2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Modeling Cases and Deaths per Million using Daily-Aggregated Facebook COVID-19 Symptom Survey Data

Sage J. Betko[1], Rishabh S. Shetty[2], Jeffrey J. Morgan[3], Prahlad G. Menon[4]

Abstract— **We develop a novel analytic approach to modeling future COVID-19 risk using COVID-19 Symptom Survey data aggregated daily by US state, joined with daily time-series data on confirmed cases and deaths. Specifically, we model N-day forward-looking estimates for per-US-state-per-day change in deaths per million (DPM) and cases per million (CPM) using a multivariate regression model to below baseline error (65% and 38% mean absolute percentage error for DPM/CPM, respectively). Additionally, we model future changes in the curvature of CPM/DPM as "increasing" or "decreasing" using a random forest classifier to above 72% accuracy. In sum, we develop and characterize models to establish a relationship between behaviors and beliefs of individuals captured via the Facebook COVID-19 Symptom Surveys and the trajectory of COVID-19 outbreaks evidenced in terms of CPM and DPM. Such information can be helpful in assessing collective risks of infection and death during a pandemic as well as in determining the effectiveness of appropriate risk mitigation strategies based on behaviors evidenced through survey responses.**

I. Introduction

The COVID-19 pandemic has been met with various mitigation strategies but the rollout of these measures as well as planning of effective regional responses to the pandemic is hinged on leveraging data-driven insights into trending risks. Organized by Catalyst@Health 2.0, Facebook Data for Good, the Delphi Group at Carnegie Mellon University (CMU), the Joint Program on Survey Methodology at the University of Maryland (UMD), the Duke Margolis Center for Health Policy, and Resolve to Save Lives, an initiative of Vital Strategies, partnered to conduct the COVID-19 Symptom Data Challenge. Opt-in surveys were conducted on a daily sample of Facebook users starting on April 6, 2020 [1]. The aggregated data adjusts for sample bias for age, gender, location, and other characteristics [2]. The Challenge asked contestants to develop novel analytic approaches to using the Symptom Survey data for earlier outbreak detection and improved situational awareness [3]. We leveraged the dataset from this challenge to develop a novel set of models for future COVID-19 risk in terms of future confirmed cases and deaths, at the daily level, grouped by US state.

II. Methods

In the present study we model N-day forward-looking COVID-19 cases and deaths (with N ranging from 1 to 24) as a function of Facebook survey data (see summary in Fig. 1). We develop two separate sets of models using this approach. The first is concerned with modeling daily new cases and deaths as a continuous response variable, which can be thought of as the slopes of the curves of cumulative cases and deaths. The second modeling approach is concerned with classifying the sign of the change in daily new cases and deaths as positive or negative, corresponding to the convexity of the trends in cases and deaths i.e. a second-order behavior.

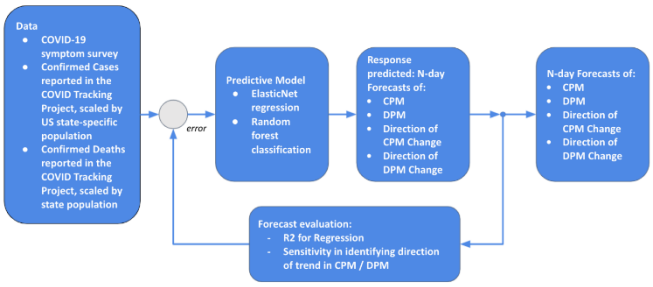

Figure 1. Summary of overall approach adopted in this study.

Facebook symptom survey data is sourced from the CMU Delphi Epidata Symptom Survey which is distributed to approximately 70,000 users daily and contains data from the 6 month period between April 12th and October 21st [3]. It has 50 survey questions / attributes in total, each with weighted and unweighted signal variants, of which we considered weighted responses for this study. Weighted values are preferred since they more closely approximate population parameters by accounting for survey bias and state demographics [4]. Specific features include percentage of respondents with self-reported symptoms of COVID-like illness (e.g., fever, with cough, shortness of breath, or anosmia ageusia) and influenza-like illness (e.g., fever, with cough, or sore throat). We average the responses to each question across age and gender demographics. Fifty states and the District of Columbia are represented in the dataset. Each geographic zone has 192 days of data available. There were a total of 9,792 samples across all geographical territories.

