PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Hindcasts and forecasts of suicide mortality in US: A modeling study

Sasikiran Kandula[1]*, Mark Olfson[2,3], Madelyn S. Gould[2,3], Katherine M. Keyes[2], Jeffrey Shaman[1]*

1 Department of Environmental Health Sciences, Columbia University, New York, New York, United States of America, **2** Department of Epidemiology, Columbia University, New York, New York, United States of America, **3** Department of Psychiatry, Columbia University, New York, New York, United States of America

* sk3542@cumc.columbia.edu (SK); jls106@cumc.columbia.edu (JS)

OPEN ACCESS

Citation: Kandula S, Olfson M, Gould MS, Keyes KM, Shaman J (2023) Hindcasts and forecasts of suicide mortality in US: A modeling study. PLoS Comput Biol 19(3): e1010945. https://doi.org/10.1371/journal.pcbi.1010945

Editor: Lusha Zhu, Peking University, CHINA

Received: July 8, 2022

Accepted: February 13, 2023

Published: March 13, 2023

Data Availability Statement: Forecasts and hindcasts outputs are provided as supplementary material. State monthly mortality counts are included as supplementary material, but can also be downloaded from CDC WONDER. Note that the public version suppresses instances where counts are below 10, and this leads to missing data during some of the months in low population states, but would not impact a vast majority of the states. Mortality data used in this study were made available to the authors under a restricted use agreement without suppression. Crisis hotline call volume are third-party data acquired under restricted use agreements and do not allow sharing

Abstract

Deaths by suicide, as well as suicidal ideations, plans and attempts, have been increasing in the US for the past two decades. Deployment of effective interventions would require timely, geographically well-resolved estimates of suicide activity. In this study, we evaluated the feasibility of a two-step process for predicting suicide mortality: a) generation of *hindcasts*, mortality estimates for past months for which observational data would not have been available if forecasts were generated in real-time; and b) generation of forecasts with observational data augmented with hindcasts. Calls to crisis hotline services and online queries to the Google search engine for suicide-related terms were used as proxy data sources to generate hindcasts. The primary hindcast model (*auto*) is an Autoregressive Integrated Moving average model (ARIMA), trained on suicide mortality rates alone. Three regression models augment hindcast estimates from *auto* with call rates (*calls*), GHT search rates (*ght*) and both datasets together (*calls_ght*). The 4 forecast models used are ARIMA models trained with corresponding hindcast estimates. All models were evaluated against a *baseline* random walk with drift model. Rolling monthly 6-month ahead forecasts for all 50 states between 2012 and 2020 were generated. Quantile score (QS) was used to assess the quality of the forecast distributions. Median QS for *auto* was better than *baseline* (0.114 vs. 0.21. Median QS of augmented models were lower than *auto*, but not significantly different from each other (Wilcoxon signed-rank test, $p > .05$). Forecasts from augmented models were also better calibrated. Together, these results provide evidence that proxy data can address delays in release of suicide mortality data and improve forecast quality. An operational forecast system of state-level suicide risk may be feasible with sustained engagement between modelers and public health departments to appraise data sources and methods as well as to continuously evaluate forecast accuracy.

Author summary

Suicide deaths in the United States have increased considerably during the last two decades. Effective deployment of interventions can benefit from the availability of timely

data publicly. Data had no personal identifying information, including originating phone number or any demographic attributes, and did not include call content. Access to search volume data through the Google Health Trends API can be requested by email to trends-api-support@google.com, or filling a form https://docs.google.com/forms/d/e/1FAIpQLSenHdGiGl1YF-7rVDDmmulN8R-ra9MnGLLs7gIIaAX9VHPdPg/viewform.

Funding: This work is funded by a grant from the National Institute of Mental Health (R01-MH121410) to KMK and JS. The funder had no role in study design; collection, analysis, and interpretation of data; preparation of the manuscript; and the decision to submit for publication.

Competing interests: I have read the journal's policy and the authors of this manuscript have the following competing interests: KMK has been compensated for expert witness work in litigation; JS and Columbia University declare partial ownership of SK Analytics; JS was a consultant for BNI. All other authors declare no competing interests.

geographically well-resolved forecasts of suicide activity. Data from the National Vital Statistics System (NVSS), the most reliable source of mortality in the US, are released in yearly increments, thus constraining the ability of a forecast system reliant solely on this dataset, in generating timely estimates of current/future suicide activity.

In this study, we explored the feasibility of generating real-time monthly state-level forecasts of suicide deaths in the US. We augmented NVSS data with two sources of suicide-related behavior — calls to crisis hotline services and online queries to the Google search engine for suicide-related terms — and evaluated their utility as proxies to traditional surveillance systems.

Our results show that forecasts from standard autoregressive models improved over benchmark models; delays in release of suicide mortality data from traditional surveillance sources can be addressed in part using proxy data; and, besides providing more timely estimates of recent suicide mortality, proxy-based hindcast estimates improved forecast quality.

Introduction

Deaths from suicide have risen for the last two decades in the United States [1,2]. Large-scale surveys have shown that besides deaths by suicide, a sizeable proportion of the population has suicidal thoughts (in 2020, 4.9 percent of adults 18 years or older; 12.2 million persons), and has planned (1.3%; 3.2 million) or attempted (0.5%; 1.2 million) suicide [3]. Several studies have quantified the differential effects of race, sex, and socioeconomic status on suicide risk and have documented higher risk among veterans and incarcerated populations [4–10]. Alongside these demographic attributes, geographical variability has also been noted with suicide mortality significantly higher in the midwest-, northwest and mountain states in the US [11].

In response, a broad range of preventive public health measures has been proposed and implemented. These encompass communication-oriented approaches to improve societal perception of suicide risk, reduce stigma, and promote help-seeking behavior, as well as measures to strengthen social connections, improve access to mental health services, and reduce access to lethal means [11,12].

While there is a significant body of research on associations between suicide mortality and individual and population level characteristics, and the relative efficacy of various public health interventions, predictive models of population risk remain relatively rare. Population level models can be useful for detecting changes in risk, either in overall trends or the emergence of suicide clusters, and can inform decisions on the timing, location and the type of interventions that are needed, or in evaluating interventions already in place. Previous studies have predominantly focused on national-level estimates in the US [13–15]. We believe that forecasts would be more actionable if they were generated with finer geographical resolution. For example, a forecasted increase in suicide deaths in a state can be used to trigger public health messaging about identifying warning signs, advertising of crisis hotline and other mental health services, and targeting of at-risk communities in that state. Therefore, we here explore the feasibility of generating real-time monthly forecasts of suicide deaths in each of the fifty states in the US.

A critical requirement for generating reliable forecasts is the availability of good quality, timely data, ideally covering a range of suicide behaviors from thoughts to deaths. However, accessing these data in real time remains challenging [12]. To focus on suicide deaths, the primary outcome in this study, the most reliable source of mortality in the US is the National

Vital Statistics System (NVSS). Deaths reported to NVSS are released in yearly increments, resulting in a lag in availability of 11 to 23 months. Although the importance of accurate and robust surveillance estimates is indisputable, a forecast system that is reliant solely on this dataset would be constrained in generating timely forecasts.

An important contribution of this study is an evaluation of two proxy sources of suicide-related behavior to address the data lag in traditional surveillance systems. Crisis hotline telephone services, one of the data sources used here, connect individuals to crisis counselors and are a critical resource to those at risk of suicide. Call volumes to crisis call centers have increased over the last two decades and multiple studies indicate their effectiveness, with callers self-reporting fewer mental health crises or suicidal states in follow-up assessments [16–18]. Similarly, researchers have hypothesized that online activity in the aggregate, such as social media use, access of suicide-seeking and suicide-prevention websites, or related queries to search engines can predict suicidal ideation at a population level [19–23]. Here, we focused on query frequencies to the Google search engine for suicide and mental health related terms. Associations of suicide deaths with crisis calls and online searches are complex and understudied with potentially large variability by location, time, and demographic characteristics. This analysis was limited to evaluating the utility of these data sources as predictive features in a forecast system and does not attempt to elucidate causal processes.

Using standard time series modeling methods and a combination of traditional and alternative data sources, we generated retrospective rolling monthly forecasts over nine years (2012-2020) for each of the 50 states in the US. This process included an intermediate step of generating *hindcast* estimates of suicide mortality, i.e. estimates for past months for which actual mortality data would not have been available if forecasts were generated in real-time. We report: a) the accuracy and reliability of the forecasts and intermediate hindcasts; b) the improvement in forecast quality from including hindcasts, overall and stratified by state and period; and, c) the relative difference in accuracy between forecasts and hindcasts from the two alternative sources of suicide-related activity.

