e.g., 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 [24]).

The order in which questions were presented was randomized across all 21 questions in the Pollfish surveys, and Survey Monkey questions were randomized within five categories that asked participants about individual NPI behaviors, the adherence to guidelines associated with community businesses, testing and quarantine, awareness, and restrictions or policies related to educational institutions (see supplemental figure 11 for an example of the survey).

II.1.4. Data acquisition and availability

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

Individual respondent data to all 21 questions for all 36 weeks is available at [6]. The data is 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 Institutional Review Board to publish these data on an open source platform.

II.2. Epidemiological data

Incident cases per epidemic week at the national level were collected from the Johns Hopkins University CSSE GitHub repository [12]. This repository stores cumulative cases per day from January 22, 2020 to 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 begins on Sunday and ends on Saturday.

II.3. Assessing whether the crowd was representative of the US

We assessed graphically whether our sample was representative of the US 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 θ_s is the probability of choosing at random a citizen registered in state s. We estimated θ_s, $\hat{\theta}_s$, as the census estimate for state s divided by the sum of census estimates for all states. The value e_s is $\widehat{\mathbb{E}(R_s)} = N\hat{\theta}_s$.

We included an estimated correlation coefficient between 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 (Suppl. Table II).

II.4. 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 $supp(X_{t,i,q}) = \{0, 1, 2, 3, 4\}$ corresponding to 5 different levels of adherence. The value 0 corresponds to no adherence or the adherence was 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 two participants are considered independent.

Mean **p**erceived **a**dherance (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 non-pharmaceutical intervention. 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 epiweek 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.

II.5. Estimating correlation between MEPA and incident cases

For each survey question, we estimated the correlation coefficient between MEPA at epiweek t and US national incident cases at: epiweek t, epiweek $t + 1$, $t + 2$, $t + 3$, and $t + 4$. A line list of the estimated correlation coefficient at each week ahead time point and 95% confidence intervals is available in supplemental table III.

II.6. Clustering questions

We fit a hierarchical clustering algorithm to all 21 MEPA time series for 2 through 10 clusters. Dissimilarity between two time series was computed using the euclidean distance. The Silhouette coefficient was used to assess the quality of fitting two clusters, three clusters, and so on up to ten clusters [39]. A dendrogram was plotted to visualize the clustering and MEPA time series were grouped, and plotted over epiweek.

II.7. Forecast models with and without crowdsourced perceptions

II.7.1. SIR plus Vector Auto Regression Moving Average

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

The SIR model estimates at time t the number of individuals that exist in a susceptible (S_t), infected (I_t), and recovered (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, R_0 and parameters $\beta > 0$ and $\gamma > 0$. We chose S_0 equal to the number of individuals in the US according to the most recent census. The initial value I_0 was set equal to the reported number of infections for the first epiweek that survey data was collected from 2020-08-30 to 2020-09-05, and R_0 was set to zero. The initial value problem above was integrated by Runge-Kutta45 and parameters β and γ were estimated by minimizing the least squares solution between I_t and the reported number of incident cases (Estimates of the SIR model at four different time points can be found in supplemental figure 10).

Residuals are generated as $\epsilon_t = c_t - I_t$ and we assume that these residuals together with one of the MEPA time series can be modeled as a vector auto regression moving average (VARMA) model. A VARMA assumes the residuals ϵ and MEPA time series M_q follow

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

where $Y_t = [\epsilon_t, m_{q,t}]'$, U_t is a random vector following a white noise process or $U_t \sim \mathcal{N}(0, \Sigma)$, the operator $\theta(L) = B_1 L + B_2 L^2 + \cdots$ and B_k is a matrix of coefficients, $\psi(L) = A_1 L + A_2 L^2 + \cdots$ and A_k is a matrix of coefficients, the operator L^j is the lag operator or $L^j Y_t = Y_{t-j}$, and we assume the covariance between any Y_s and Y_t is fixed and equal to Σ.

The optimal number of lags for θ and for ψ was estimated every week through each of the 36 weeks by computing the AIC for models fit with all combinations of one through three lags for θ and one through three lags for ψ. The combination that resulted in the lowest AIC was picked.

II.7.2. SIR+Random Forest+VARMA

To incorporate all MEPA time series into a model we first fit an SIR to the original time series and computed the residuals $\epsilon_t = c_t - I_t$. Next, we train a random forest regression f with $5,000$ trees where the desired output is ϵ_t as a function of ϵ_{t-1}, and all the MEPA time series values, smoothed using LOWESS, with a lag of one. The residuals $\delta_t = e_t - f(e_{t-1}, \hat{M}_{1,t-1}, \hat{M}_{2,t-1}, \cdots, \hat{M}_{21,t-1})$—where $\hat{M}_{q,t}$ is the LOWESS smoothed

MEPA time series value for question q at time t—are computed and are assumed to follow an ARIMA process, or $\theta(L)\delta_t = \psi(L)u_t$. Lags were chosen at each week based on AIC in the same manner as with the above SIR plus VARMA model.

II.7.3. Control model

Our control model follows the same SIR "detrending" of the original incident cases time series and then fits an ARIMA to the residuals. The ARIMA follows a similar approach as the VARMA model when modeling $Y_t \sim \epsilon_t$.

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

where $u_t \sim \mathcal{N}\left(0, \sigma^2\right)$. The only addition to this model is that we may first difference Y_t. A first difference computes $d_t = \nabla Y_t = Y_t - Y_{t-1}$, fits the model above, generates forecasts of $d_{t+1}, d_{t+2}, \cdots$ and then recovers $Y_{t+\ell}$ by computing $Y_{(t+\ell)-1} + d_{(t+\ell)}$.

The ARIMA process is a first attempt model in many time series applications. If models that includes MEPA variables cannot improve upon the above SIR plus ARIMA 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 [40].

II.8. Predictive scoring

Forecasts were scored using the weighted interval score (WIS) over K central quantiles [5].

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

where the interval score (IS_{α_k}) is

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

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, and m represents the median or 0.50 quantile, and c is eventually reported truth [17]. Weight w_0 equals $\frac{1}{2}$ and $w_k = \frac{\alpha_k}{2}$.

The weighted interval score (and interval score) are negatively sensed—larger values indicate worse predictive performance compared to smaller values. The best possible weighted interval score is zero and the worst possible weighted interval score is positive infinity.

III. RESULTS

III.1. Survey platform response rates

SurveyMonkey surveys received an average of 236.06 compensated responses per week ($SD = 81.14$) and an average of 88.80 volunteer responses per week ($SD = 22.68$), revealing that response rates for paid surveys was higher but more variable across weeks than volunteer survey responses. Pollfish surveys received an average of 272.55 compensated responses per week ($SD = 7.80$) and volunteer responses were not collected on the Pollfish system. (Fig. 1A).

III.2. Question response rates

The mean percent of questions that a participant answered was 87.89% ($SD = 6.15\%$) (Fig.1B). Questions 1 through five 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 response with an average response rate of 78.63%($SD = 2.09\%$).

III.3. Representative sampling

States from which most responses were collected include California (9.45%; 956/10,120), New York (8.66%; 876/10,120), Pennsylvania (6.70%; 678/10,120), Texas (6.37%; 645/10,120), and Florida (4.51%; 456/10,120).

The correlation between observed frequency of responses and expected frequency was 0.90 (95CI = $[0.84, 0.94]$; $P < 0.001$) and suggests that 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 supplemental Table II for the observed proportion, expected proportion, and relative difference between the two).

Seven states deviated from expected response rates by more than 9 standard deviations. Four states were under represented: Mississippi, Puerto Rico, Florida, and Texas, and three states were over represented: Minnesota, Pennsylvania, and New York (Fig. 2). Pennsylvania (PA) was the most over represented state.

When both compensated and volunteer responses are included, the response frequency in PA is 10 standard deviations above the expected and when volunteer responses are removed the response frequency decreases to 3.5 standard deviations below expectation.

III.4. MEPA over time

Mean perception of adherence (MEPA) increased the most from the beginning of the survey to the end for the following three questions: question 21 ($\Delta M_{21} = M_{21,36} - M_{21,1} = 1.29$) that asked participants about knowledge of their state policies and whether "Violations of COVID restrictions result in fines or police enforcement", question 11 ($\Delta M_{11} = 1.24$) that asked how frequently community members follow recommendations to seek "[...] antibody testing to detect prior infection", and for question 14 ($\Delta M_{14} = 0.57$)

that asked participants how frequently members of their community quarantine after traveling (Fig. 3A.).

MEPA decreased the most from the start to the end of the survey for question 7 (ΔM_7 = -1.07) that asked participants how frequently restrictions are placed on visiting senior living facilities, for question 4 (ΔM_4 = -0.83) that asked how frequently restaurants have reduced seating capacity, and question 9 (ΔM_9 = -0.81) that asked about the frequency of special protection in hospitals when treating patients with COVID-19.

The standard deviation between MEPA values at the beginning of the survey period ($SD_{Beginning} = 0.89$) was larger than the standard deviation between MEPA values at the end of the survey period (SD_{End} = 0.33) (Fig. 3A.). The mean MEPA value take over all twenty one questions remained similar over the course of the survey ($M_{Beginning} = 3.15, M_{End} = 3.14$).

the estimated correlation between MEPA values at time t and $t - \ell$ was greater than 0.35 for lags of up to 4 weeks ($\ell = 3$) for a majority of MEPA time series (Fig. 3B.) and suggests many MEPA time series contain more structure than a random walk. Responses to five survey questions had a mean absolute autocorrelation greater than 0.2: Question 3 ([...] staying at home), Question 4 ([...] restaurants complying with CDC recommendations to have reduced seating), Question 9 ([...] special protection in hospitals areas that treat COVID patients), Question 10 ([...] get tested for active virus), and Question 11 ([...] get antibody testing to detect prior infection). Mean absolute autocorrelations for these five questions across 34 lagged weeks were above 0.2. A more detailed view of auto correlation for a lag of one week is available in supplemental figure 9.