Targets (cumulative cases and deaths per-state) are sourced from covidtracking.com and scaled by state population to produce cases per million (CPM) and deaths per million (DPM). State populations are sourced from the US Census Bureau's 2019 estimates [5]. CPM and DPM were de-

*Research originated as part of the COVID-19 Symptom Data Challenge

[1] Undergraduate Student in Statistics and Machine Learning at Carnegie Mellon University, Pittsburgh, PA 15213 USA (email: sbetko@cmu.edu).
[2] Graduate Student in Actuarial Science at Columbia University, New York, NY 10027 USA (email: rss2226@columbia.edu).
[3] With iPower, LLC and a Graduate Student in Biomedical Engineering at the Catholic University of America, Washington, DC 20064 USA (phone 540-845-7249; email: 19morgan@cua.edu).
[4] Bioengineering Professor with the University of Pittsburgh (email: prm44@pitt.edu).

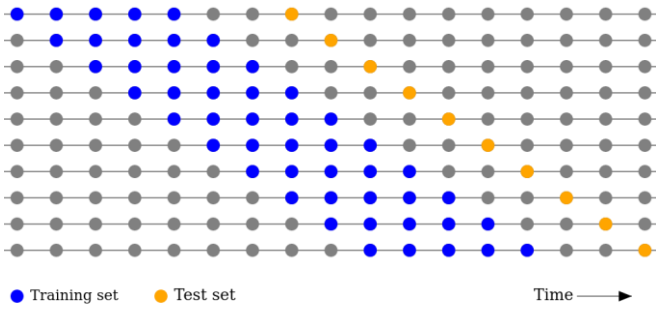

Figure 2. Example of a time-series cross-validation with a rolling forecast origin on a fixed size window for N-day forward looking response validation. In this example, N=3. Gray units indicate dates not included in a training or test set.

trended via first-order differencing with a lag of one day, to produce daily new CPM and DPM. However, standard instance-based or tree-based methods such as the *k*-nearest neighbors algorithm for regression or decision tree algorithms cannot produce target values outside of the range of their training set. Therefore, these models struggle with forecasting "waves" of unseen CPM or DPM magnitudes in certain states, even if they are capable of learning appropriate predictor-response relationships within the low magnitude regime of their training set. Hence, we used a popular implementation of a regularized linear model (ElasticNet) available in Scikit-Learn, which is capable of extrapolation beyond the training set, to predict future new CPM and DPM [6]. Additionally, we impose a strong prior on the output of the linear model and all forecasts are evaluated as $\hat{y}_t = \max(f(\boldsymbol{x}_t), 0)$, where f is the learned function and $\boldsymbol{x}_t$ is the feature vector corresponding to time t. The latter ensures that's forecasts never emerge as negative numbers.

To create the response variable for our second model to classify the sign of the change in daily new cases and deaths as positive or negative, categorical response variables were created via second-order differencing for the same for CPM and DPM respectively. A positive change in new cases and deaths on a per-state-per-day basis corresponds to the "increasing" label, and a negative value to "decreasing" label. We train a random forest classifier to model these categorical response variables, separately for CPM and DPM and on a per-state basis.

In the 2020 COVID-19 Symptom Data Challenge, we originally reported performance metrics over a single holdout dataset. This strategy for assessing forecasting performance privileges smaller choices of holdout set size as larger holdout sets contain larger forecasting horizons (i.e., the duration between the last training sample and subsequent test samples). However, smaller holdout sets yield more variance in the accuracy measure, making a single train/test split insufficient to capture estimator performance. To encourage more stable results, and to simulate a real-world deployment scenario of a forecasting engine, in this study we make use of a time series cross-validation scheme as described in Athanasopoulos [7] (see Fig. 2). This procedure involves a series of test sets, each consisting of a single observation for each state. Each test has a corresponding training set consisting only of prior observations. Observations from the first 30 days are not included in any test set since training sets smaller than 30 days produce unreliable predictions. This procedure is commonly

Table 1. Frequency table for second order change in CPM and DPM per-state, per-day (i.e., the change in the *daily new* CPM or DPM, reported in each state).