Materials and methods

Suicide mortality rate

Records of all-cause deaths were obtained from the US National Vital Statistics System (NVSS) [24] and suicide deaths were identified using *International Classification of Diseases*, *Tenth Revision* underlying cause-of-death codes X60-X84, Y87.0, and U03 [25]. Monthly suicide counts in each state were calculated using the decedent's county of residence and month of death. Annual state population estimates were obtained from the Bridged-Race Intercensal (2005-2010) [26] and Postcensal (2011-2020) [27] datasets and were assumed to remain unchanged during a calendar year. Monthly suicide mortality rates were calculated as suicide deaths per 100,000 population.

Crisis hotline call rates

The 988 Suicide and Crisis Lifeline [Lifeline; https://988lifeline.org/], a network of over 200 round-the-clock toll-free centers, is the primary hotline for suicidal crisis and emotional distress counseling services in the US, accessible by phone as well as chat/text. Logs of calls routed to Lifeline centers were used to estimate the date of each call and caller location (inferred from the first six digits of the phone number) and aggregated to calculate monthly state-level call volumes. As above, annual state populations were used to estimate call rates per 100,000 population. As call rates are not normally distributed, a log transformation was applied. Access to Lifeline call volumes is a possibility but does not currently exist.

Google Health Trends (GHT) API

The GHT API provides estimates of the proportion of user sessions on the Google search engine that included a query for a specified term. These estimates can be stratified by geography (country, state, etc.) and time (month, week, or day), and a historical record since 2004 is available. To identify terms whose search frequency can predict suicide mortality rates, we relied on prior studies that identified 6 categories comprising 111 suicide-related terms — *suicide seeking* (e.g. commit suicide), *suicide prevention* (e.g. suicide hotline), *suicide neutral* (e.g. suicides), *mood and anxiety* (e.g. depressed), *psychosis* (e.g. delusion) and *stressor or trauma* (e.g. social isolation) [23,28–40] (see S1 Appendix for a list of terms, by category). For each state, monthly search rates for a term category, defined as the proportion of user sessions from the state during a month that included one or more terms in the category, were retrieved and logit transformed. This choice is in part motivated by previous analyses that estimated influenza from search rates and found a logit transformation on search rates useful due to an approximately linear relationship between predictor and response in the logit space [41–43].

Monthly Google Health Trends (GHT) data are available at the end of each month, with a lag of less than one week. Access to the data feed must be requested through Google (see *Data availability*).

Forecast generation

All-cause public use and restricted use mortality datasets are released by NVSS in one-year increments, usually in December of the following year, resulting in a 11 to 23 month lag for monthly suicide mortality records. For example, mortality data for all of 2020 were released in December 2021, resulting in a lag of 11 months for deaths that occurred in December 2020 and a lag of 23 months for those in January 2020. Provisional mortality counts for certain causes [44] may be available sooner, with varying degrees of completeness.

Forecast generation at month m proceeds in two steps: a) generation of *hindcast* estimates of suicide mortality for the time period between month m and the last available real observation (at earliest *m-12*); and b) generation of 6-month ahead forecasts using observations and hindcast estimates up to month m. Formally, let y_k denote the observed suicide mortality rate in month k, Y_k the time series of rates up to month k, $(y_1,\cdots,y_k)$; $\bar{y}_{m-l}$ the hindcast estimate for month *m-l* generated at month m and $\bar{Y}_l$ the time series $(\bar{y}_{m-l},\cdots,\bar{y}_m)$. At month m, hindcast estimates for l past months were generated using Y_k, and forecast estimates for h months in the future, $\hat{y}_{m+1},\ldots,\hat{y}_{m+h}$, were generated using the time series $(Y_k, \bar{Y}_l)$. If the last available observation is for month k, $l = m\text{-}k\text{-}1$, and as noted above, $12 \leq l \leq 24$ and $1 \leq h \leq 6$

Model specification

Models used in this study are primarily based on the Autoregressive Integrated Moving Average (ARIMA) [45,46] approach and are described in detail in Text A in S1 Text. The primary hindcast model (*auto*) was trained on the time series of suicide mortality rates alone, and included components to capture trend and seasonality, the latter modeled with Fourier terms. Three models augment hindcast estimates from *auto* by including call rates (*calls*), GHT search rates (*ght*) and both datasets together (*calls_ght*). All 4 models were evaluated against a *baseline* persistence model with no trend or seasonality components, but which accounts for the average change per time step (random walk with drift [47,48]).

The models used to generate 6-month ahead forecasts were identical to the *auto* model described above and differed only in the data used to train the models — *auto*, *calls*, *ght* and

calls_ght forecasts were generated using respective hindcast estimates appended to observed mortality rates. Change in hindcast/forecast quality relative to the *baseline* can be interpreted as the effect of a more careful modeling of the characteristics of the time series, while the difference of the augmented models, *calls*, *ght* and *calls_ght*, relative to *auto* can be interpreted as the advantage from incorporating more timely surveillance proxies.

Implementation of all models are available in *R* [49] packages *forecast [50]* and *fable [48]*.

Experimental setup

The models were trained independently for each state using data solely pertaining to that state. Monthly data from all three sources were available for years 2007-2020 for all 50 states in the US. Suicide mortality data for 2020, which were set aside as a test set, were not used in model selection or hyperparameter tuning. We defined 2012 through 2019 as the validation period and generated rolling forecasts beginning January 2012, incrementing the training window one month at a time, and using only (and all) the data that would have been available were the estimates generated in real time (Fig A in S1 Text). For example, to generate retrospective forecasts at the end of January 2012, suicide mortality data from January 2007 through December 2010, and call rates and search rates for January 2007 through January 2012 were used to first generate monthly hindcast estimates for January 2011 through January 2012, and these estimates appended to mortality observations in order to generate forecasts for months February 2012 through July 2012. Hindcast models at the end of February 2012, were retrained with one additional month's call and search rates, and forecast models were retrained with the new hindcast estimates. This iterative process was terminated at the end of 2019.

Model performance is reported separately for the validation (2012-2019) and test periods (2020), the latter being a better measure of model performance as it was withheld from both models and modelers.

Evaluation metrics

Forecast models were used to generate quantile estimates, $\hat{y}_{\alpha,m+h}$, at 23 levels, $\alpha = \{0.01, 0.025, 0.05, 0.1, 0.15, \ldots, 0.95, 0.975, 0.99\}$, and the median estimate, $\hat{y}_{.5,m+h}$, was used as a point estimate. Probabilistic and point estimates for hindcasts, $\bar{y}_{\alpha,m-l}$ and $\bar{y}_{.5,m-l}$ were defined analogously (see Text B in S1 Text).

Accuracy of forecast point estimates was evaluated with mean absolute proportionate error (MAPE), where the accuracy of forecasts generated at month m was calculated as $MAPE\_F_m = \frac{1}{h}\sum_h \frac{abs(y_h - \hat{y}_h)}{y_h}$. Accuracy of probabilistic forecasts were evaluated using quantile score (QS), calculated as $QS\_F_m = \sum_\alpha QS\_F_{\alpha,m}$, where

$$QS\_F_{\alpha,m} = 2 * \alpha * (y_h - \hat{y}_{\alpha,h}) * \mathbf{1}(y_h \geq \hat{y}_{\alpha,h}) + 2 * (1-\alpha) * (\hat{y}_{\alpha,h} - y_h) * \mathbf{1}(y_h < \hat{y}_{\alpha,h}),$$

and $\mathbf{1}()$ denotes the indicator function, and α denotes the quantile level [47,51]. MAPE and quantile score of hindcasts were calculated analogously.

Both metrics are non-negative and can be interpreted as penalty measures with a higher value indicating an inferior estimate; a value of 0 indicates a perfect estimate. Summary measures are reported by aggregating (mean) across states and/or years and months. For each pair of models, Wilcoxon signed rank test was used to assess whether the difference in model quality per each metric was statistically significant [52].

In order to calculate model performance relative to a reference model, a *relative* measure was calculated. Relative quantile score (RQS) of forecasts generated with model *A* relative to

reference model R is given by:

$$RQS\_F = \frac{100}{|s| * |m|} \sum_{s,m} \frac{QS\_F^A_{s,m} - QS\_F^R_{s,m}}{QS\_F^A_{s,m} + QS\_F^R_{s,m}}$$

where s and m denote the state and month at which forecasts were generated, respectively. RQS_F has well-defined bounds of [−100, 100] and is computable when at least one of the two models is not perfect ($QS > 0$). A negative RQS indicates an improvement over reference model.

Calibration of the hindcast and forecast probabilistic estimates were assessed by inspecting observations against estimated quantile distributions (see Text C in S1 Text). These were visualized as probability plots and the deviation from the diagonal was interpreted as a measure of miscalibration [53].