III.5. Clustering questions according to similarities in responses over time

MEPA time series were grouped into four clusters (Fig. 4A. and B.), one cluster of MEPA questions with values between 2.5 and 3.5 (i.e. low to medium adherence; Fig. 4C.), a second cluster of questions that decreased over time (Fig. 4D.), a third cluster with values near 2.25 at the beginning of the survey and that increased over time (Fig. 4E.), and a fourth cluster with values near 1.25 at the start of the survey and that increased over time, ending above 2.50 by the end of the survey (Fig. 4F.).

Cluster quality as measured by the silhouette coefficient (Sil.) was highest when grouping MEPA time series into four clusters, however the Sil. for four clusters was similar to the Sil. for two and three clusters (Fig. 4A.). In the cluster in Fig. 4C there may exist two clusters—one with increasing adherence over time and a second cluster with decreasing adherence over time.

MEPA time series within the same cluster asked participants about similar adherence behaviors. Questions corresponding to avoidance behavior (questions 2, 12, and 15) were more similar to one another than 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 (question 2 and 13), and awareness at the state level (questions 16 and 17).

For autocorrelations between one through four week ahead MEPA responses across the different clusters, see supplement Figure 12.

III.6. Correlation between perceptions of adherence and reported incident cases

The estimated correlation between the MEPA time series representing responses to the question "What percent of people in your community do you notice are usually wearing a mask in public?" and one week ahead incident cases was $\rho = -0.43$; 95CI = $[-0.67, -0.10]$. For two week ahead incident cases the correlation was $\rho = -0.44$ $[-0.68, -0.12]$, for three weeks ahead was $\rho = -0.42$ $[-0.67, -0.08]$, and for four weeks ahead was $\rho = -0.37$ $[-0.64, -0.02]$ (Fig. 5 row two). The MEPA time series for the question "In your state, what percent of colleges are closed or holding only remote classes?" has an estimated correlation of $\rho = 0.62$ $[0.35, 0.79]$ with one week ahead cases and $\rho = 0.52$ $[0.21, 0.73], 0.41$ $[0.08, 0.67], 0.28$ $[-0.09, 0.57]$ for two through four week ahead reported incident cases at the US national level (Fig. 5 row 19). Correlation coefficients and 95 percent confidence intervals for each question are available in supplement table III.

III.7. Out-of-sample improvement in forecasting with crowdsourced MEPA

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

The proportion of times a forecast that includes a single MEPA time series generates a smaller (improved) WIS score compared to a model that does not use MEPA, is above 50% for the majority of adherence questions for forecast horizons of one through three weeks ahead (Fig. 7). MEPA most improves two week ahead forecasts. 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" improve 76%; 95CI [58%, 94%] of two week ahead forecasts. For three weeks ahead, the question "What percent of people usually stay home" improved 76%; 95CI = [58%, 94%] of forecasts and the machine learning model that incorporated all adherence questions improved 76% 95CI=[58%, 94%] of forecasts. Including MEPA data improves four week ahead forecasts minimally and for only a small set of questions.

Including MEPA data improves, compared to the control model, one through four week ahead forecast accuracy (i.e. reduces WIS) at and after the peak reported number of incident cases (Fig. 8). One week ahead forecasts show consistent, small gains in forecast accuracy over time (Fig. 8A.). Two and three week ahead forecasts show large gains in forecast accuracy at and just after the peak number of incident cases time (Fig. 8B. and C.), and four week ahead improvements in forecast accuracy appear near the peak number of cases (Fig. 8D.).

IV. DISCUSSION

We found that crowdsourced perceptions of adherence to non-pharmaceutical intervention correlated with 1-4 week ahead incident cases at the US national level, and that including perceptual data into a computational model improved one through three week ahead forecast accuracy.

This work supports the hypothesis that a crowd may be able to assign realistic probabilities to outcomes about community adherence to NPI, 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 [18, 31]. Past literature about our ability to make accurate probabilistic predictions is mixed. Some past work suggests a crowd may not be able to map environmental cues to accurate probabilities of outcomes [33], while other work has shown people's statistical intuitions may overlap with the statistics of their environment [18].

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 more available in their environment, such as the proportion of people wearing masks, and responded less often to questions that were less available in the 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. Last, strong correlations between weekly responses to specific NPI questions indicates that participants' judgment in the present study was consistent (Fig. 9). 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 [48]. Seemingly irrelevant information can have large impacts on judgment. For example, when asked to complete an irrelevant task such as writing down the last two digits of their social security numbers 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 [1]. Such findings underscore the importance of carefully crafting judgment questions to avoid activating judgment biases. Human judgment data must also be inspected for quality—participants in the present 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 survey items closely. Last, 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 epi- or pandemic. Participation rates in the present study tended to increase throughout the data collection period, which created difficulties 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 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 [1]. Finally, variable sampling rates across states created difficulties 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.

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, respondents could be asked to judge compliance specifically at the level of their county, then these judgments could be input into 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.

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

V. ACKNOWLEDGEMENTS

We thank 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 the support of the data collection. We also wish to thank all participants who offered their perceptions over the survey period.

VI. TABLES

Question
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. Hairdresser and barber to be open with restrictions
7. Visitors to senior living facilities to be restricted
8. Commonly touched surfaces to be sanitized
9. Special protection in hospitals 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

TABLE I: A listing of 21 "core" questions that were presented to participants on every survey from September 6, 2020 to April 30, 2021. Questions asked for participant's perceptions of community and local businesses' adherence to non-pharmaceutical interventions and awareness of public health guidelines.

VII. FIGURES

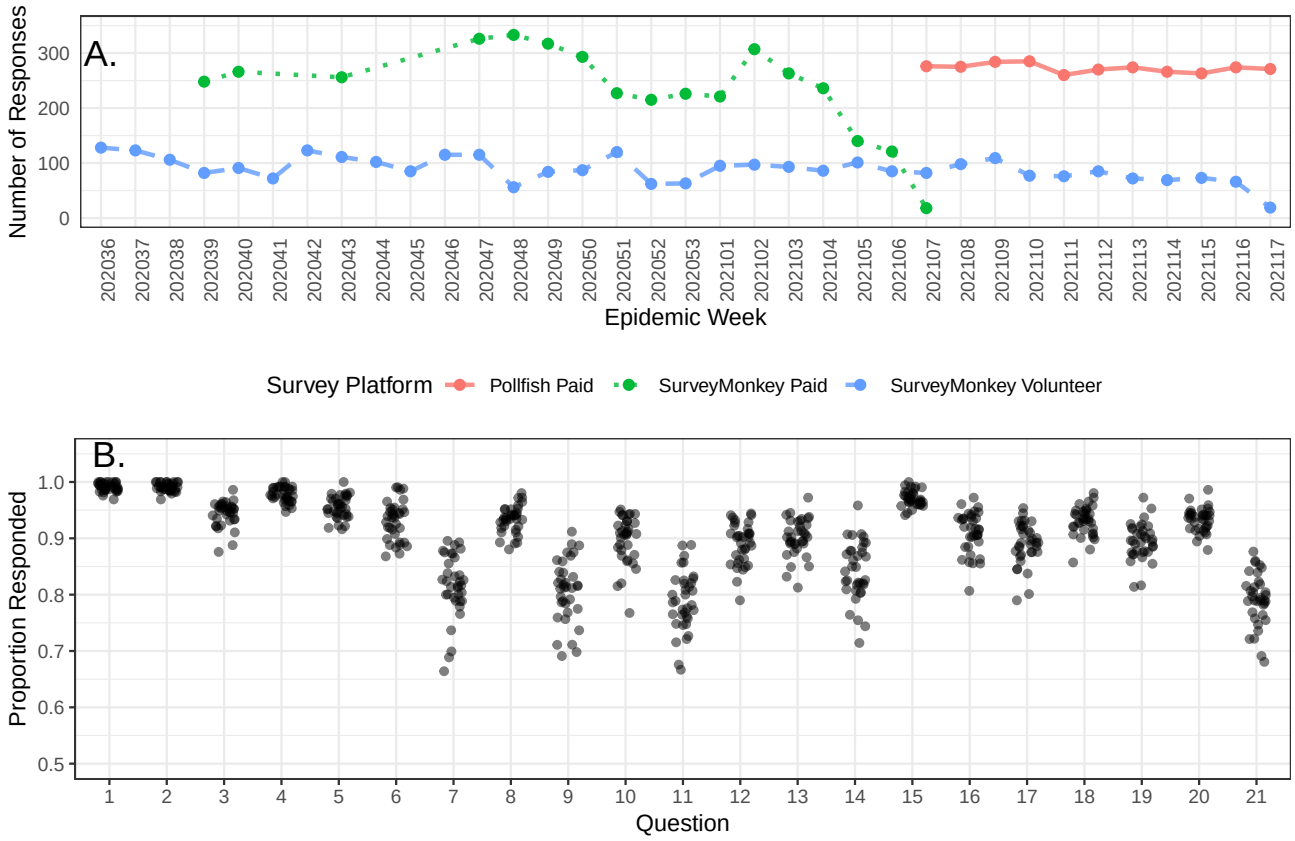


FIG. 1: (A.) The number of participant responses per epidemic week for the Pollfish platform (red) and for those who submitted responses on Survey Monkey who were compensated (green) and who were volunteer (blue). (B.) The proportion of participants who responded to each question in a given epidemic week. Volunteers made consistent contributions each week as did the Pollfish compensated participants, 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 NPI 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).