	CPM	DPM
Increasing	6117	6616
Decreasing	6032	5533

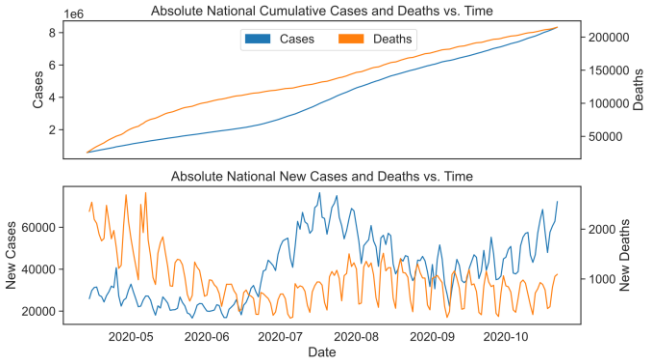

Figure 3. Absolute national cumulative cases and deaths (TOP); and absolute national daily new cases and deaths (BOTTOM), over time in our dataset.

referred to as evaluation on a rolling forecasting origin, as the "origin" at which the forecast is based moves forward in time.

One-step forecasts of CPM/DPM may not be as relevant as multi-step forecasts of several days or weeks. In this case, the cross-validation procedure based on a rolling forecasting origin is modified to allow errors from multi-day horizons to be used. Fig. 2 illustrates the series of training and test sets for a specific case of N=3 days ahead for forecasts, where blue observations form the training sets, and orange observations form the test sets.

The forecast accuracy is computed by averaging the mean absolute percentage error (MAPE) over the test sets, as defined in Equation 1 where y_t is the actual value at time t, and $\hat{y}_t$ is the forecasted value at time t. All models, including baselines, were trained and evaluated individually for each state. Hence, the final reported MAPE is the mean MAPE across all states.

$$M = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{y_t - \hat{y}_t}{y_t}\right| \quad (1)$$

Since the MAPE is a measure of relative error, this measure is conducive to comparing regression performance across states with varying levels of CPM/DPM. When evaluating regression accuracy across states, the same absolute error for a state experiencing a high level of CPM/DPM contributes less error to the final metric than for a state with comparatively lower CPM/DPM.

III. Results

Cumulative CPM and DPM were differenced with successive days on a per-state basis to produce per-state-per-day change in cases and deaths. This raw data is summarized in Fig. 3. The distribution of change in daily new CPM and DPM per-state-per-day, counting days where there was an increase in daily new CPM/DPM as opposed to those that

Table 2. Mean absolute percentage error (MAPE) in N-Day forward-looking forecasts, aggregated across states, per day, for N ranging from 1 to 24 days in future. N's corresponding to minimum MAPE, for each predicted response, are highlighted in grey.

N-Ahead	CPM		DPM	
	ElasticNet	*Baseline*	*ElasticNet*	*Baseline*
1	0.89	1.08	1.09	1.04
2	0.94	1.09	1.10	1.05
3	0.83	1.10	1.34	1.06
4	1.04	1.10	1.19	1.06
5	1.15	1.11	2.41	1.06
6	0.98	1.11	1.83	1.08
7	2.01	1.12	1.08	1.08
8	1.44	1.13	1.03	1.08
9	1.20	1.15	1.24	1.10
10	0.71	1.16	1.21	1.11
11	0.87	1.16	1.09	1.12
12	0.63	1.17	1.97	1.12
13	0.79	1.18	1.94	1.13
14	1.23	1.19	1.05	1.13
15	0.57	1.21	1.52	1.13
16	0.67	1.23	1.82	1.16
17	0.72	1.24	1.74	1.18
18	0.61	1.24	0.87	1.18
19	0.42	1.25	**0.65**	1.19
20	**0.38**	1.27	0.84	1.20
21	0.57	1.29	3.50	1.21
22	0.79	1.31	1.21	1.23
23	0.88	1.34	1.10	1.24
24	0.68	1.36	1.17	1.24

experienced a decrease in daily new CPM/DPM, is summarized in Table 1.

In Table 2, we measure the mean absolute percentage error (MAPE) for CPM change and DPM change modeled N (1 to 24) days in advance, where the baseline model predicts the mean response value seen in the training set on a per-state basis. Confusion matrices for N=N0 days in advance are presented in Table 3, where N0 corresponds to MAPE = $MAPE_{min}$ from Table 2. Classification of change in daily new DPM with a 19-day forecasting horizon had an average of 74% accuracy in detecting an increasing event and 75% accuracy in detecting a decreasing event. Modeling of change in daily new CPM with a 20-day forecasting horizon had an average of 72% accuracy in detecting an increasing or a decreasing event.

Figs. 4 and 5 are N=20 and N=19 days day future prediction (CPM change and DPM change) run continuously out-of-sample on a moving window basis with comparison with actuals. Integrated versions of this for CPM and DPM time-series comparison averaged per-state-per-day are shown in Fig. 4. The predicted and actual daily new CPM and DPM for the top 4 states with leading case-counts (California, New York, Texas, and Florida) are shown in Fig. 5.