Results

Fig 1 shows pairwise correlation between suicide mortality rates, call rates and six GHT search rates, estimated across all locations and over the entire study period (n = 8400; 50 locations * 168 months). All exogenous rates were found to have statistically significant, but small (|Spearman's rho| < 0.15) correlations with mortality rates. Higher correlations were observed between call rates, and search rates for mood/anxiety (Spearman's rho = 0.40, $p <$ 1e-6) and

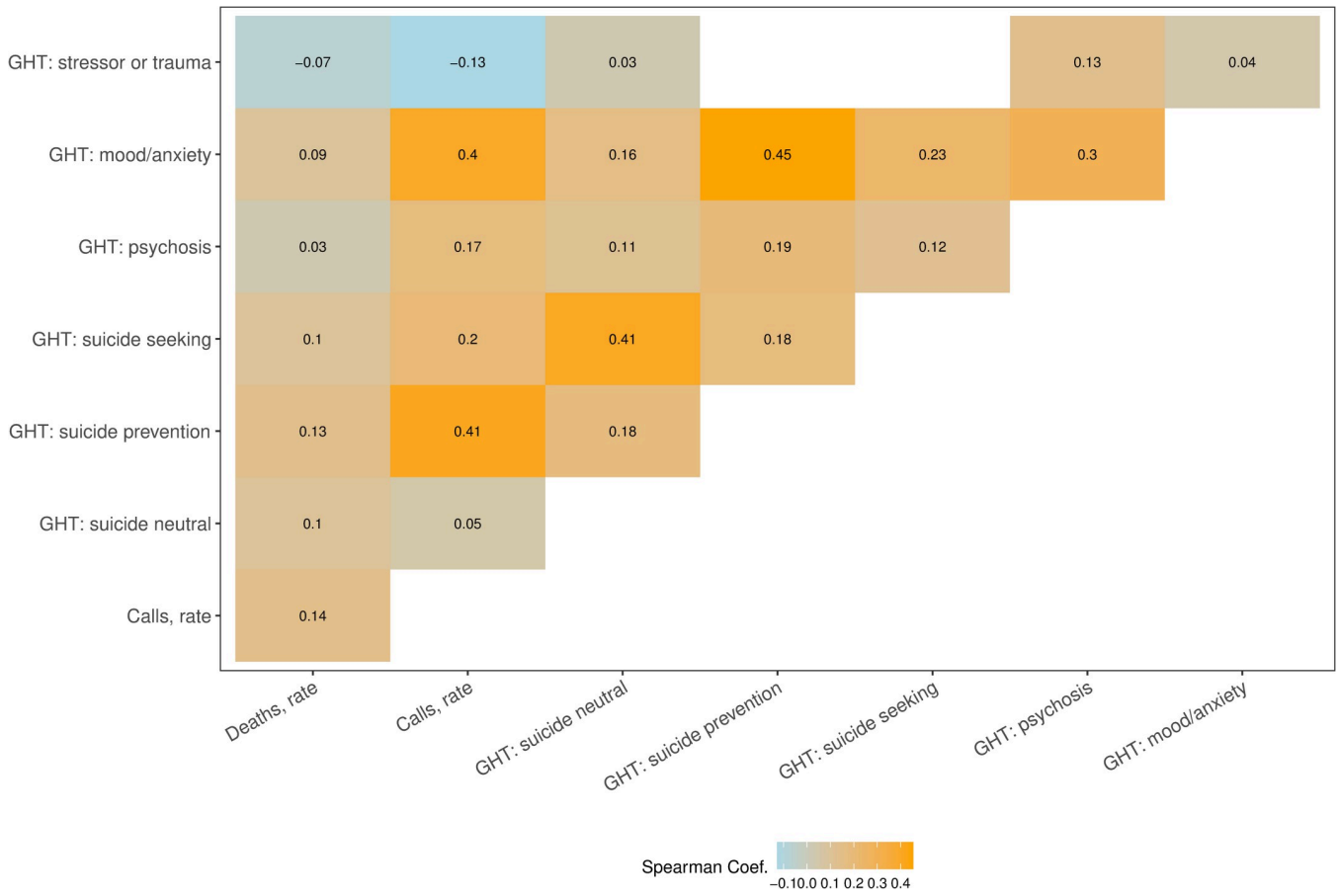

Fig 1. Pairwise Spearman correlation for each pair of variables. Correlations found to be not significant ($p > .05$) are not shown.

https://doi.org/10.1371/journal.pcbi.1010945.g001

suicide prevention (0.41, $p < 1e\text{-}6$) terms, as well as among search rate variables, particularly between mood/anxiety and suicide seeking (0.23, $p < 1e\text{-}6$), psychosis (0.3, $p < 1e\text{-}6$) and suicide prevention (0.45, $p < 1e\text{-}6$) terms.

Validation period (2012-2019)

Fig 2 shows improved forecast quality when trend and seasonality in mortality rates were included in the models — the median quantile score for *auto* is lower than *baseline* (0.114 vs. 0.21), with an increase in spread (interquartile range: 0.28 vs. 0.21), and noticeably longer left tail indicating a higher proportion of good forecasts. Informing the models with near real-time proxy data further improved forecast quality. Augmented models *calls*, *ght* and *calls_ght* all had a lower median quantile score than *auto*, but their median scores were nearly indistinguishable from each other. This was further verified using two-sided Wilcoxon signed rank test where statistically significant differences ($p < 1e\text{-}4$) were observed between quantile scores of *auto* and *baseline* and between *auto* and each of the three augmented models, but not among any pair of augmented models (see S1 Text for results using MAPE as evaluation metric).

While the improvement in *auto* relative to *baseline* was also evident in the intermediate hindcasts, *calls*, *ght* and *calls_ght* were not significantly different from *auto* (Fig B in S1 Text). Hindcasts from *calls* were found to have better quantile scores than from the other two augmented models.

Disaggregating forecasts by state showed consistent improvement of the augmented models' quantile scores relative to both *baseline* and *auto* (RQS) across all states, with some variability in magnitude (Figs 3 and C in S1 Text for MAPE). In each state, RQS' of the augmented models were nearly identical, suggesting little complementarity of the data sources and hence limited justification for their simultaneous inclusion in the models. This is in contrast to the differential state-wise improvement seen with hindcasts of the augmented models relative to *auto*, indicative of differences in predictive skill, and hence utility, of the data sources in generating hindcasts (Fig D in S1 Text).

No clear difference in RQS was apparent when forecasts were stratified by month, with the possible exception of months May-July. A decrease in RQS of *auto* relative to *baseline* near the end of the validation period was also observed, possibly due to the plateauing of suicide mortality rates from 2018, leading to a disruption in the historical trend on which *auto* relies (Fig 4).

Fig 5 demonstrates that hindcasts and forecasts from the augmented models have good calibration, with their corresponding probability plots nearly tracking the diagonal. While hindcasts from *auto* were also well-calibrated, their forecast distributions appear to have inadequate coverage for extreme instances of low and high mortality rates (also see Fig E in S1 Text). In contrast, both forecasts and hindcasts from *baseline* were miscalibrated, with hindcasts exhibiting a strong bias (forecasts skewed low/left) and a lack of sharpness.

Test period (2020)

Quantile score of forecasts and hindcasts from all models (except *baseline*) were higher for 2020 than in the validation period (Figs 6 and F in S1 Text). The largest deterioration in forecast quality occurred in the *auto* model, with the model underperforming *baseline* in 19 states. This is possibly an extension of the decrease in accuracy at the end of the validation period, due to lower suicide mortality beginning 2018. As seen in the calibration plots (Fig G in S1 Text), all models that included a trend component overestimated the mortality rate, with the miscalibration particularly evident for the *auto* model forecasts; augmented models continued to outperform the *baseline* and *auto* models.