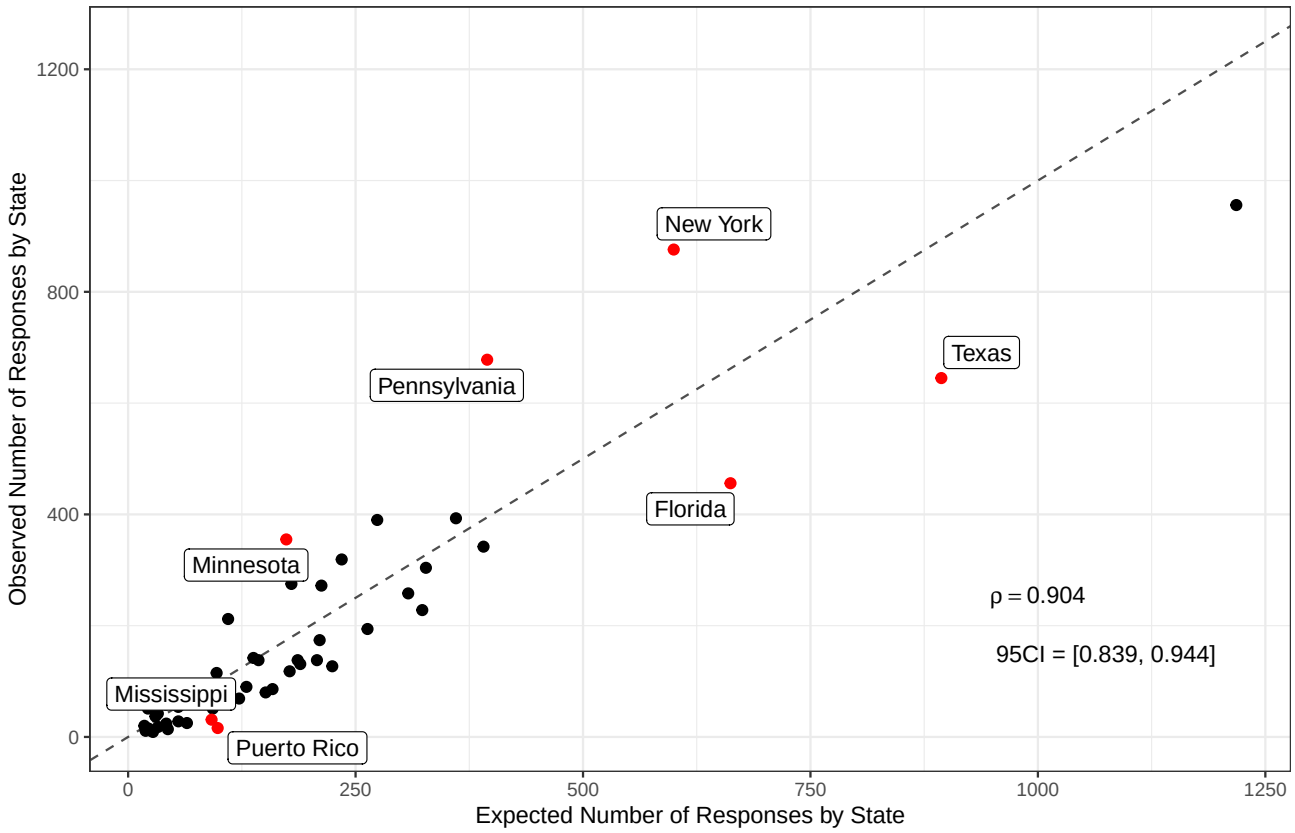


FIG. 2: The number of observed responses to the surveys summed over the survey period (vertical axis) compared to the expected number of total responses according to the census (horizontal axis). A dashed line indicates if the observed and expected number of responses equal one another. A select number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

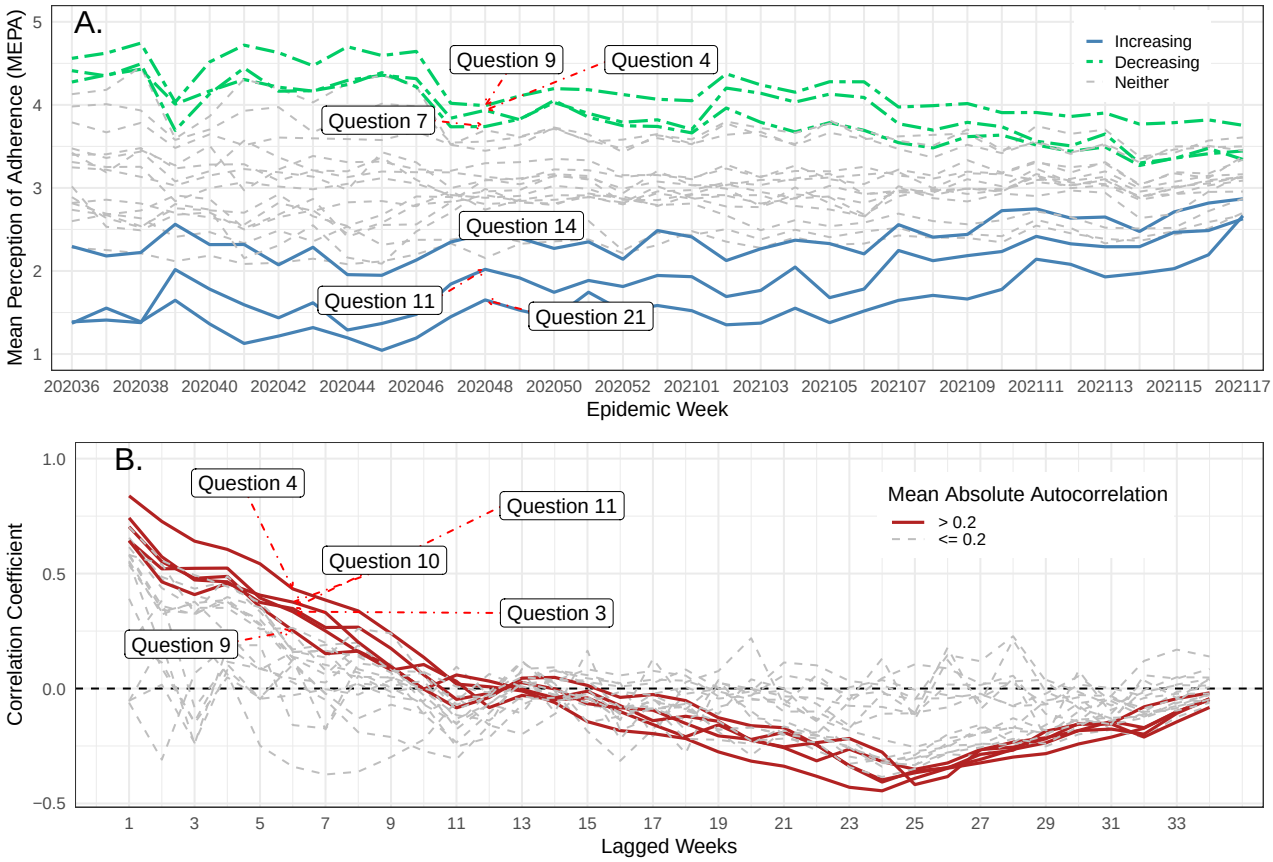


FIG. 3: (A.) Mean perception of adherence—MEPA—for twenty one questions asked over the survey period. (B.) Auto correlation for all twenty one MEPA time series for a lag of one to 18 weeks. Perceptions of adherence for questions that asked about state policies (question 21) and antibody testing practices (question 11) increased over the survey period while perception of adherence decreased for questions that asked about restrictions placed on senior living facilities (question 7) and on restaurants (question 4). Mean absolute autocorrelations for these 5 questions across 34 lagged weeks were above 0.2. The estimated correlation between MEPA values at time t and $t - \ell$ was greater than 0.35 for lags of up to 4 weeks ($\ell = 4$) for a majority of MEPA time series. MEPA time series appear to contain more structure then a random walk, suggesting crowdsourced perceptions may be a useful signal for predicting incident cases.

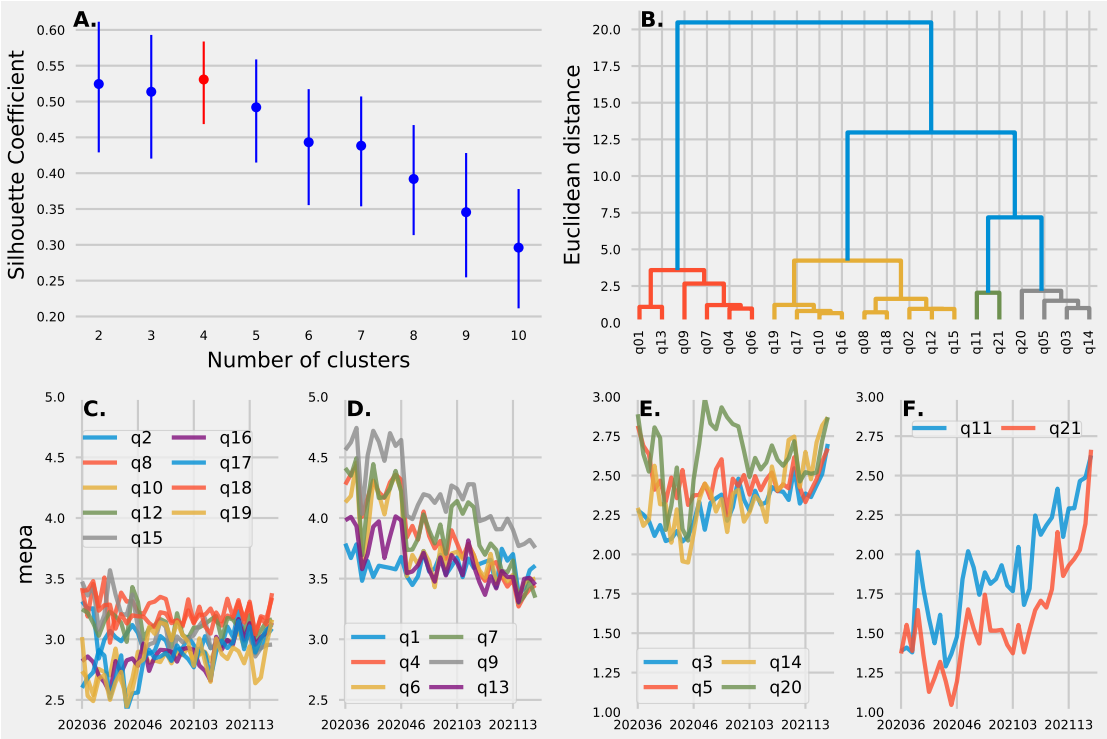


FIG. 4: Hierarchical clustering of the twenty one MEPA time series using euclidean distance as a measure of dissimilarity between two time series. (A.) Silhouette coefficients for two through ten clusters of MEPA time series. (B.) Dendrogram that reports questions on the horizontal axis and dissimilarity between individual questions or clusters on the vertical axis. (C.-F.) MEPA time series clustered into four 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 non-pharmaceutical intervention.