Table 3. Confusion matrices for random forest classification of 20-day advance and 19-day advance second-order change for CPM and DPM, respectively.

	CPM		DPM	
Increasing (Actual)	4540	2543	5864	1996
Decreasing (Actual)	2321	4796	2592	3748
	Increasing (Predicted)	Decreasing (Predicted)	Increasing (Predicted)	Decreasing (Predicted)

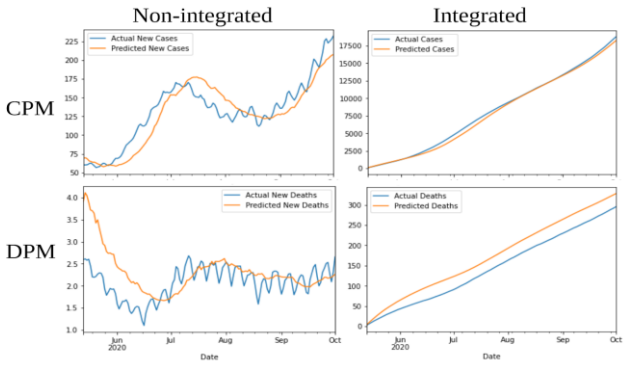

Figure 4. Integrated (RIGHT) and non-integrated (LEFT), predicted and actual CPM (TOP) DPM (BOTTOM) vs. Time, averaged per-state-per-day.

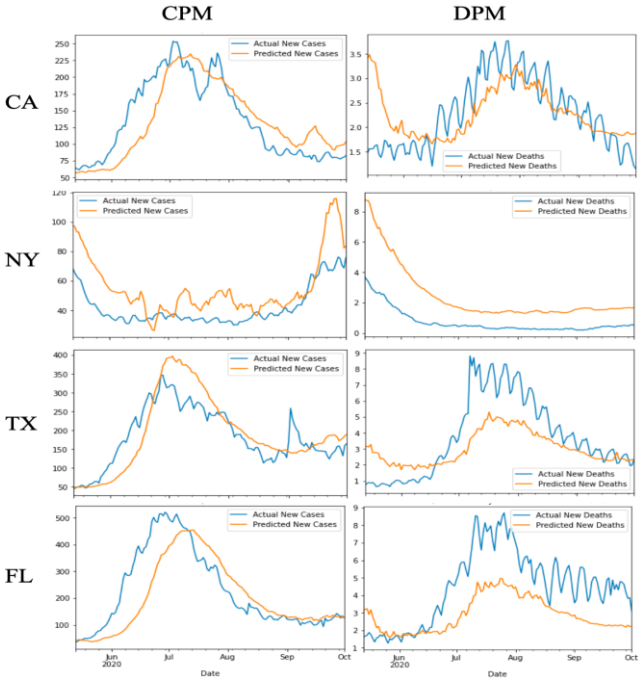

Figure 5. 20-day future prediction of daily new CPM (LEFT) and 19-day future prediction of DPM (RIGHT) run continuously (i.e., with a 20-day and 19-day forecasting horizon) in case-count leading states.

IV. Discussion

Of the many decisions that may affect the spread of COVID-19, both on an individual level and on a public health level, uncontrolled spread is the likely outcome if people do not appreciably change their behavior or vaccination status. Critical decisions regarding the effectiveness of appropriate risk mitigation strategies, or simply the assessment of individual risks of infection or death are better informed by prediction models that provide insight. Models built on symptom surveys offer the ability to incorporate surveillance data from individual respondents that collectively represent individuals in varying stages of infection, even as early as before they go to a drug store, testing facility, or health care provider. Further, Facebook makes these surveys available on a daily basis, providing a unique opportunity to utilize these survey data as a proxy for other reactive indicators such as the outcomes of laboratory test results, provided a suitable mathematical model is built, validated and adopted in practice. Public health officials could act on the information

from forecasts to make rules or recommendations about the extent of opening schools and businesses.