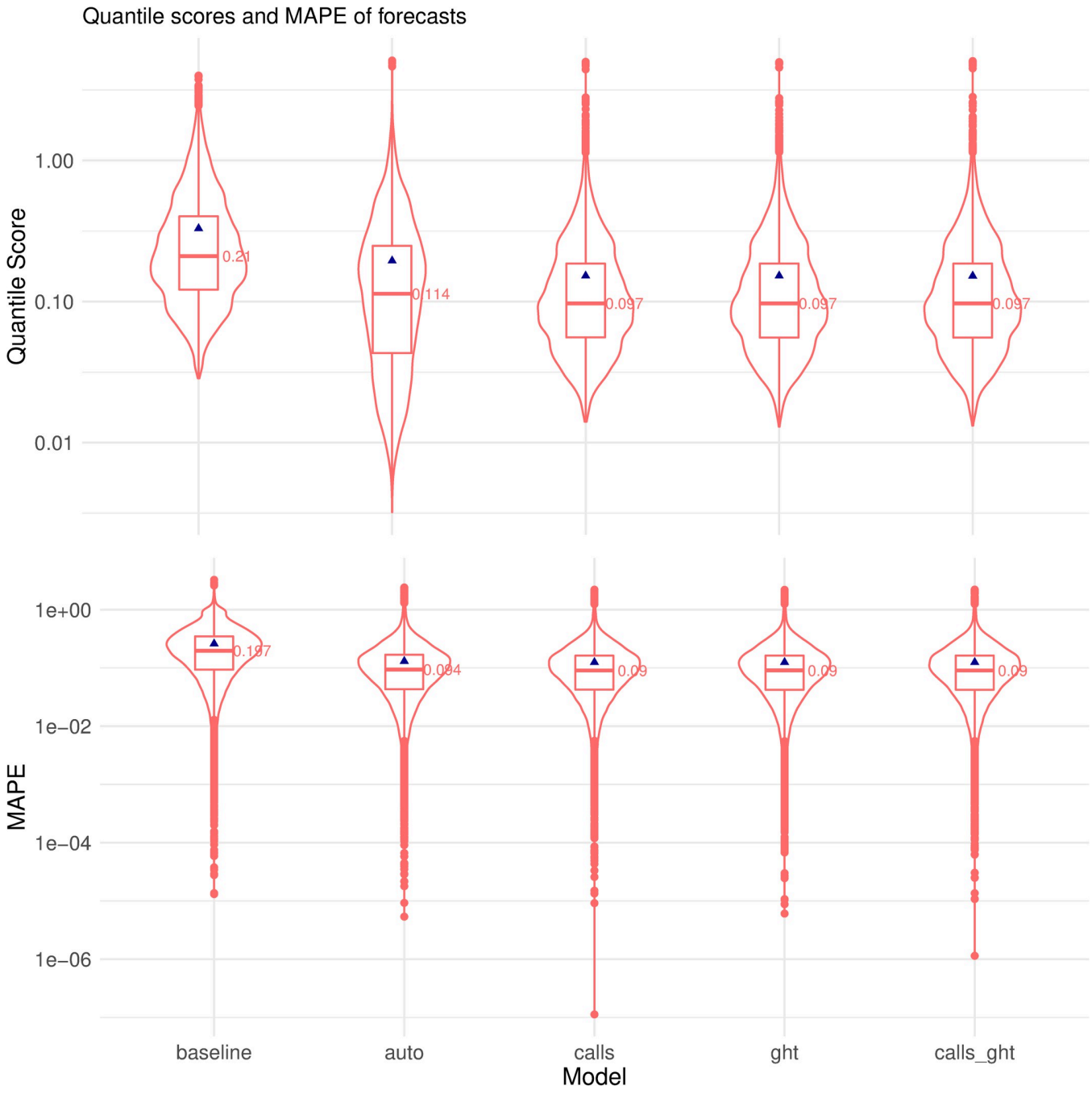

Fig 2. Boxplots of quantile scores and MAPE of forecasts from 5 models across all states and months. Blue points show mean estimate. *p*-value for Wilcoxon signed rank test on quantile scores: *calls*/*ght*=0.64; *calls*/*calls_ght* = 0.86; *ght*/*calls_ght* = 0.35. *p*-value for Wilcoxon signed rank test on MAPE: *calls*/*calls_ght* = 0.08; *ght*/*calls_ght* = 0.49. All other model pairs were statistically significant ($p < 1e\text{-}4$).

https://doi.org/10.1371/journal.pcbi.1010945.g002

Multi-model ensembles

We evaluated the utility of an ensemble of models by averaging forecasts from three alternative methods to the fixed ARIMA model described above — a neural network based model, an exponential trend smoothing model, and a more flexible ARIMA approach that searches the

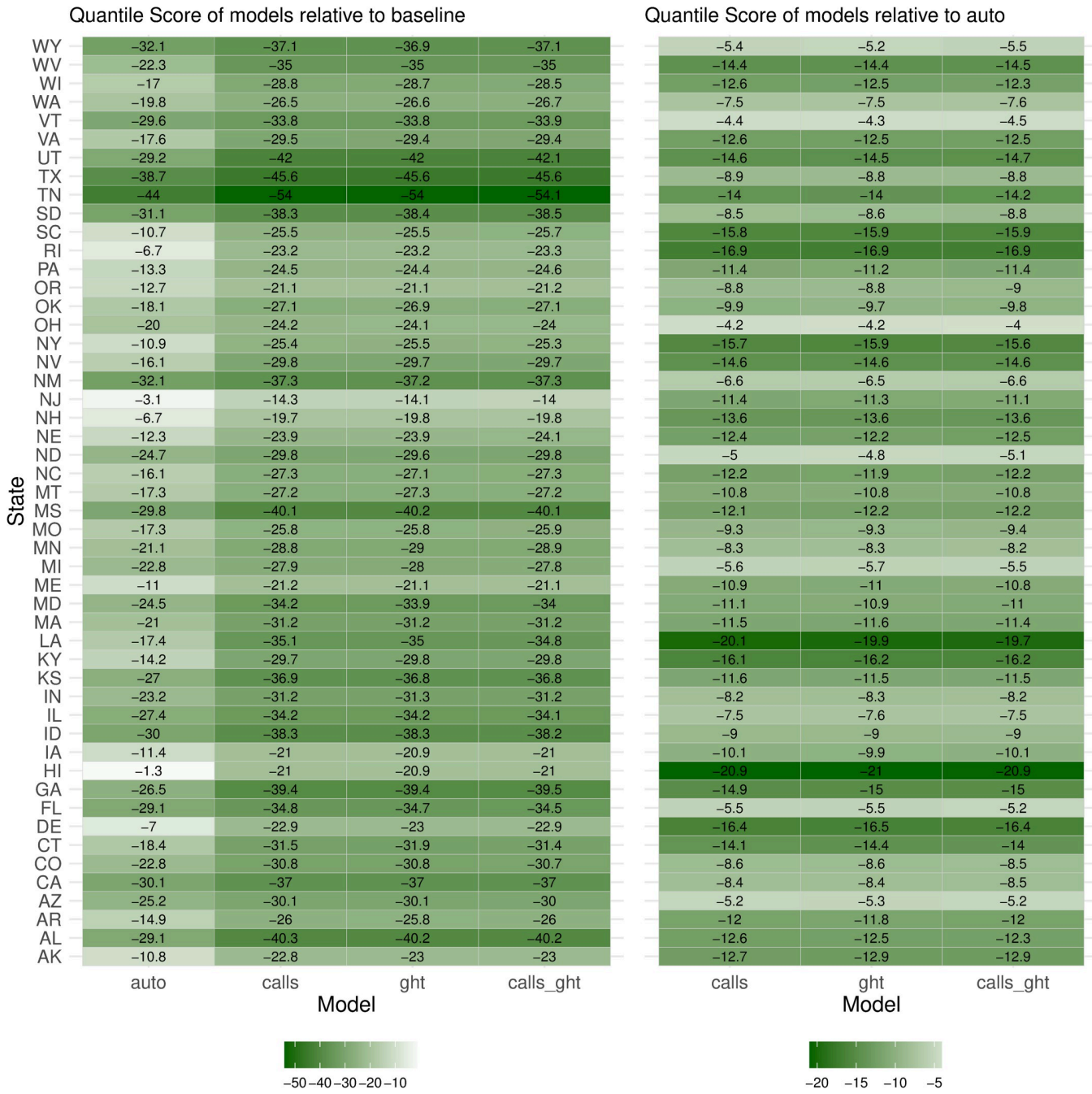

Fig 3. Quantile scores of forecasts from augmented models relative to *baseline* model (left) and relative to *auto* model (right), by state. The relative quantile score (RQS), has a range of -100 to 100, a negative value indicating a better forecast than the reference and a positive value indicating a worse forecast. The color lightness (light to dark) represents the magnitude of difference from reference, with a darker shade implying greater improvement.

https://doi.org/10.1371/journal.pcbi.1010945.g003

parameter space at each month to identify best fit (Text D in S1 Text). During the validation period, the forecast quality of the ensemble was found to be not statistically different from the primary model used in this study. Encouragingly, however, during the test period two of the component models and the ensemble overall had better forecasts (statistically significant) than the fixed ARIMA model (Fig H in S1 Text). Calibration also improved.