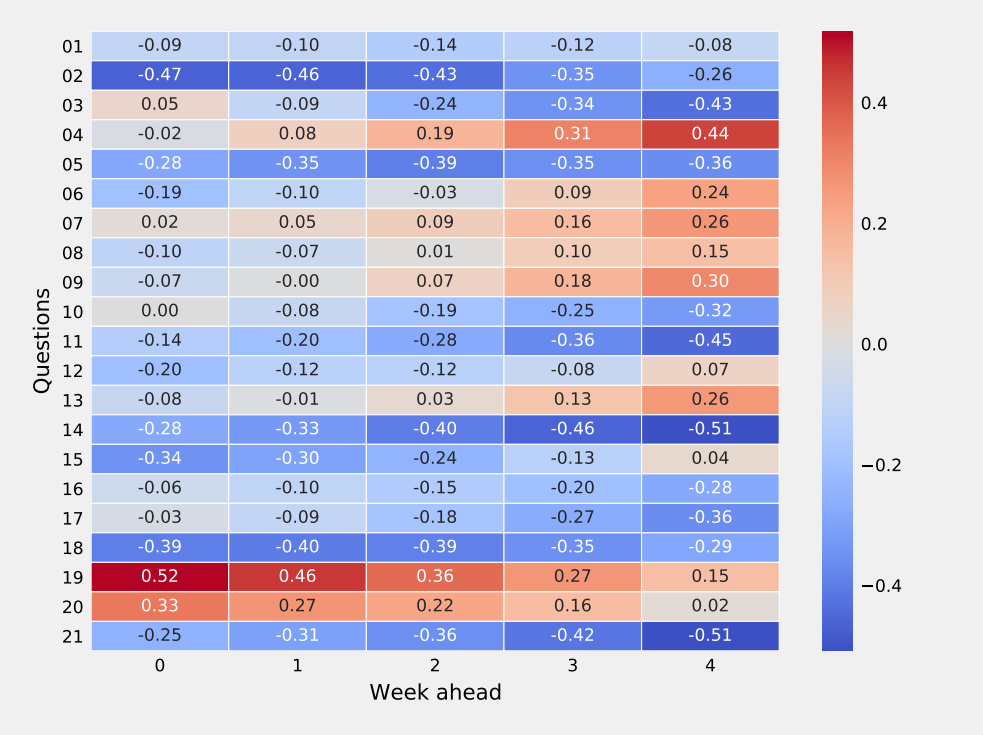


FIG. 5: Linear correlation between twenty one MEPA time series associated with questions about perception of adherence and one through four week ahead incident cases at the US national level. The correlation between question two which asked "What percent of people in your community do you notice are usually wearing a mask in public?" and 1-4 week ahead incident cases is -0.26 or lower, and the correlation between question nineteen which asked "In your state, what percent of colleges are closed or holding only remote classes." and cases is 0.27 or higher for one through three week ahead cases. Select crowdsourced perceptions of adherence to non-pharmaceutical interventions correlate with short range and long range reported incident cases at the national level.

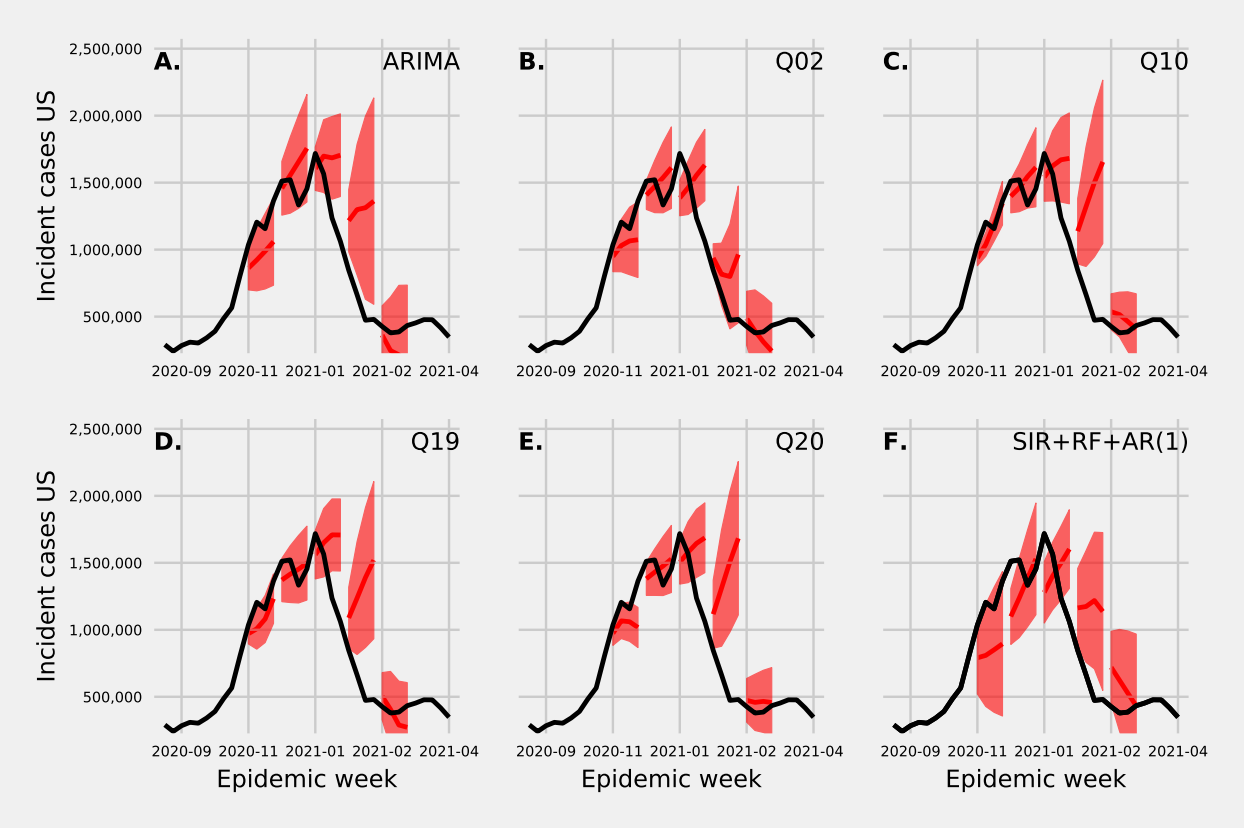


FIG. 6: Forecasts of US national incident cases 1-4 week ahead at six time points throughout the survey period by first fitting an SIR model and then modeling the residuals by (A.) fitting an auto regressive model with one lag, (B.)-(E.) fitting a vector auto regression moving average that includes the residual time series and MEPA values for select questions, and (F.) fitting a random forecast to residuals that includes MEPA values for all questions asked of participants plus an AR(1) model.

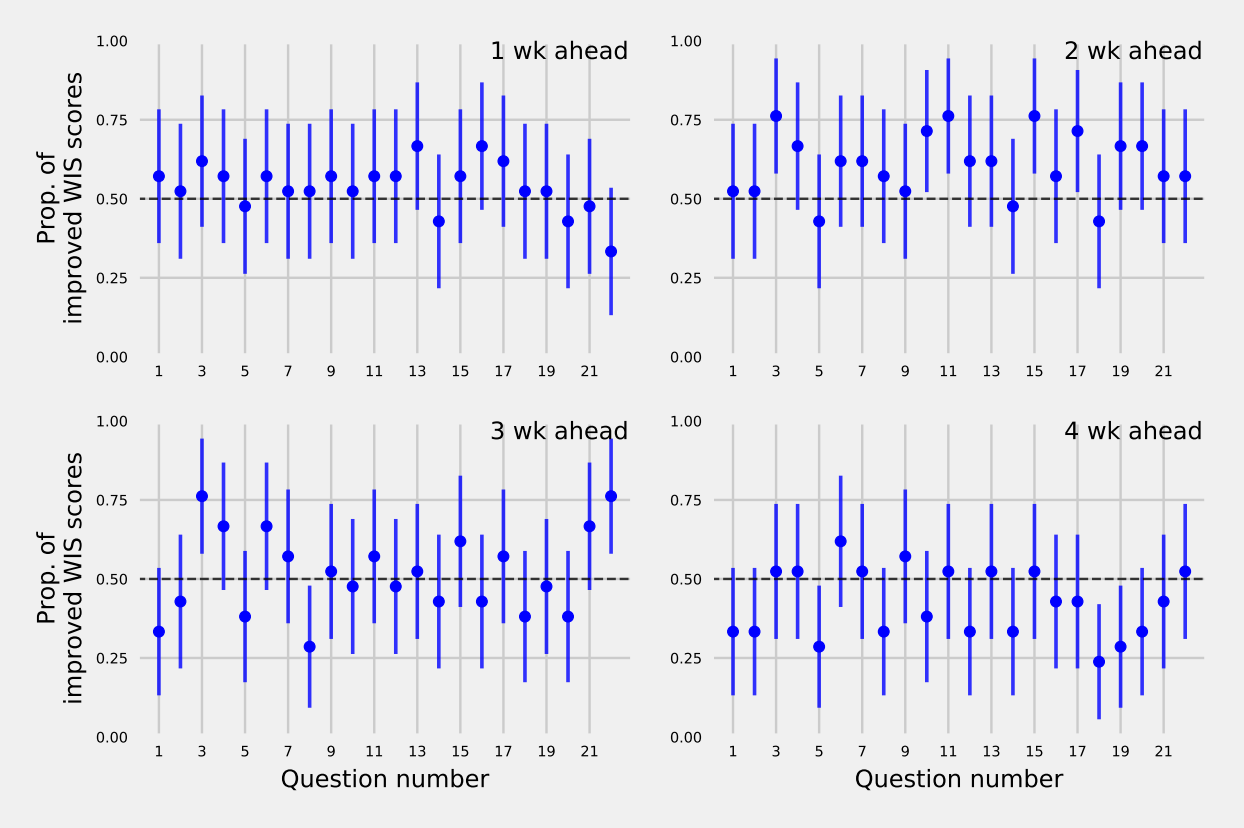


FIG. 7: The proportion and 95CI of WIS scores that were improved (smaller) for a SIR plus VARMA model that included MEPA time series one through twenty one compared to the the control SIR model without using a MEPA time series for forecasts one through four weeks ahead. An additional model, to the right of model number 21, is an SIR model plus a random forecast that includes all twenty one MEPA time series and an ARIMA to model residuals. The majority of MEPA time series improve one and two week ahead forecasts of incident cases. A smaller number of MEPA time series improve three week ahead forecasts and four week ahead forecasts are improved only modestly.

