Appendix to: DeepCOVID: An Operational Deep Learning-driven Framework for Explainable Real-time COVID-19 Forecasting

Alexander Rodríguez[*], Anika Tabassum[*†], Jiaming Cui[*], Jiajia Xie[*], Javen Ho[*], Pulak Agarwal[*], Bijaya Adhikari[‡], B. Aditya Prakash[*]

[*]College of Computing, Georgia Institute of Technology, Atlanta, USA
[†]Department of Computer Science, Virginia Tech, Arlington, USA
[‡]Department of Computer Science, University of Iowa, Iowa City, USA
{arodriguezc, jiamingcui1997, jxie, jho67, pulak.agarwal, badityap}@gatech.edu, anikat1@vt.edu, bijaya-adhikari@uiowa.edu

A1: Data Module (More Details)

Detailed Description of Signals (C2)

The extended summary of the data signals used in DEEPCOVID can be found in Table 1. Next, we give a brief description about what each type of signal represents, its time and spatial granularity, and rationale behind using such signals.

1. *Line list* signals (DS1): These signals are derived from records that report who, when and where facets of an infected person; thus, they are directly related to the disease spread. This type includes the number of persons infected, hospitalized, and recovered (or deceased). Others report % of visits to the emergency room (ER) due to influenza-like illness (ILI) symptoms and COVID-like illness (CLI). Signal 8 represents the difference between the observed deaths (overall, not only COVID-related) from the expected deaths during specific time periods.
2. *Testing* signals (DS2): This type of signals reflect social policy (local leader's efforts for escalating testing) and social behavior, e.g. people with fever caused by any other disease may request to be tested for COVID. Here, we have signals that report the number of tests (total and negative), number of emergency facilities and health care providers reporting data in *DS1*. The latter two are measures of the perceived importance of the disease by health care providers, as it has been previously observed in influenza forecasting (Brooks et al. 2015).
3. *Crowdsourced symptomatic* signals (DS3): This signal is collected from individuals using Kinsa digital thermometers at home who present fever and influenza-like symptoms, which have a significant overlap with COVID-like symptoms. This signal is potentially an indirect measure of both reported and unreported COVID cases.
4. *Mobility* signals (DS4): These signals are collected from the record of people visits to points of interest (POI) in different regions. Google collects the daily change of visits in multiple categories of POI compared with the period January 3-February 6, 2020 (signals 14-20 in Table 1). We also collected a daily change of visits from Apple (signal 22), which indicates the relative volume of direction requests in the geographic map compared to January 13 across different US states. Mobility signals implicitly show the impact of different non-pharmaceutical interventions (NPI) and change of policies adopted by different states.
5. *Exposure based* signals (DS6): The signals measure social contacts in different groups of individuals. The records have been collected from tracking the overlapping location of distinct smartphone devices in commercial venues. They estimate the signals considering standard sample (signal 21) and removing biased samples (signal 22), i.e., removing the devices from samples which have less movement due to stay-at-home order. The signals implicitly indicate the impact of social contacts and NPI interventions on COVID-19 cases.
6. *Social Media* signals (DS6): Facebook collects the daily percentage of people with covid-like-illness (CLI%) and influenza-like-illness (ILI%)(see signals 23-24 in Table 1) across national level and different states of US. They estimate this percentage from the records of voluntary surveys based on disease symptoms.

Addressing missing values in incident hospitalizations (C5)

While collecting incident hospitalizations for forecasting target *T2*, we observed that daily records of 11 states (CA, DC, TX, IL, LA, PA, MI, MO, NC, NV, DE) are missing. Instead, there are other signals such as the current records of hospitalizations and ICU patients. However, assuming current hospitalizations as target *T2* would affect our forecast by overestimating incidences. Hence, our goal was to find a way to estimate *T2* from the current hospitalizations for these 11 states. We tried a compartmental model based on recovered cases and deaths to address our goal. However, we found this challenging to do with the available data because recovered cases are not only considering people that was hospitalized, and the number of deaths is being reported with a different delay than hospitalizations (because they come from different sources). Therefore, as we could not include recoveries nor

Type of signal	Description	Signals	Rationale
(DS1) *Line list*	Directly related to the disease spread (derived from records of who, when, and where a person got infected) (COVID-Tracking 2020; JHU 2020)	1. Confirmed cases; 2. UCI beds currently occupied; 3. People on ventilation; 4. Recovered; 5. Hospitalization rate (COVID-Net); 6. ILI% ER visits; 7. CLI% ER visits; 8. Excess Deaths;	Traditional surveillance for tracking patients and symptoms (CDC 2020)
(DS2) *Testing*	Related to social policy and behavioral considerations (COVID-Tracking 2020)	9. People tested; 10. Negative cases; 11. Emergency facilities reporting; 12. Number of providers;	To capture changing screening and diagnostic artifacts on surveillance
(DS3) *Crowdsourced symptomatic*	Collected from individuals using Kinsa digital thermometers at home (Miller et al. 2018)	13. Digital thermometer readings provide ILI%;	Syndromic symptomatic surveillance
(DS4) *Mobility*	Indicate people visits in multiple POIs (Google 2020) (Apple 2020)	14. Retail and recreation; 15. Grocery and pharmacy; 16. Parks; 17. Transit stations; 18. Residential 19. Workplaces; 20. Overall-region-based	Evidence of changing spatio-temporal contact patterns due to non pharma. interventions and behavior changes
(DS5) *Exposure based*	Collected from tracking overlapping location of distinct smartphones in commercial venues (Couture et al. 2020).	21-22. Device exposures (normal & adjusted);	Measure social contacts and direct potential exposures
(DS6) *Social Surveys*	Facebook symptomatic survey (CMU-Delphi 2020)	23. CLI%; 24. ILI%	Measuring related symptomatic burden

Table 1: Overview of data signals used in DEEPCOVID. (ILI=Influenza like Illness; CLI=COVID like Illness)

deaths, we assumed that after a week only a fraction β of the current hospitalizations at time $t-7$ would stay in the hospital. Thus, we used

$$hosp_{inc}(t) = hosp_{cur}(t) - \beta \cdot hosp_{cur}(t-7)$$

where t is time in days, $hosp_{inc}$ are the incident hospitalizations, and $hosp_{cur}$ are the current hospitalizations. We used grid-search and found $\beta = 0.5$ closely matches with the ground-truth incident hospitalizations of the states that do not have missing values.