Verelst et al 2016 conducted a systematic review of Behavioural Change Models (BCMs) from 2010-2015 for infectious disease transmission and observed that most BCMs are purely theoretical and were constructed independent of empirical observations [8]. BCMs have often used mechanistic models to explore the effect of behavioral changes on outbreak trajectory by modifying the Kermack-McKendrick susceptible-exposed-infected recovered (SEIR) model or allowing the transmission parameters to change based on changes in contact rates [8,9,10]. Researchers have reported the effect on COVID-19 transmission dynamics in Korea and the Daegu/Gyeongbuk area resulting from reduced transmission rates due to behavioral changes in a portion of the individuals as they become more fearful of the disease when incident rates grow higher [9]. Modelers have also reported the effect of the timing and duration of policy changes on the COVID-19 Growth Rate in various locations [11]. This paper differs from these studies in that it incorporates estimates of the extent of behavioral changes from surveys of millions of people viz. survey respondents.

While our approach to capturing behavior in models for COVID19 cases and deaths is novel, there is scope for adding more attributes of behaviors evidenced from data beyond survey responses. For instance, Google produces Community Mobility Reports to allow public health officials to capture changes in movement trends across different places where community members may visit (e.g., transit stations, grocery and pharmacy stores, and parks) [12]. Future work could incorporate Community Mobility Reports in our models.

Being able to predict whether the number of cases will increase or decrease had important implications under the "Opening Up America Again" proposed by the Trump Administration One of the proposed criteria to be satisfied referred to the number of cases, requiring either a "downward trajectory of documented cases within a 14-day period" or a "downward trajectory of positive tests as a percent of total tests within a 14-day period (flat or increasing volume of tests)." [13] Models such as the ones described in this manuscript will be applicable to such strategies. Further, our data evidences that the response time between behaviors of individuals and peaks in cases or deaths evidenced in our reactively collected data is between 19 and 20 days.

V. Conclusion

The Facebook Symptom Survey data has allowed us to validate the feasibility of accurately forecasting DPM and CPM. This project has focused on making predictions and visualizations for population-normalized cases and deaths in US states. Our methodology could be extended to other geographic regions where Facebook has collected surveys. Results are presented for nowcasting and N-day-ahead forecasting of cases and deaths by US state and additionally on a county level, for greater granularity of reporting. While our manuscript defines a framework for modeling normalized cases and deaths as a time-series, the model will optimally require to be refitted on new data until the present date of any given forecast for optimal performance.

Acknowledgment

The authors thank the organizers of the COVID-19 Symptom Data Challenge and Anne Gibbon for providing visualization products for our submission in the Challenge. J. J. M thanks Dr. Otto Wilson and Dr. Binh Tran for guidance in infectious disease modeling.

References

1. The COVID-19 Symptom Data Challenge. https://www.symptomchallenge.org/
2. The COVID-19 Symptom Data Challenge/Challenge Background. https://www.symptomchallenge.org/background
3. The COVID Tracking Project from The Atlantic, https://covidtracking.com/
4. CMU Delphi Group. The CMU Delphi Epidata API. https://cmu-delphi.github.io/delphi-epidata/api/covidcast-signals/fb-survey
5. US Census Bureau. (2019, January 1). National State Estimates. US Census Bureau. https://www.census.gov/data/datasets/time-series/demo/popest/2010s-state-total.html
6. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. The Journal of Machine Learning Research, 12, 2825-2830
7. Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: Principles and Practice. OTexts.
8. Verelst, F., Willem, L., & Beutels, P. (2016). Behavioural change models for infectious disease transmission: a systematic review (2010–2015). Journal of The Royal Society Interface, 13(125), 20160820.
9. Kim, S., Seo, Y. B., & Jung, E. (2020). Prediction of COVID-19 transmission dynamics using a mathematical model considering behavior changes in Korea. Epidemiology and health, 42, e2020026. https://doi.org/10.4178/epih.e2020026.
10. Acuña-Zegarra, M. A., Santana-Cibrian, M., & Velasco-Hernandez, J. X. (2020). Modeling behavioral change and COVID-19 containment in Mexico: A trade-off between lockdown and compliance. Mathematical biosciences, 325, 108370 https://doi.org/10.1016/j.mbs.2020.108370
11. Courtemanche, C., Garuccio, J., Le, A., Pinkston, J., & Yelowitz, A. (2020). Strong Social Distancing Measures In The United States Reduced The COVID-19 Growth Rate. Health affairs (Project Hope), 39(7), 1237–1246. https://doi.org/10.1377/hlthaff.2020.00608
12. COVID-19 Community Mobility Reports. https://www.google.com/covid19/mobility/
13. The White House, Centers for Disease Control and Prevention. Guidelines Opening America Again. https://trumpwhitehouse.archives.gov/openingamerica/