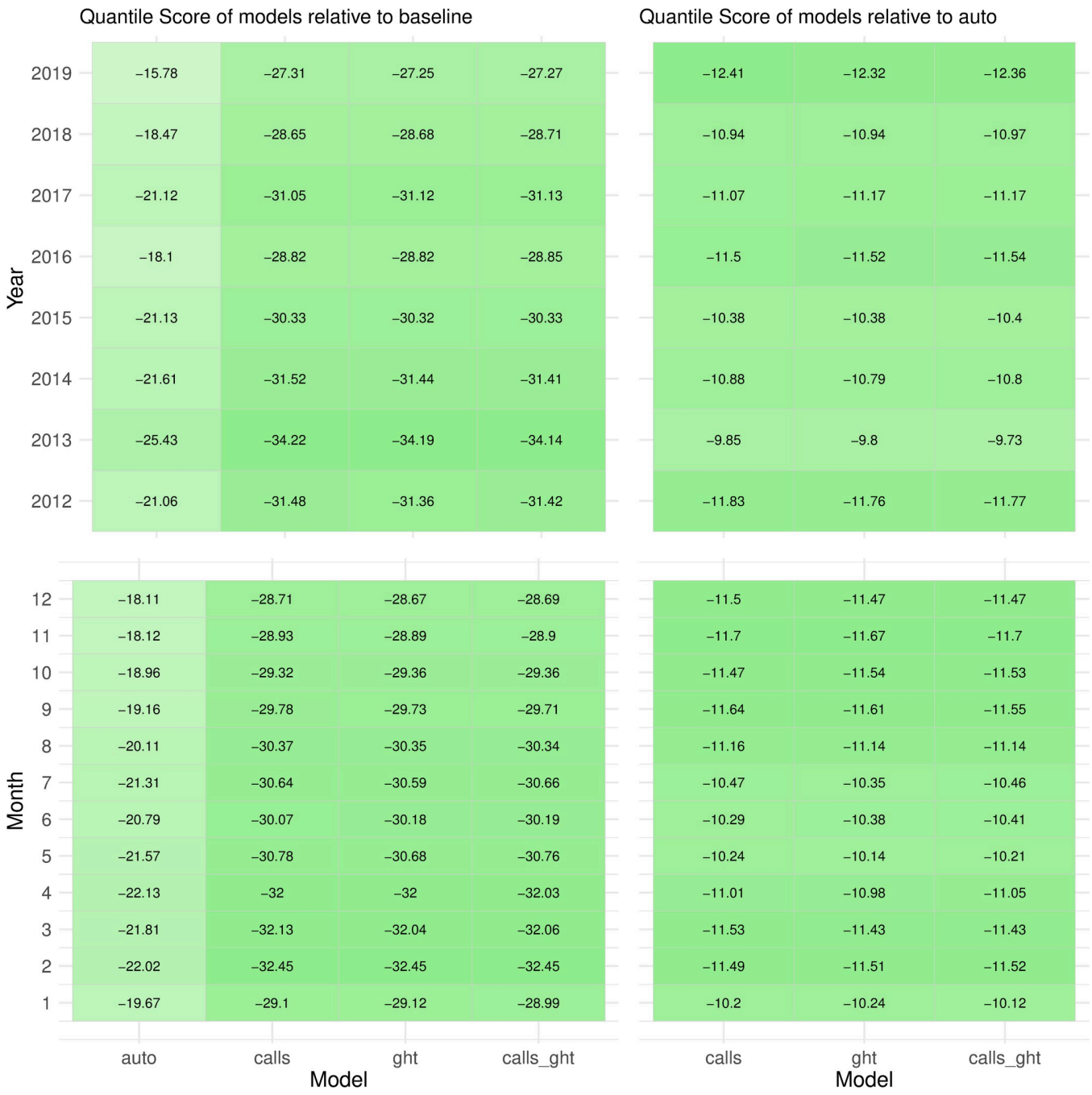

Fig 4. Quantile scores of forecasts from augmented models relative to *baseline* model (left column) and relative to *auto* model (right column), by year (top row) and month (bottom row). The relative quantile score (RQS), has a range of -100 to 100, a negative value indicating a better forecast than the reference and a positive value indicating a worse forecast. The color lightness (light to dark) represents the magnitude of difference from reference, with a darker shade implying greater improvement.

https://doi.org/10.1371/journal.pcbi.1010945.g004

Discussion

This study aimed to evaluate the feasibility of generating monthly 6-month ahead forecasts of suicide mortality in US states. We have shown that forecasts from standard autoregressive models improved over benchmark models. We have also demonstrated that delays in release

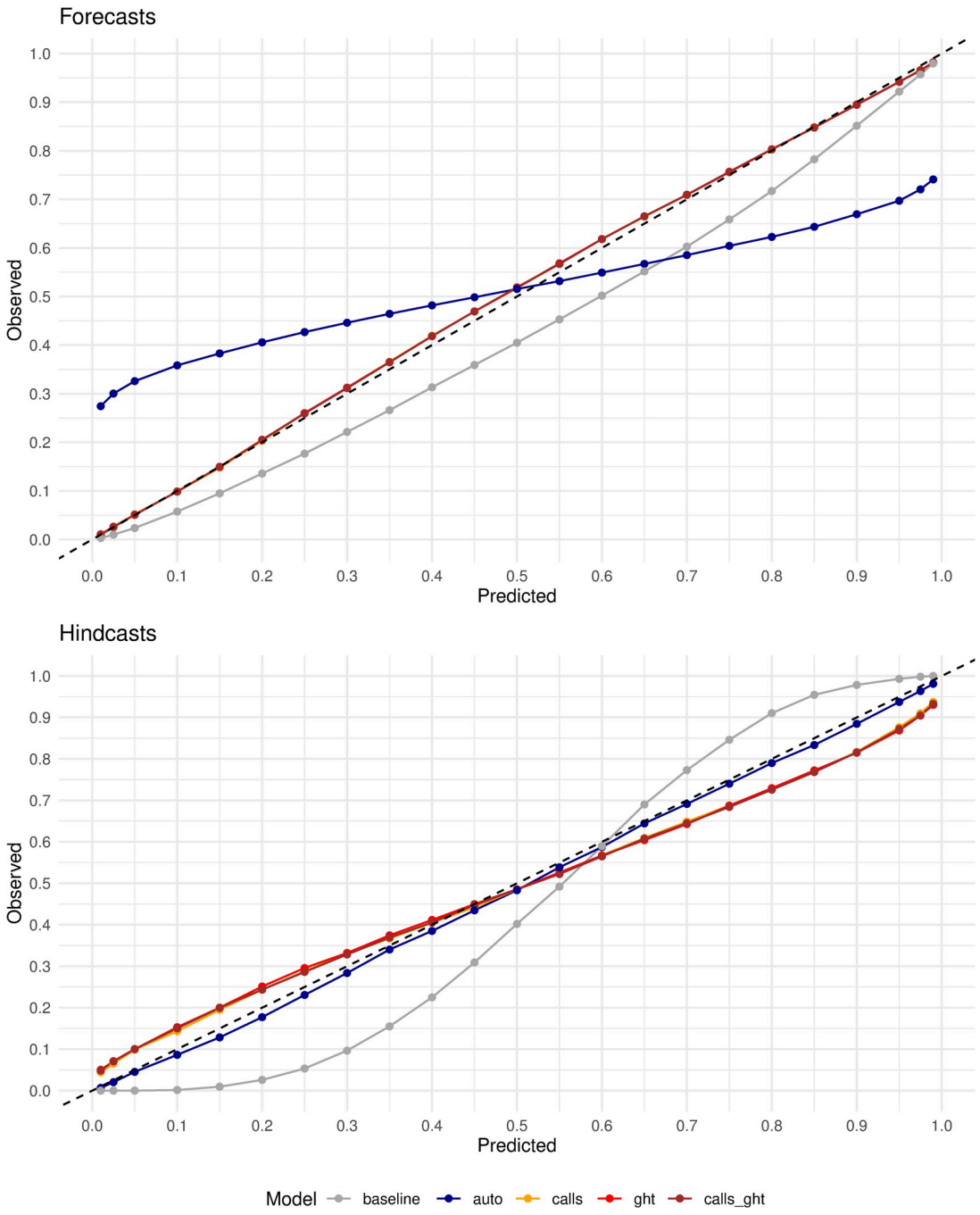

Fig 5. Calibration plot of forecasts (top) and hindcasts (bottom). Forecasts from *auto* (Cramer's distance [53]=0.03) and to a lesser degree *baseline* (0.006) appear to be miscalibrated, while the remaining three models have similar and better calibration (5e-4). On the other hand, hindcasts from *auto* have the best calibration (7e-4) hindcasts and *baseline* the least calibrated (.015); the augmented models have similar, good calibration (3e-3).

https://doi.org/10.1371/journal.pcbi.1010945.g005

of suicide mortality data from traditional surveillance sources can be partially addressed by using proxy data, and inclusion of proxy-based hindcast estimates, besides providing more timely estimates of recent suicide mortality, improved forecast quality and rendered forecasts more sensitive to changes in long-term trends (as evidenced by performance during the 2020