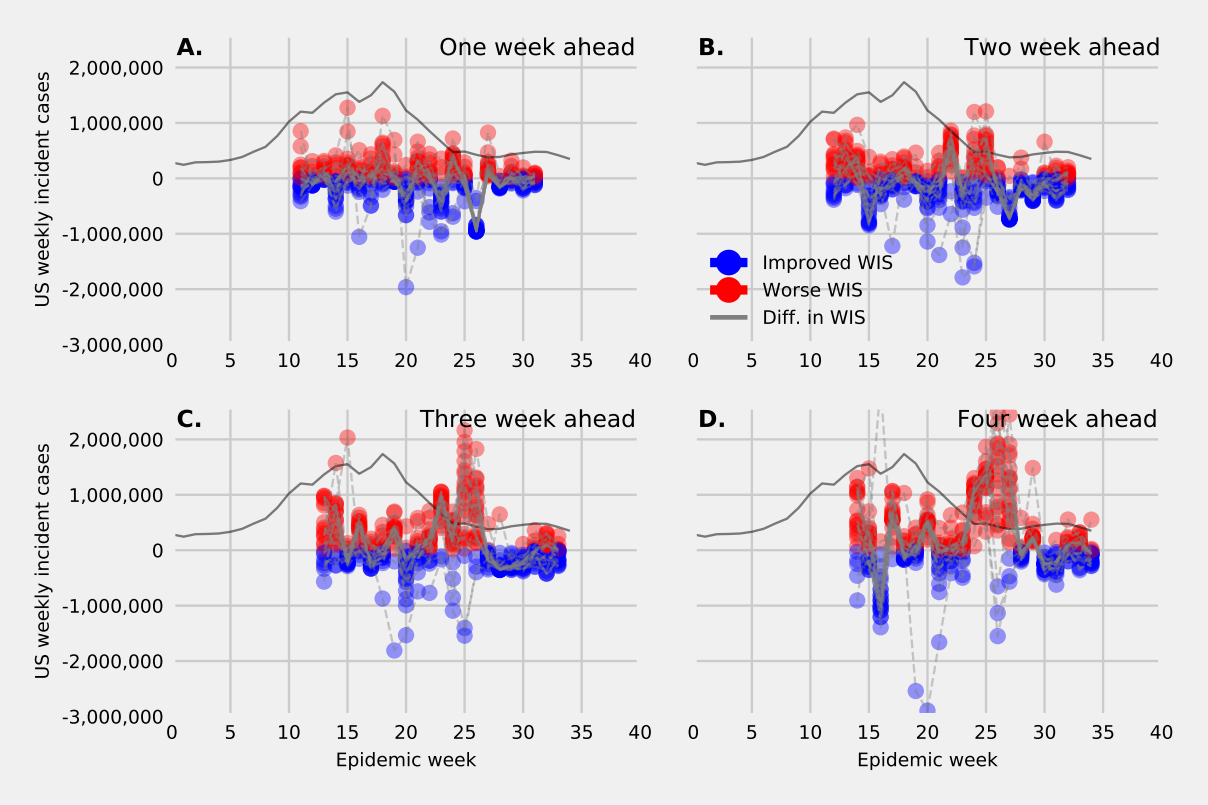


FIG. 8: The difference in WIS scores for (A.) one week ahead, (B.) two week, (C.) three week, and (D.) four week ahead forecasts of US national incident cases between models that include one MEPA time series and the control model that uses only past incident case data to produce a forecast. The differences in WIS correspond to the forecasted epidemic weeks, not when the forecast was generated. The reported number of incident cases at the US national level is in (grey). A point represents the difference in WIS at this 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 NPI improves predictions at and after peak incident cases.

Appendix A: Expected and Observed state sampling of participants

State	Observed Proportion	Expected Proportion	Relative Difference
Alabama	.0084	.0149	-0.44
Alaska	.0016	.0022	-0.29
Arizona	.0134	.0222	-0.40
Arkansas	.0054	.0092	-0.42
California	.1007	.1204	-0.16
Colorado	.0124	.0175	-0.29
Connecticut	.0223	.0109	1.06
Delaware	.0039	.0030	0.31
District Of Columbia	.0054	.0022	1.50
Florida	.0481	.0654	-0.27
Georgia	.0320	.0323	-0.01
Hawaii	.0015	.0043	-0.66
Idaho	.0057	.0054	0.05
Illinois	.0360	.0386	-0.07
Indiana	.0145	.0205	-0.29
Iowa	.0121	.0096	0.26
Kansas	.0072	.0089	-0.19
Kentucky	.0150	.0136	0.10
Louisiana	.0145	.0142	0.03
Maine	.0061	.0041	0.49
Maryland	.0145	.0184	-0.21
Massachusetts	.0287	.0210	0.36
Michigan	.0272	.0304	-0.11
Minnesota	.0374	.0172	1.18
Mississippi	.0033	.0091	-0.64
Missouri	.0138	.0187	-0.26
Montana	.0019	.0033	-0.42
Nebraska	.0079	.0059	0.34
Nevada	.0068	.0094	-0.27
New Hampshire	.0025	.0041	-0.39
New Jersey	.0411	.0271	0.52
New Mexico	.0026	.0064	-0.59
New York	.0923	.0593	0.56
North Carolina	.0240	.0320	-0.25
North Dakota	.0013	.0023	-0.46
Ohio	.0414	.0356	0.16
Oklahoma	.0073	.0121	-0.40
Oregon	.0095	.0128	-0.26
Pennsylvania	.0714	.0390	0.83
Puerto Rico	.0017	.0097	-0.83
Rhode Island	.0044	.0032	0.37
South Carolina	.0091	.0157	-0.42
South Dakota	.0009	.0027	-0.65
Tennessee	.0183	.0208	-0.12
Texas	.0680	.0883	-0.23
Utah	.0071	.0098	-0.28
Vermont	.0012	.0019	-0.39
Virginia	.0204	.0260	-0.21
Washington	.0336	.0232	0.45
West Virginia	.0030	.0055	-0.46
Wisconsin	.0290	.0177	0.63
Wyoming	.0021	.0018	0.20

TABLE II: The observed and expected proportion according to the US national census, and relative difference between observed and expected proportions over the 36 week survey period stratified by state.

Appendix B: Autocorrelation of one week for MEPA time series

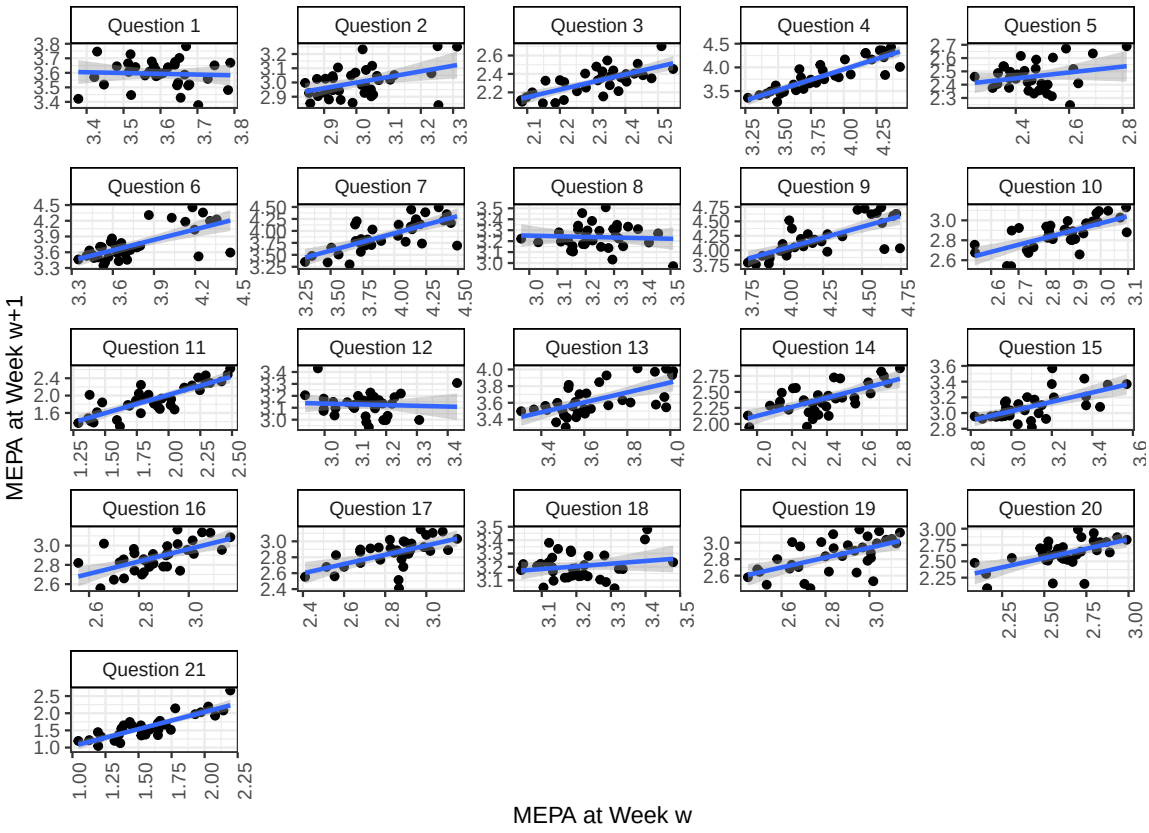


FIG. 9: Auto correlation between MEPA at week $w+1$ and MEPA at week w for all twenty one questions over the survey period. The correlation is high for the majority of MEPA time series, suggesting MEPA could be modeled by an auto regressive process and contains more structure than a random walk.