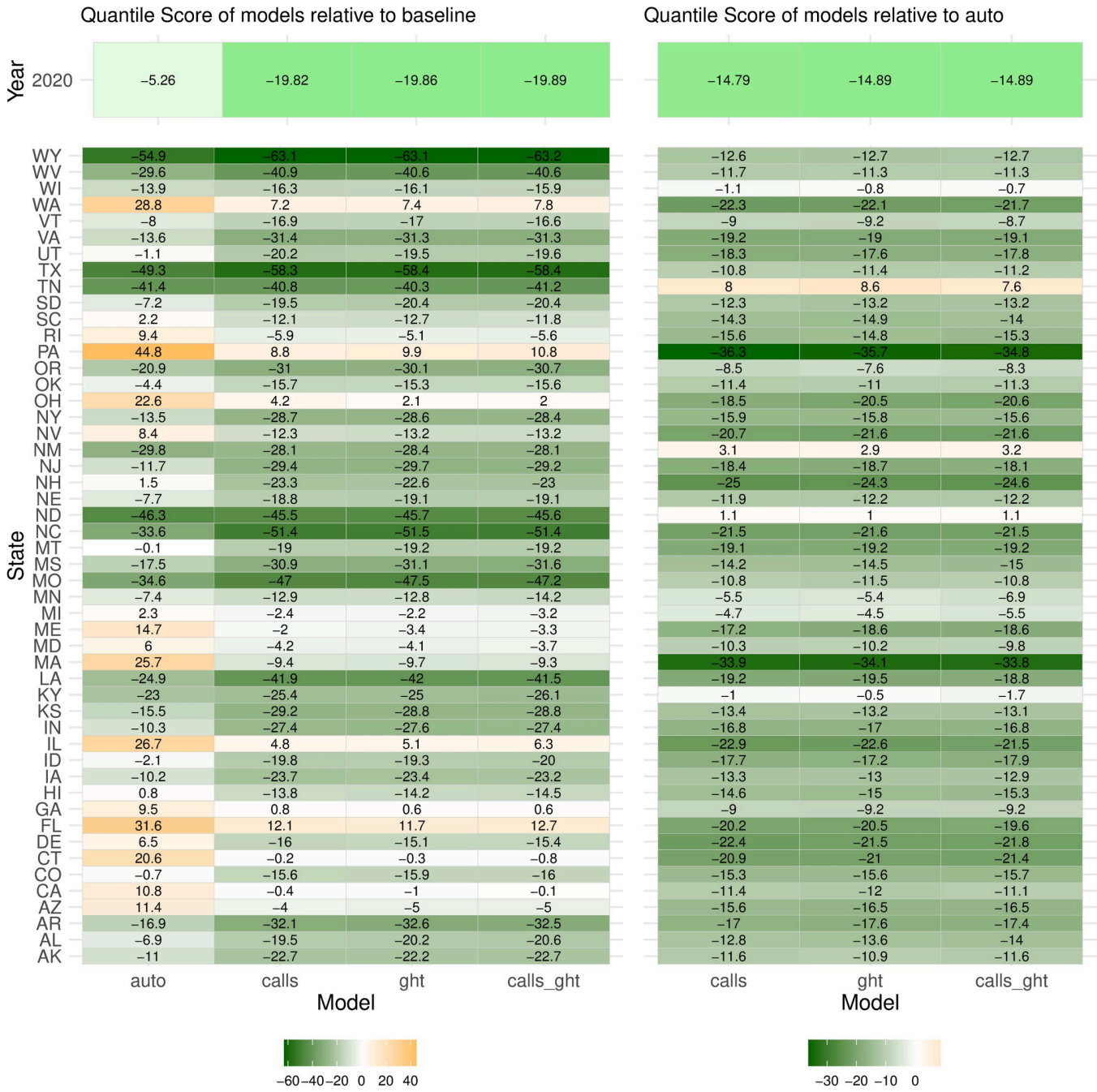

Fig 6. Quantile scores of forecasts from augmented models relative to *baseline* model (left) and relative to *auto* model (right), by state, during the test period (January 2020 – December 2020). Note that forecasts generated for the latter half of the year cannot be fully evaluated until mortality data for 2021 are available (for example, of forecasts generated at August 2020, 5-month ahead (Jan 2021) and 6-month ahead (Feb 2021) could not be evaluated).

https://doi.org/10.1371/journal.pcbi.1010945.g006

test period). Forecasts from augmented models were also better calibrated even when their advantages, as assessed by aggregate measures of error, were less clear.

The choice of ARIMA as the primary model framework was motivated by our prior experience with this method, availability of robust implementations, relative conceptual simplicity and computational efficiency. While hyperparameter tuning and data transformations were

handled to some extent by the *fable* software package, a more thorough exploration may further improve forecast quality. More recently developed time series modeling approaches, such as recurrent neural networks, have shown marked improvement in some domains [54,55], and their utility with the relatively short time series available here could be tested.

Although ARIMA-based forecast models can generate predictions at much longer horizons, the quality of the forecasts tends to degrade the farther ahead they project. The choice of a 6-month forecast horizon was believed to provide a reasonable trade-off between forecast quality and practical public health utility, and is in part influenced by forecast systems of influenza and other respiratory infections where a 4 time step horizon is commonly used.

The performance of the multi-model ensemble is in line with findings from prior studies in disease modeling [56–61] and other domains that ensembles often match or exceed individual model performance in the absence of a single reliably superior component model. Operationally, deploying an ensemble over a single model is likely to yield a more consistently good forecast quality. Modeling frameworks that capture suicide processes and mechanisms – thoughts, plans and attempts – would serve as valuable complements to the statistical models described here and need to be pursued as an important addition to ensembles [62,63].

Forecasts would be more actionable if they can be generated at sub-state resolutions, and/or tailored to specific population sub-groups (for example, young adults or marginalized communities). Models to flag emerging clusters among population subgroups would also be useful in deploying targeted interventions. Neither of the two data sources used in this study support stratifying by demographic attributes (such as age, sex and race/ethnicity), but aggregation at sub-state resolutions is possible, and reliability of such estimates remains to be investigated.

This study has several limitations. The models did not include socioeconomic or clinical predictors of mortality rate. We have described such models elsewhere [64] and note that the applicability of such methods would be contingent on the timely availability of covariate data. Similarly, better predictive models may be possible through inclusion of suicide-related information from neighboring states or the US overall rather than treating each state as an isolated entity. In addition, the precise physical locations of Lifeline callers were not available and our inference of location from caller area code could have introduced errors among mobile phone callers who relocated from their home state. Errors in mortality rate estimates are possible due to inconsistencies in death certification across states and study period, and potential undercounting of suicide deaths among certain racial/ethnic minority groups [65]. Additionally, the test period overlapped with the first year of the COVID-19 pandemic, and although the pandemic does not appear to have increased the annual suicide burden in most states [66], differences during some months, as well as a change in the relation between suicide mortality rates and call/search rates could have impacted the evaluation.

While the different skill metrics we report provide reasonable evidence of the utility of the proposed approach, a formal comparison with forecasts based on human expert judgement may be necessary to assess the value of automated approaches. Similarly, although we demonstrated the viability of a forecast system for state-level suicide mortality in a retrospective setting, there are challenges to operationalizing such a system. Lifeline call data are not public, and real-time access is uncertain. With the launch of a national suicide and mental health crisis number (988), nationwide call logs would be potentially warehoused in a central system, but no data sharing plans have been disclosed, possibly owing to confidentiality concerns. Prior studies have examined the possibility of using other auxiliary non-surveillance sources, such as social media posts with mixed results [19–22], to predict population-level suicide risk.

As illustrated by the complementarity of call and search rates for hindcast estimates, models built using multiple data sources may be better able to capture a wider range of intrinsic features and provide more resilience against data characteristics of a single source. While real-

time population-level forecast systems of suicide deaths hold promise, they would benefit from access to multiple historical suicide-related datasets to train models and timely release of up-to-date data.

Our results indicate that a simple autoregressive model built solely on public mortality data can substantially improve over baseline estimation approaches at the state-level and possibly at finer geographic scales such as county or city. GHT data are free and relatively easy to access over some of the other sources noted above and can improve forecast quality. Selection of a modeling approach, identification and evaluation of data sources on suicide behavior, and appraisal of the tradeoffs between complexity/expense and forecast accuracy depend on the context in which the models are deployed and used, and can benefit from sustained engagement between modelers and public health departments.

Supporting information

S1 Text. Text A: Hindcasts and forecasts. Text B: Generating quantile distributions and assessing point and probabilistic forecasts. Text C: Assessing model calibration. Text D: Multi-model ensembles. Fig A. Schematic representation of the time periods covered by observations, hindcasts and forecasts. At the current time, all observational data available are used to train the hindcast model and estimate mortality for past months without mortality observational data. Forecast models are trained on a time series stitched together with both mortality data and hindcast estimates. Fig B. Boxplots of quantile score and MAPE of hindcasts from 5 models across all states and months. Blue points show mean estimate. *p*-value for Wilcoxon signed rank test on quantile scores: *auto*/*calls*=0.3*; auto*/*ght*=0.43; *auto*/*calls_ght* = 0.47; *ght*/*calls_ght* = 0.87. *p*-value for Wilcoxon signed rank test on MAPE: *auto*/*calls* = 0.23; *ght*/*calls_ght* = 0.52. All other model pairs are statistically significant ($p < 1e\text{-}4$). Fig C. MAPE of forecasts relative to the baseline model (left) and relative to the auto model built with no real-time proxy data (right). Fig D. Quantile scores of hindcasts from augmented models relative to *baseline* model (left) and relative to *auto* model (right). Fig E. Hindcasts and 6-month ahead forecasts for 2019 in California. Distribution shown for the first month in each subpanel is the hindcast estimate and the rest are forecasts. Fig F. Quantile scores of hindcasts from augmented models relative to *baseline* model (left) and relative to *auto* model (right), during the test period (January 2020 – December 2020). Fig G. Calibration plot for forecasts (top) and hindcasts (bottom), during the test period (January 2020 – December 2020). Fig H. Quantile scores of forecasts from ensembles of augmented models relative to *baseline* model (left) and relative to *auto* model (right), by state, during the test period (January 2020 – December 2020).
(DOCX)

S1 Appendix. List of search terms used to query Google Health Trends API, by category.
(CSV)

S1 Data. All files are R datasets. Mortality.Rds: Monthly suicide deaths observed in each state during the study period. Nowcasts.Rds: Model hindcast estimates. Forecasts.Rds: Model forecast estimates.
(ZIP)

Acknowledgments

The authors thank Alena Goldstein, Johnathan Higgins and Sean Murphy of the Vibrant Emotional Health, 988 Suicide and Crisis Lifeline for access to call dataset as well as valuable feedback on the manuscript.

Author Contributions

Conceptualization: Sasikiran Kandula, Mark Olfson, Madelyn S. Gould, Katherine M. Keyes, Jeffrey Shaman.

Formal analysis: Sasikiran Kandula.

Funding acquisition: Katherine M. Keyes, Jeffrey Shaman.

Investigation: Sasikiran Kandula, Jeffrey Shaman.

Methodology: Sasikiran Kandula, Mark Olfson, Madelyn S. Gould, Katherine M. Keyes, Jeffrey Shaman.

Project administration: Katherine M. Keyes, Jeffrey Shaman.

Software: Sasikiran Kandula.

Supervision: Katherine M. Keyes, Jeffrey Shaman.

Validation: Sasikiran Kandula.

Visualization: Sasikiran Kandula.

Writing – original draft: Sasikiran Kandula, Katherine M. Keyes, Jeffrey Shaman.

Writing – review & editing: Sasikiran Kandula, Mark Olfson, Madelyn S. Gould, Katherine M. Keyes, Jeffrey Shaman.

References

1. Centers for Disease Control and Prevention. National Center for Injury Prevention and Control. Web-based Injury Statistics Query and Reporting System (WISQARS)2021. Available from: www.cdc.gov/injury/wisqars/index.html.
2. Hedegaard H, Warner M. Suicide mortality in the United States, 1999-2019. NCHS Data Brief. 2021; 398. https://stacks.cdc.gov/view/cdc/101761. PMID: 33663651
3. Substance_Abuse_Mental_Health_Services_Administration. Key substance use and mental health indicators in the United States: results from the 2019 National Survey on Drug Use and Health. 2020.
4. Franklin JC, Ribeiro JD, Fox KR, Bentley KH, Kleiman EM, Huang X, et al. Risk factors for suicidal thoughts and behaviors: A meta-analysis of 50 years of research. Psychological bulletin. 2017; 143 (2):187. https://doi.org/10.1037/bul0000084 PMID: 27841450
5. Keyes KM, Kandula S, Olfson M, Gould MS, Martínez-Alés G, Rutherford C, et al. Suicide and the agent–host–environment triad: leveraging surveillance sources to inform prevention. Psychological medicine. 2021; 51(4):529–37. https://doi.org/10.1017/S003329172000536X PMID: 33663629
6. Olfson M, Wall M, Wang S, Crystal S, Gerhard T, Blanco C. Suicide following deliberate self-harm. American Journal of Psychiatry. 2017; 174(8):765–74. https://doi.org/10.1176/appi.ajp.2017.16111288 PMID: 28320225
7. Olfson M, Wall M, Wang S, Crystal S, Liu S-M, Gerhard T, et al. Short-term suicide risk after psychiatric hospital discharge. JAMA psychiatry. 2016; 73(11):1119–26. https://doi.org/10.1001/jamapsychiatry.2016.2035 PMID: 27654151
8. Ursano RJ, Kessler RC, Heeringa SG, Cox KL, Naifeh JA, Fullerton CS, et al. Nonfatal suicidal behaviors in US Army administrative records, 2004–2009: Results from the Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS). Psychiatry. 2015; 78(1):1–21.
9. National Research Council. Firearms and violence: a critical review. 2005 0309091241.
10. Rehkopf DH, Buka SL. The association between suicide and the socio-economic characteristics of geographical areas: a systematic review. Psychological medicine. 2006; 36(2):145–57. https://doi.org/10.1017/S003329170500588X PMID: 16420711
11. Martínez-Alés G, Jiang T, Keyes KM, Gradus JL. The recent rise of suicide mortality in the United States. Annual review of public health. 2021; 43. https://doi.org/10.1146/annurev-publhealth-051920-123206 PMID: 34705474

12. Iskander JK, Crosby AE. Implementing the national suicide prevention strategy: Time for action to flatten the curve. Preventive medicine. 2021; 152:106734. https://doi.org/10.1016/j.ypmed.2021.106734 PMID: 34344523

13. Choi D, Sumner SA, Holland KM, Draper J, Murphy S, Bowen DA, et al. Development of a machine learning model using multiple, heterogeneous data sources to estimate weekly US suicide fatalities. JAMA network open. 2020; 3(12):e2030932–e. https://doi.org/10.1001/jamanetworkopen.2020.30932 PMID: 33355678

14. Lee KS, Lee H, Myung W, Song G-Y, Lee K, Kim H, et al. Advanced daily prediction model for national suicide numbers with social media data. Psychiatry investigation. 2018; 15(4):344. https://doi.org/10.30773/pi.2017.10.15 PMID: 29614852

15. Won H-H, Myung W, Song G-Y, Lee W-H, Kim J-W, Carroll BJ, et al. Predicting national suicide numbers with social media data. PloS one. 2013; 8(4):e61809. https://doi.org/10.1371/journal.pone.0061809 PMID: 23630615

16. Gould MS, Lake AM. Suicide Prevention and 988: Effectiveness of the National Suicide Prevention Lifeline. Alexandria, VA: National Association of State Mental Health Program Directors., 2021.

17. Gould MS, Kalafat J, HarrisMunfakh JL, Kleinman M. An evaluation of crisis hotline outcomes. Part 2: Suicidal callers. Suicide and Life-Threatening Behavior. 2007; 37(3):338–52. https://doi.org/10.1521/suli.2007.37.3.338 PMID: 17579545

18. Gould MS, Lake AM, Galfalvy H, Kleinman M, Munfakh JL, Wright J, et al. Follow-up with callers to the National Suicide Prevention Lifeline: Evaluation of callers' perceptions of care. Suicide and Life-Threatening Behavior. 2018; 48(1):75–86. https://doi.org/10.1111/sltb.12339 PMID: 28261860

19. De Choudhury M, Kiciman E, Dredze M, Coppersmith G, Kumar M, editors. Discovering shifts to suicidal ideation from mental health content in social media. Proceedings of the 2016 CHI conference on human factors in computing systems; 2016. https://doi.org/10.1145/2858036.2858207 PMID: 29082385

20. Eichstaedt JC, Smith RJ, Merchant RM, Ungar LH, Crutchley P, Preoțiuc-Pietro D, et al. Facebook language predicts depression in medical records. Proceedings of the National Academy of Sciences. 2018; 115(44):11203–8. https://doi.org/10.1073/pnas.1802331115 PMID: 30322910

21. Jashinsky J, Burton SH, Hanson CL, West J, Giraud-Carrier C, Barnes MD, et al. Tracking suicide risk factors through Twitter in the US. Crisis. 2014. https://doi.org/10.1027/0227-5910/a000234 PMID: 24121153

22. O'dea B, Wan S, Batterham PJ, Calear AL, Paris C, Christensen H. Detecting suicidality on Twitter. Internet Intervention. 2015; 2(2):183–8.

23. Lee J-Y. Search trends preceding increases in suicide: A cross-correlation study of monthly Google search volume and suicide rate using transfer function models. Journal of affective disorders. 2020; 262:155–64. https://doi.org/10.1016/j.jad.2019.11.014 PMID: 31733460

24. National Center for Health Statistics. Detailed Mortality, All counties, 1999-2019 as compiled from data provided by the 57 vital statistics jurisdictions through the Vital Statistics Cooperative Program 2019.

25. World Health Organization. The International Statistical Classification of Diseases and Health Related Problems ICD-10: Tenth Revision. Volume 1: Tabular List: World Health Organization; 2004.