Appendix C: Correlation between MEPA and US national incident cases

Question	$\hat{\rho}$	Lower CI	Upper CI
Week 0			
1	-0.094	-0.414	0.247
2	-0.467	-0.692	-0.158
3	0.051	-0.287	0.378
4	-0.018	-0.349	0.317
5	-0.285	-0.565	0.053
6	-0.194	-0.495	0.149
7	0.021	-0.315	0.352
8	-0.104	-0.423	0.237
9	-0.068	-0.392	0.272
10	0.002	-0.331	0.335
11	-0.136	-0.449	0.207
12	-0.203	-0.502	0.140
13	-0.080	-0.403	0.260
14	-0.280	-0.561	0.059
15	-0.342	-0.606	-0.009
16	-0.059	-0.385	0.280
17	-0.029	-0.359	0.307
18	-0.394	-0.642	-0.069
19	0.518	0.223	0.726
20	0.335	0.002	0.601
21	-0.249	-0.538	0.092
Week 1			
1	-0.099	-0.418	0.242
2	-0.459	-0.687	-0.148
3	-0.089	-0.410	0.252
4	0.084	-0.257	0.406
5	-0.348	-0.610	-0.017
6	-0.100	-0.419	0.241
7	0.046	-0.292	0.373
8	-0.071	-0.395	0.269
9	-0.003	-0.336	0.330
10	-0.082	-0.404	0.258
11	-0.201	-0.501	0.142
12	-0.125	-0.440	0.218
13	-0.011	-0.343	0.324
14	-0.328	-0.596	0.006
15	-0.296	-0.573	0.041
16	-0.101	-0.420	0.241
17	-0.094	-0.415	0.247
18	-0.403	-0.649	-0.081
19	0.458	0.147	0.686
20	0.273	-0.066	0.556
21	-0.305	-0.579	0.031

Week 2			
1	-0.144	-0.455	0.199
2	-0.433	-0.670	-0.117
3	-0.237	-0.528	0.105
4	0.185	-0.158	0.488
5	-0.393	-0.642	-0.069
6	-0.025	-0.355	0.311
7	0.089	-0.252	0.410
8	0.005	-0.329	0.338
9	0.071	-0.268	0.395
10	-0.185	-0.488	0.158
11	-0.275	-0.557	0.064
12	-0.124	-0.439	0.219
13	0.033	-0.303	0.363
14	-0.400	-0.647	-0.077
15	-0.245	-0.534	0.096
16	-0.148	-0.459	0.195
17	-0.182	-0.486	0.161
18	-0.390	-0.640	-0.065
19	0.364	0.035	0.622
20	0.225	-0.117	0.519
21	-0.356	-0.616	-0.026
Week 3			
1	-0.116	-0.433	0.226
2	-0.347	-0.610	-0.016
3	-0.341	-0.605	-0.008
4	0.310	-0.026	0.583
5	-0.355	-0.615	-0.024
6	0.091	-0.250	0.412
7	0.160	-0.183	0.468
8	0.096	-0.245	0.416
9	0.181	-0.162	0.485
10	-0.252	-0.540	0.089
11	-0.365	-0.622	-0.036
12	-0.081	-0.403	0.259
13	0.128	-0.214	0.442
14	-0.460	-0.688	-0.150
15	-0.127	-0.442	0.215
16	-0.198	-0.498	0.145
17	-0.270	-0.553	0.070
18	-0.352	-0.613	-0.021
19	0.268	-0.072	0.552
20	0.157	-0.186	0.466
21	-0.418	-0.659	-0.098
Week 4			
1	-0.085	-0.407	0.256
2	-0.259	-0.545	0.081
3	-0.434	-0.670	-0.118
4	0.440	0.125	0.674

5	-0.357	-0.617	-0.027
6	0.237	-0.105	0.528
7	0.264	-0.075	0.549
8	0.148	-0.194	0.459
9	0.300	-0.037	0.576
10	-0.320	-0.591	0.014
11	-0.447	-0.679	-0.134
12	0.068	-0.272	0.392
13	0.258	-0.083	0.544
14	-0.509	-0.720	-0.211
15	0.040	-0.297	0.368
16	-0.279	-0.560	0.060
17	-0.358	-0.618	-0.028
18	-0.289	-0.568	0.049
19	0.151	-0.192	0.461
20	0.023	-0.313	0.354
21	-0.507	-0.719	-0.209

TABLE III: Estimated correlation coefficients ($\hat{\rho}$) and 95% confidence intervals between MEPA time series and one through four week ahead US national incident cases.

Appendix D: SIR model fit to US national incident cases

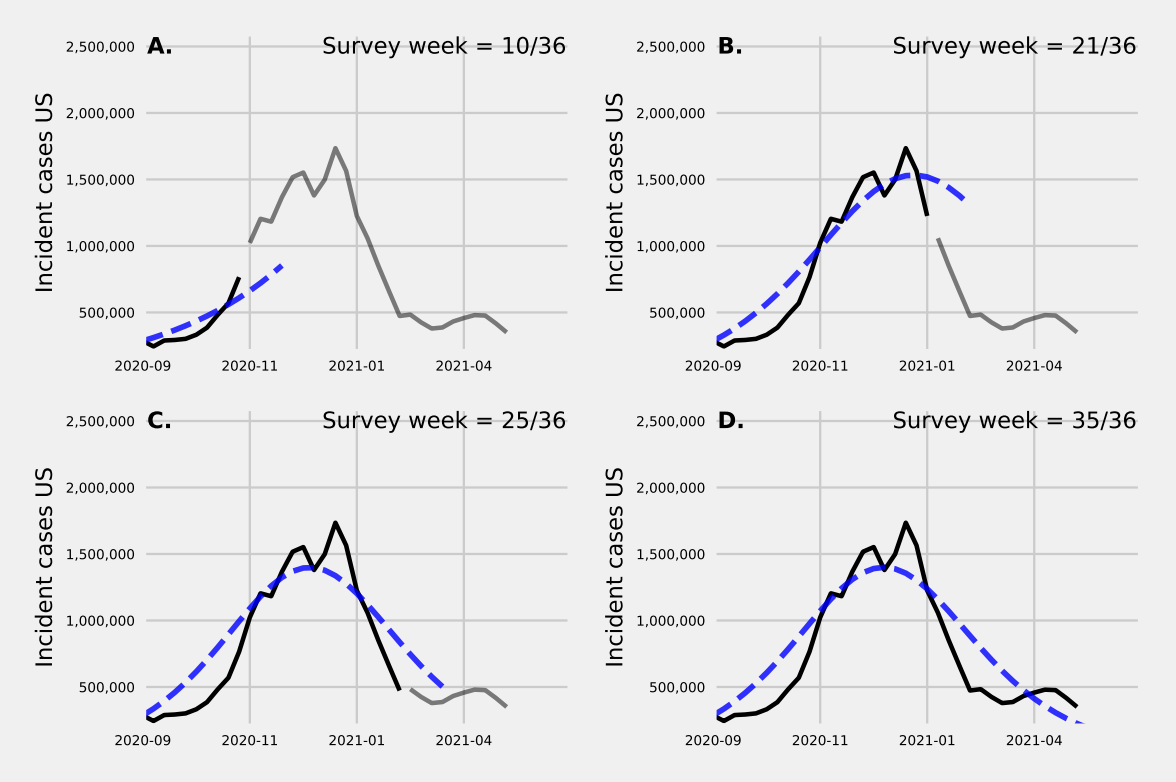


FIG. 10: The number of US national incident cases that are observed during a specific week (black), future unobserved incident cases (gray), and a simple SIR model (blue) fitted at survey week (A.) 10 which corresponds to epidemic 202036, survey week (B.) 21, (C.) 25, and (D.) 35 out of the 36 weeks that this survey was conducted.

Appendix E: Survey Template

CMMP

Copy of Responses to COVID-19 In Your Community Survey

This weekly survey collects your opinions on how your community is responding to COVID-19. Whether you are out and about every day or haven't left home except for groceries since March, *your insight is valuable*.

We are collecting responses from many people across the U.S. and Canada. The *"Wisdom of the Crowd"* works by adding up people's knowledge to get a complete picture. So everyone's observation helps, even if you don't think you are an expert in how your community is responding to COVID-19.

Thanks for your time. **Please enter your email address at the end to receive the weekly report and get signed up to take next week's survey. To see how we use these survey results, go to our website.**

Note: We ask you to rate practices as 0%, 20%, 40%, 60%, 80%, 100%. We would want you to think of it as 0% to 10% gets a 0%, 10% to 30% gets a 20%, 30% to 49% gets a 40%, 51% to 70% gets a 60%, 70% to 90% gets an 80%, and 90% and over gets a 100% response. We left out 50%. If your answer is 50%, please think about it and decide if you can make that a 49% or a 51%.

1. **What percent of people in your community do you notice are usually:**
(Note: Please adjust your answer for partial compliance. So, for example, if someone covers just one of nose and mouth with mask, treat that as 50% compliance for people who do that.)