26. National Center for Health Statistics. Bridged-race intercensal estimates of the resident population of the United States for July 1, 2000-July 1, 20092019; (September 1, 2021). Available from: https://www.cdc.gov/nchs/nvss/bridged_race.htm.

27. National Center for Health Statistics. Vintage 2020 Postcensal estimates of resident population of the United States for April 1, 2010 - July 1, 20202020; (September 21, 2021). Available from: https://www.cdc.gov/nchs/nvss/bridged_race/data_documentation.htm.

28. Arendt F, Scherr S. Optimizing online suicide prevention: A search engine-based tailored approach. Health communication. 2017; 32(11):1403–8. https://doi.org/10.1080/10410236.2016.1224451 PMID: 27739876

29. Bruckner TA, McClure C, Kim Y. Google searches for suicide and risk of suicide. Psychiatric services. 2014; 65(2):271–2. https://doi.org/10.1176/appi.ps.201300211 PMID: 24492910

30. Gunn JF III, Lester D. Using google searches on the internet to monitor suicidal behavior. Journal of affective disorders. 2013; 148(2-3):411–2. https://doi.org/10.1016/j.jad.2012.11.004 PMID: 23182592

31. Hagihara A, Miyazaki S, Abe T. Internet suicide searches and the incidence of suicide in young people in Japan. European archives of psychiatry clinical neuroscience. 2012; 262(1):39–46. https://doi.org/10.1007/s00406-011-0212-8 PMID: 21505949

32. Kristoufek L, Moat HS, Preis T. Estimating suicide occurrence statistics using Google Trends. EPJ data science. 2016; 5:1–12. https://doi.org/10.1140/epjds/s13688-016-0094-0 PMID: 32355600

33. Ma-Kellams C, Or F, Baek JH, Kawachi I. Rethinking suicide surveillance: Google search data and self-reported suicidality differentially estimate completed suicide risk. Clinical Psychological Science. 2016; 4(3):480–4.

34. McCarthy MJ. Internet monitoring of suicide risk in the population. Journal of affective disorders. 2010; 122(3):277–9. https://doi.org/10.1016/j.jad.2009.08.015 PMID: 19748681

35. Page A, Chang S-S, Gunnell D. Surveillance of Australian suicidal behaviour using the internet? Australian New Zealand Journal of Psychiatry. 2011; 45(12):1020–2. https://doi.org/10.3109/00048674.2011.623660 PMID: 22034830

36. Recupero PR, Harms SE, Noble JM. Googling suicide: surfing for suicide information on the Internet. The Journal of clinical psychiatry. 2008; 69(6):7856. PMID: 18494533

37. Sueki H. Does the volume of Internet searches using suicide-related search terms influence the suicide death rate: Data from 2004 to 2009 in Japan. Psychiatry clinical neurosciences. 2011; 65(4):392–4. https://doi.org/10.1111/j.1440-1819.2011.02216.x PMID: 21569178

38. Till B, Niederkrotenthaler T. Surfing for suicide methods and help: content analysis of websites retrieved with search engines in Austria and the United States. The Journal of clinical psychiatry. 2014; 75 (8):20534. https://doi.org/10.4088/JCP.13m08861 PMID: 25099284

39. Tran US, Andel R, Niederkrotenthaler T, Till B, Ajdacic-Gross V, Voracek M. Low validity of Google Trends for behavioral forecasting of national suicide rates. PloS One. 2017; 12(8):e0183149. https://doi.org/10.1371/journal.pone.0183149 PMID: 28813490

40. Yang AC, Tsai S-J, Huang NE, Peng C-K. Association of Internet search trends with suicide death in Taipei City, Taiwan, 2004–2009. Journal of affective disorders. 2011; 132(1-2):179–84. https://doi.org/10.1016/j.jad.2011.01.019 PMID: 21371755

41. Culotta A. Lightweight methods to estimate influenza rates and alcohol sales volume from Twitter messages. Language resources evaluation. 2013; 47(1):217–38.

42. Kandula S, Hsu D, Shaman J. Subregional nowcasts of seasonal influenza using search trends. Journal of medical Internet research. 2017; 19(11):e7486. https://doi.org/10.2196/jmir.7486 PMID: 29109069

43. Lampos V, Miller AC, Crossan S, Stefansen C. Advances in nowcasting influenza-like illness rates using search query logs. Scientific reports. 2015; 5(1):1–10. https://doi.org/10.1038/srep12760 PMID: 26234783

44. Ahmad F, Rossen L, Sutton P. Provisional drug overdose death counts. National Center for Health Statistics. Statistics, Centers for Disease Control and Prevention (CDC); 2021.

45. Brockwell PJ, Davis RA. Introduction to time series and forecasting: Springer; 2002.

46. Durbin J, Koopman SJ. Time series analysis by state space methods: OUP Oxford; 2012.

47. Hyndman RJ, Athanasopoulos G. Forecasting: principles and practice: OTexts; 2018.

48. O'Hara-Wild M, Hyndman R, Wang E, Caceres G, Hensel T, Hyndman T. fable: Forecasting models for tidy time series. 2021. 2021.

49. R Core Team: A language environment for statistical computing. R Foundation for Statistical Computing: version 3.5. 0. 2018.

50. Hyndman R, Athanasopoulos G, Bergmeir C, Caceres G, Chhay L, O'Hara-Wild M, et al. forecast: Forecasting functions for time series and linear models; 2020. R package version 8.12.

51. Bentzien S, Friederichs P. Decomposition and graphical portrayal of the quantile score. Quarterly Journal of the Royal Meteorological Society. 2014; 140(683):1924–34.

52. Hollander M, Wolfe DA. Nonparametric statistical methods 1999.

53. Wang SY, Stark A, Ray E, Bosse N, Reich NG, Sherratt K, et al. covidHubUtils: Utility functions for the COVID-19 forecast hub. 0.1.6 ed 2021.

54. Hewamalage H, Bergmeir C, Bandara K. Recurrent neural networks for time series forecasting: Current status and future directions. International Journal of Forecasting. 2021; 37(1):388–427.

55. Smyl S. A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting. International Journal of Forecasting. 2020; 36(1):75–85.

56. Cramer EY, Ray EL, Lopez VK, Bracher J, Brennen A, Castro Rivadeneira AJ, et al. Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States. Proceedings of the National Academy of Sciences. 2022; 119(15):e2113561119. https://doi.org/10.1073/pnas.2113561119 PMID: 35394862

57. Johansson MA, Apfeldorf KM, Dobson S, Devita J, Buczak AL, Baugher B, et al. An open challenge to advance probabilistic forecasting for dengue epidemics. Proceedings of the National Academy of Sciences. 2019; 116(48):24268–74.

58. Reich NG, Brooks LC, Fox SJ, Kandula S, McGowan CJ, Moore E, et al. A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the United States. Proceedings of the National Academy of Sciences. 2019; 116(8):3146–54. https://doi.org/10.1073/pnas.1812594116 PMID: 30647115

59. Reich NG, McGowan CJ, Yamana TK, Tushar A, Ray EL, Osthus D, et al. Accuracy of real-time multimodel ensemble forecasts for seasonal influenza in the US. PLoS computational biology. 2019; 15(11): e1007486.

60. Yamana TK, Kandula S, Shaman J. Superensemble forecasts of dengue outbreaks. Journal of The Royal Society Interface. 2016; 13(123):20160410. https://doi.org/10.1098/rsif.2016.0410 PMID: 27733698

61. Yamana TK, Kandula S, Shaman J. Individual versus superensemble forecasts of seasonal influenza outbreaks in the United States. PLoS computational biology. 2017; 13(11):e1005801. https://doi.org/10.1371/journal.pcbi.1005801 PMID: 29107987

62. Brauer F. Compartmental models in epidemiology. Mathematical epidemiology: Springer; 2008. p. 19–79.

63. Keeling MJ, Rohani P. Modeling infectious diseases in humans and animals. Modeling infectious diseases in humans and animals: Princeton university press; 2011.

64. Kandula S, Martinez-Alés G, Rutherford C, Gimbrone C, Olfson M, Gould MS, et al. County-level estimates of suicide mortality in the USA: a modelling study. Lancet Public Health. 2023 Jan 23:S2468–2667(22)00290-0. https://doi.org/10.1016/S2468-2667(22)00290-0 PMID: 36702142

65. Rockett IR, Wang S, Stack S, De Leo D, Frost JL, Ducatman AM, et al. Race/ethnicity and potential suicide misclassification: window on a minority suicide paradox? BMC psychiatry. 2010; 10(1):1–8. https://doi.org/10.1186/1471-244X-10-35 PMID: 20482844

66. Ehlman DC. Changes in suicide rates—United States, 2019 and 2020. Morbidity Mortality Weekly Report. 2022; 71.