	None/Not Adopted	Few / 20%	Some / 40%	Many / 60%	Most / 80%	All / 100%	Don't know
Wearing a mask in public	○	○	○	○	○	○	○
Maintaining social distance	○	○	○	○	○	○	○
Staying at home	○	○	○	○	○	○	○

2. **How common is it in your community for:**

	None/Not Adopted	Few / 20%	Some / 40%	Many / 60%	Most / 80%	All / 100%	Don't know
Restaurants complying with CDC recommendations to have reduced seating	○	○	○	○	○	○	○
Businesses to be closed – work from home only	○	○	○	○	○	○	○
Hairdresser and barber complying with CDC restrictions	○	○	○	○	○	○	○
Visitors to senior living facilities to be restricted	○	○	○	○	○	○	○
Commonly touched surfaces to be sanitized	○	○	○	○	○	○	○
Special protection in hospitals areas that treat COVID patients	○	○	○	○	○	○	○

3. **In your community, how common is it for people to follow recommendations or requirements to:**

	None/Not Adopted	Few / 20%	Some / 40%	Many / 60%	Most / 80%	All / 100%	Don't know
Get tested for active virus	○	○	○	○	○	○	○
Get antibody testing to detect prior infection	○	○	○	○	○	○	○
Quarantine people who have been in close contact with people with positive tests	○	○	○	○	○	○	○
Quarantine people with positive tests	○	○	○	○	○	○	○
Quarantine travelers from higher infection places	○	○	○	○	○	○	○
Limit large gatherings of people	○	○	○	○	○	○	○

4. **How many people in your community are aware of:**

	None/Not Adopted	Few / 20%	Some / 40%	Many / 60%	Most / 80%	All / 100%	Don't know
Local level of COVID infections	○	○	○	○	○	○	○
Statewide targets for reducing COVID spread	○	○	○	○	○	○	○
Local approach to limiting COVID spread	○	○	○	○	○	○	○

5. **In your state what percent of:**

	None/Not Adopted	Few / 20%	Some / 40%	Many / 60%	Most / 80%	All / 100%	Don't know
Colleges are closed or holding only remote classes	○	○	○	○	○	○	○
Schools (K-12) are closed or holding only remote classes	○	○	○	○	○	○	○
Violations of COVID restrictions result in fines or police enforcement	○	○	○	○	○	○	○

6. How soon would you expect do long distance travel by air?

- ○ By the end of April
- ○ May
- ○ June
- ○ Third Quarter
- ○ Fourth Quarter
- ○ 2022
- ○ I don't expect to travel by air
- ○ Don't know

7. What will make you comfortable with traveling by air? (Please check all that apply)

- ☐ Getting vaccinated myself
- ☐ Lots of people getting vaccinated at my destination
- ☐ Reaching Herd Immunity
- ☐ COVID mostly eliminated
- ☐ CDC says it is safe
- ☐ Governor says it is safe
- ☐ Airline says it is safe
- ☐ Other (please specify)

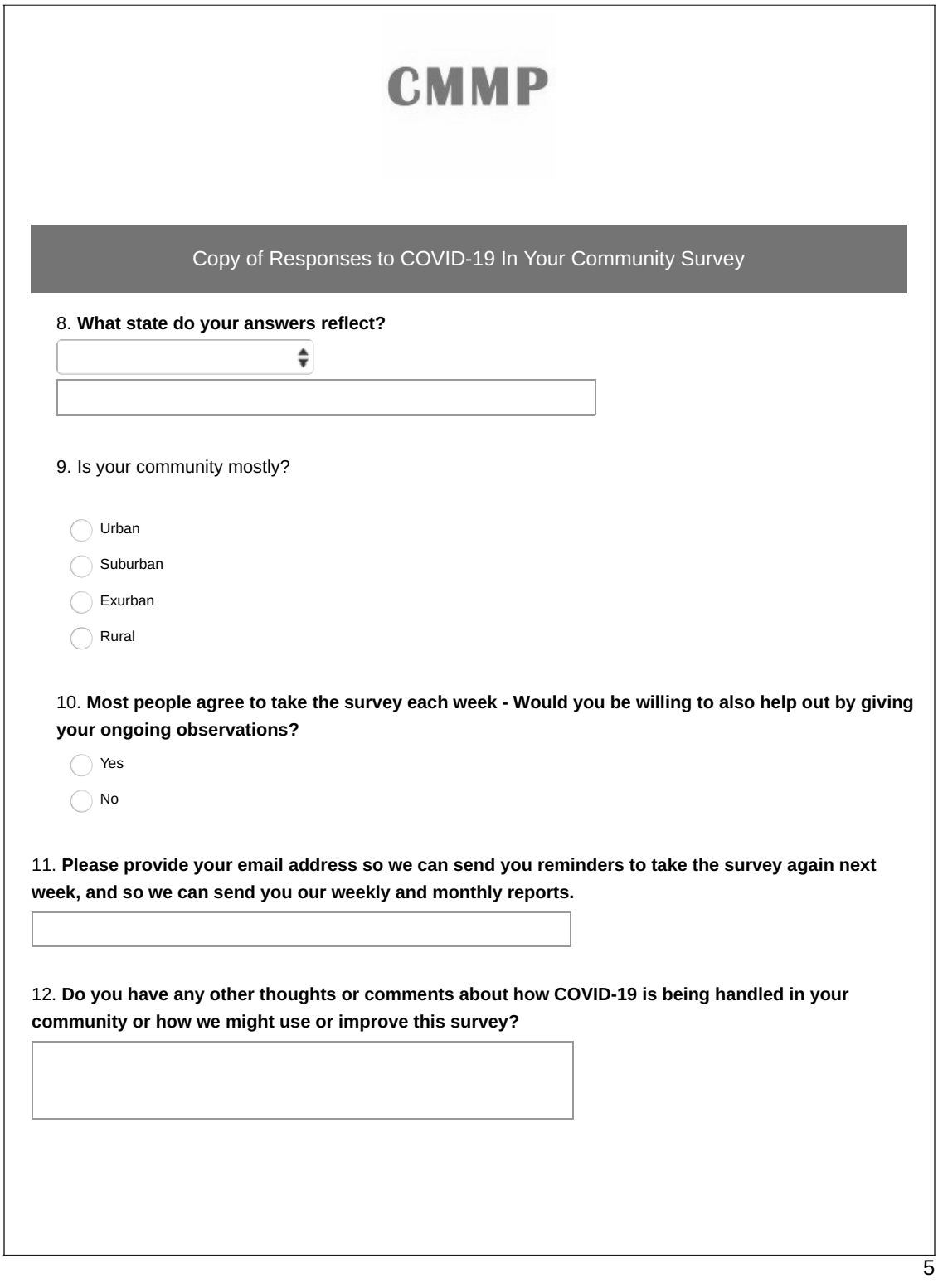


FIG. 11: Survey provided to participants to capture perceptions of adherence to non-pharmaceutical adherence.

Appendix F: Bivariate relationships between question clusters and incident cases

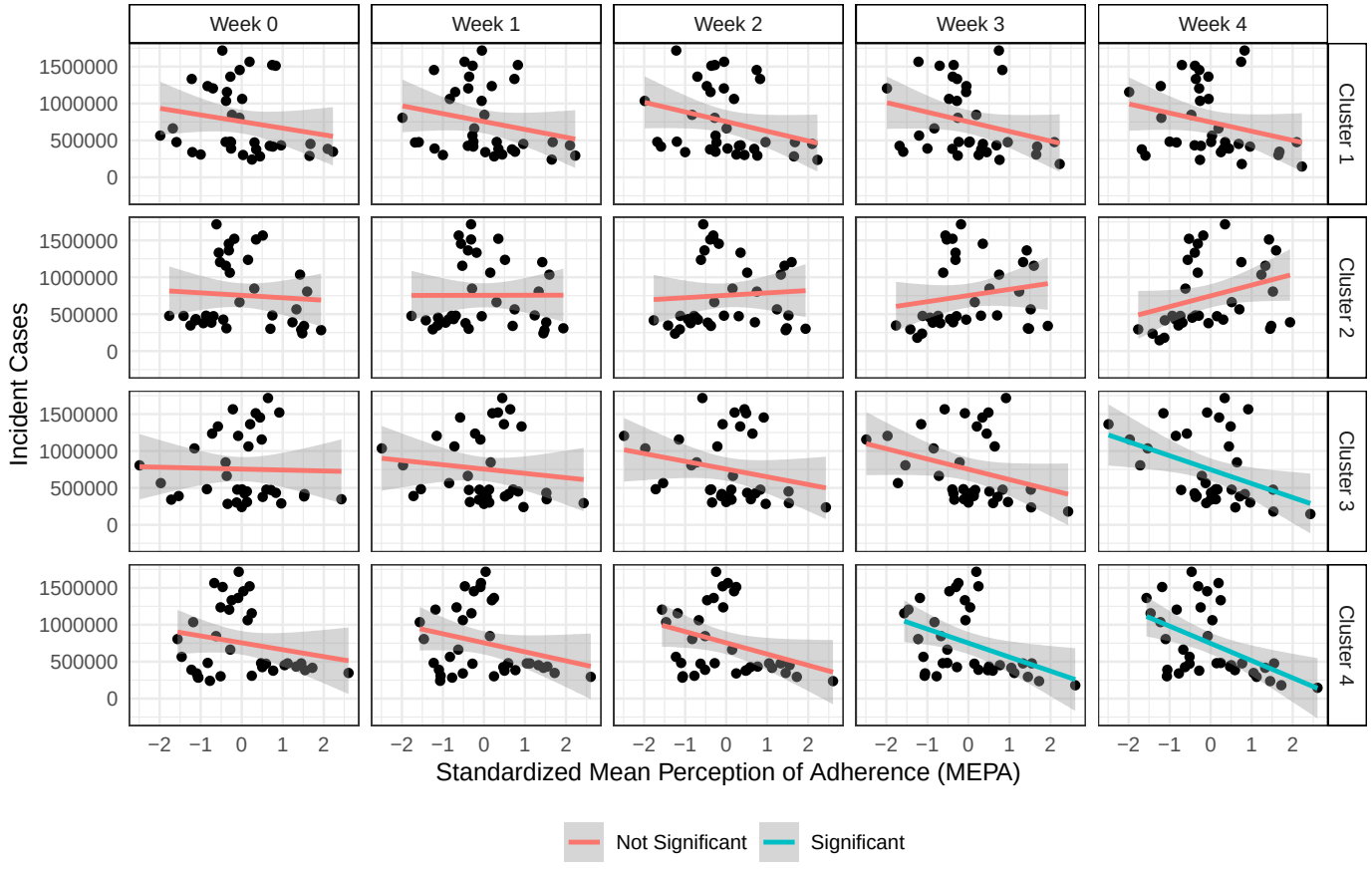


FIG. 12: Bivariate relationships between four clusters of MEPA time series and one through four week ahead incident cases at the US national level.

[1] Dan Ariely, George Loewenstein, and Drazen Prelec. "coherent arbitrariness": Stable demand curves without stable preferences. *The Quarterly journal of economics*, 118(1):73–106, 2003.

[2] Sujin Bae, Eunyoung Sung, and Ohbyung Kwon. Accounting for social media effects to improve the accuracy of infection models: combatting the covid-19 pandemic and infodemic. *European Journal of Information Systems*, 30(3):342–355, 2021.

[3] 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.

[4] 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.

[5] Johannes Bracher, Evan L Ray, Tilmann Gneiting, and Nicholas G Reich. Evaluating epidemic forecasts in an interval format. *PLoS computational biology*, 17(2):e1008618, 2021.

[6] David Braun, Daniel Ingram, David Ingram, Bilal Khan, Jessecae Marsh, and Thomas McAndrew. Incorporating crowdsourced perceptions of human behavior into computational forecasts of us national incident cases of covid-19., 2022. Access at https://doi.org/10.17605/OSF.IO/FVD4H.

[7] Mark A Burgman, Marissa McBride, Raquel Ashton, Andrew Speirs-Bridge, Louisa Flander, Bonnie Wintle, Fiona Fidler, Libby Rumpff, and Charles Twardy. Expert status and performance. *PloS one*, 6(7):e22998, 2011.

[8] Robert T Clemen and Robert L Winkler. Combining probability distributions from experts in risk analysis. *Risk analysis*, 19(2):187–203, 1999.

[9] Roger Cooke et al. *Experts in uncertainty: opinion and subjective probability in science*. Oxford University Press on Demand, 1991.

[10] Estee Y Cramer, Evan L Ray, Velma K Lopez, Johannes Bracher, Andrea Brennen, Alvaro J Castro Rivadeneira, Aaron Gerding, Tilmann Gneiting, Katie H House, Yuxin Huang, et al. Evaluation of individual and ensemble probabilistic forecasts of covid-19 mortality in the us. *Medrxiv*, 2021.

[11] Sara Y Del Valle, Benjamin H McMahon, Jason Asher, Richard Hatchett, Joceline C Lega, Heidi E Brown, Mark E Leany, Yannis Pantazis, David J Roberts, Sean Moore, et al. Summary results of the 2014-2015 darpa chikungunya challenge. *BMC infectious diseases*, 18(1):1–14, 2018.

[12] Ensheng Dong, Hongru Du, and Lauren Gardner. An interactive web-based dashboard to track covid-19 in real time. *The Lancet infectious diseases*, 20(5):533–534, 2020.

[13] Michelle V Evans, Tad A Dallas, Barbara A Han, Courtney C Murdock, and John M Drake. Data-driven identification of potential zika virus vectors. *elife*, 6:e22053, 2017.

[14] 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.

[15] Sebastian Funk, Marcel Salathé, and Vincent AA Jansen. Modelling the influence of human behaviour on the spread of infectious diseases: a review. *Journal of the Royal Society Interface*, 7(50):1247–1256, 2010.

[16] Gerd Gigerenzer and Wolfgang Gaissmaier. Heuristic decision making. *Annual review of psychology*, 62:451–482, 2011.

[17] Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):243–268, 2007.

[18] Thomas L Griffiths and Joshua B Tenenbaum. Optimal predictions in everyday cognition. *Psychological science*, 17(9):767–773, 2006.

[19] Bogdan Grigore, Jaime Peters, Christopher Hyde, and Ken Stein. Explicit: a feasibility study of remote expert elicitation in health technology assessment. *BMC medical informatics and decision making*, 17(1):1–10, 2017.

[20] Anca M Hanea, Victoria Hemming, and Gabriela F Nane. Uncertainty quantification with experts: present status and research needs. *Risk Analysis*, 2021.

[21] Victoria Hemming, Mark A Burgman, Anca M Hanea, Marissa F McBride, and Bonnie C Wintle. A practical guide to structured expert elicitation using the idea protocol. *Methods in Ecology and Evolution*, 9(1):169–180, 2018.

[22] Hideo Hirose and Liangliang Wang. Prediction of infectious disease spread using twitter: A case of influenza. In *2012 Fifth International Symposium on Parallel Architectures, Algorithms and Programming*, pages 100–105. IEEE, 2012.

[23] Rouba Ibrahim, Song-Hee Kim, and Jordan Tong. Eliciting human judgment for prediction algorithms. *Management Science*, 67(4):2314–2325, 2021.

[24] David Ingram and Daniel Ingram. Covid-19 mitigations in the U.S. September 2020 to April 2021. Technical report, Society of Actuaries, 2021. Access at https://www.soa.org/globalassets/assets/files/resources/research-report/2021/covid-eight-month-report.pdf.

[25] Michael A Johansson, Karyn M Apfeldorf, Scott Dobson, Jason Devita, Anna L Buczak, Benjamin Baugher, Linda J Moniz, Thomas Bagley, Steven M Babin, Erhan Guven, et al. An open challenge to advance probabilistic forecasting for dengue epidemics. *Proceedings of the National Academy of Sciences*, 116(48):24268–24274, 2019.

[26] Daniel Kahneman and Amos Tversky. Subjective probability: A judgment of representativeness. *Cognitive psychology*, 3(3):430–454, 1972.

[27] Daniel Kahneman and Amos Tversky. On the psychology of prediction. *Psychological review*, 80(4):237, 1973.

[28] Pei-Ying Kobres, Jean-Paul Chretien, Michael A Johansson, Jeffrey J Morgan, Pai-Yei Whung, Harshini Mukundan, Sara Y Del Valle, Brett M Forshey, Talia M Quandelacy, Matthew Biggerstaff, et al. A systematic review and evaluation of zika virus forecasting and prediction research during a public health emergency of international concern. *PLoS neglected tropical diseases*, 13(10):e0007451, 2019.

[29] 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.

[30] 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.

[31] Thomas McAndrew, Nutcha Wattanachit, Graham C Gibson, and Nicholas G Reich. Aggregating predictions from experts: A review of statistical methods, experiments, and applications. *Wiley Interdisciplinary Reviews: Computational Statistics*, 13(2):e1514, 2021.

[32] Daniel J McDonald, Jacob Bien, Alden Green, Addison J Hu, Nat DeFries, Sangwon Hyun, Natalia L Oliveira, James Sharpnack, Jingjing Tang, Robert Tibshirani, et al. Can auxiliary indicators improve covid-19 forecasting and hotspot prediction? *Proceedings of the National Academy of Sciences*, 118(51), 2021.

[33] Michelle McDowell and Perke Jacobs. Meta-analysis of the effect of natural frequencies on bayesian reasoning. *Psychological bulletin*, 143(12):1273, 2017.

[34] Philip M Polgreen, Forrest D Nelson, George R Neumann, and Robert A Weinstein. Use of prediction markets to forecast infectious disease activity. *Clinical Infectious Diseases*, 44(2):272–279, 2007.

[35] 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.

[36] Evan L Ray, Nutcha Wattanachit, Jarad Niemi, Abdul Hannan Kanji, Katie House, Estee Y Cramer, Johannes Bracher, Andrew Zheng, Teresa K Yamana, Xinyue Xiong, et al. Ensemble forecasts of coronavirus disease 2019 (covid-19) in the us. *MedRXiv*, 2020.

[37] Gabriel Recchia, Alexandra LJ Freeman, and David Spiegelhalter. How well did experts and laypeople forecast the size of the covid-19 pandemic? *PloS one*, 16(5):e0250935, 2021.

[38] Nicholas G Reich, Craig J McGowan, Teresa K Yamana, Abhinav Tushar, Evan L Ray, Dave Osthus, Sasikiran Kandula, Logan C Brooks, Willow Crawford-Crudell, Graham Casey Gibson, et al. Accuracy of real-time multimodel ensemble forecasts for seasonal influenza in the us. *PLoS computational biology*, 15(11):e1007486, 2019.

[39] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[40] Skipper Seabold and Josef Perktold. Statsmodels: Econometric and statistical modeling with python. In *Proceedings of the 9th Python in Science Conference*, volume 57, page 61. Austin, TX, 2010.

[41] Tara Kirk Sell, Kelsey Lane Warmbrod, Crystal Watson, Marc Trotochaud, Elena Martin, Sanjana J Ravi, Maurice Balick, and Emile Servan-Schreiber. Using prediction polling to harness collective intelligence for disease forecasting. *BMC public health*, 21(1):1–9, 2021.

[42] Anuj K Shah and Daniel M Oppenheimer. Heuristics made easy: an effort-reduction framework. *Psychological bulletin*, 134(2):207, 2008.

[43] Cuihua Shen, Anfan Chen, Chen Luo, Jingwen Zhang, Bo Feng, Wang Liao, et al. Using reports of symptoms and diagnoses on social media to predict covid-19 case counts in mainland china: Observational infoveillance study. *Journal of medical Internet research*, 22(5):e19421, 2020.

[44] Mark T Spence and Merrie Brucks. The moderating effects of problem characteristics on experts' and novices' judgments. *Journal of marketing Research*, 34(2):233–247, 1997.

[45] Thomas R Stewart, Paul J Roebber, and Lance F Bosart. The importance of the task in analyzing expert judgment. *Organizational behavior and human decision processes*, 69(3):205–219, 1997.

[46] Amos Tversky and Daniel Kahneman. Availability: A heuristic for judging frequency and probability. *Cognitive psychology*, 5(2):207–232, 1973.

[47] Amos Tversky and Daniel Kahneman. Evidential impact of base rates. Technical report, Stanford Univ Ca Dept Of Psychology, 1981.

[48] Amos Tversky and Daniel Kahneman. Rational choice and the framing of decisions. In *Multiple criteria decision making and risk analysis using microcomputers*, pages 81–126. Springer, 1989.

[49] Cécile Viboud, Kaiyuan Sun, Robert Gaffey, Marco Ajelli, Laura Fumanelli, Stefano Merler, Qian Zhang, Gerardo Chowell, Lone Simonsen, Alessandro Vespignani, et al. The rapidd ebola forecasting challenge: Synthesis and lessons learnt. *Epidemics*, 22:13–21, 2018.

[50] Robert L Winkler, Yael Grushka-Cockayne, Kenneth C Lichtendahl Jr, and Victor Richmond R Jose. Probability forecasts and their combination: A research perspective. *Decision Analysis*, 16(4):239–260, 2019.

[51] Markus Wübben and Florian v Wangenheim. Instant customer base analysis: Managerial heuristics often "get it right". *Journal of Marketing*, 72(3):82–93, 2008.

[52] Shihao Yang, Mauricio Santillana, and Samuel C Kou. Accurate estimation of influenza epidemics using google search data via argo. *Proceedings of the National Academy of Sciences*, 112(47):14473–14478, 2015.