A2: Prediction Module (More Details)

Hyperparameter tuning. At the beginning of the pandemic, as data was very scarce, architecture and hyperparameter search were guided by domain expertise on how the epidemic curve should evolve. When more data was available, we were able to create a validation set to tune these parameters.

Training data. Details of our setup (size of data, number of observations) can be inferred from the information described in the paper (number of signals in Table 1, target temporal granularity in Section 3). We designed our method to work throughout the evolution of the pandemic and we use the entire time series. As the pandemic progresses, we get more instances to train and we do notice the performance improves.

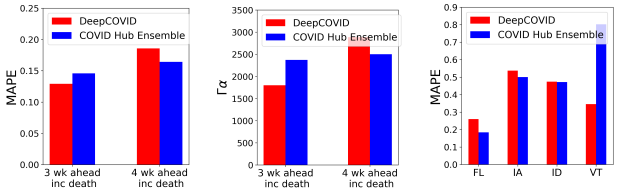
A3: Additional Empirical Results

In this section, we present complementary results and one more observation.

In the main paper, we wanted to emphasize short-term forecasting skill of DEEPCOVID, thus, we showed 1- and 2-weeks ahead performance separately. In Fig. 1(a-b) of this appendix, we present 3- and 4- weeks results showcasing that our longer-term performance is not compromised. In US National, we are able to outperform the official ensemble in 3-wk ahead in both point estimate performance (MAPE) and probabilistic performance ($\Gamma\alpha$ with $\alpha = 0.7$). Further evidence is presented in Fig. 1(c), where 1-4 week ahead performance in other states are consistent with the featured ones in the main paper.

Observation 6 DEEPCOVID *helps to determine change in importance of signals as the spread of the disease progresses.*

With our explainability module, we can communicate temporal insights about the importance of signals (goal **G4**).



(a) 3-4 wk ahead (b) 3-4 wk ahead (c) 1-4 wk ahead

Figure 1: (a-b) DEEPCOVID is close to the official ensemble in 3-4 wk ahead performance in US National, measured with MAPE and probabilistic metric $\Gamma\alpha$ with $\alpha = 0.7$. (c) More examples that our focus on short-term predictions does not compromise longer-term (1-4 wk ahead) performance.

For instance, we analyzed mobility in predictions for US National and California, two regions where mobility was found to have a positive contribution (see Table 2 in paper). We considered two periods: (P1) May to June, when stay-at-home orders were lifted in most states, businesses reopened, and mobility signals increased; (P2) July to August, a period when most mobility signals have already stabilized. We noticed a high contribution of mobility during (P1) in both US National and California, and low or non-existing in (P2). This observation suggests that data driven models need to be regularly updated to reflect the changing dynamics of the disease spread.

A3: Data Revisions

Several of our data sources, especially disease surveillance data from health agencies, report an initial value that undergoes several rounds of revisions to reach a stable value, a process which typically lasts several weeks. As also noted by (Reich et al. 2019), our experiments suggest that these data revisions have an impact into real-time forecasting performance.

To illustrate this issue, we computed the revision error $|v_w - v_s|/v_s$, where v_w is the value at revision week w ($w = 0$ denotes first release) and v_s is the stable value. When we average this error over several past observations, we get a time series over revision weeks, which allows us to see how many revision weeks on average it takes for a signal to stabilize. In Fig. 2 of this appendix, we can notice that even our ground truth target, reported incident death (JHU),

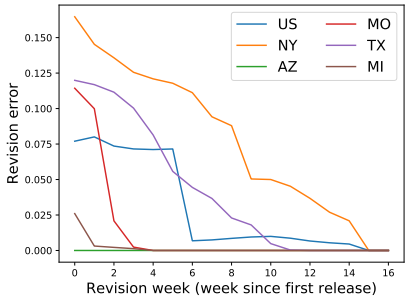


Figure 2: Average revision error for ground truth reported deaths from JHU (our forecasting target). We can see there are a few clusters of geographical regions. Arizona (AZ) and Michigan (MI) have none to minor data revisions problems, which are resolved promptly. Montana (MO) has large revision error, but it stabilizes rapidly. US National and Texas (TX), and New York have large revision error and are slow in stabilizing, taking up to 15 weeks.

exhibit this issue, which implies measures of performance may be unreliable till data stabilizes.

References

Apple. 2020. COVID-19 - Mobility Trends Reports. URL https://www.apple.com/covid19/mobility.

Brooks, L. C.; et al. 2015. Flexible modeling of epidemics with an empirical Bayes framework. *PLoS computational biology* 11(8): e1004382.

CDC. 2020. *Coronavirus Disease 2019 (COVID-19)*. URL https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/.

CMU-Delphi. 2020. COVID-19 Symptom Survey (Delphi Epidata API). URL https://cmu-delphi.github.io/delphi-epidata/.

Couture, V.; et al. 2020. Measuring movement and social contact with smartphone data: a real-time application to COVID-19. Technical report, National Bureau of Economic Research.

COVID-Tracking. 2020. The COVID Tracking Project. URL https://covidtracking.com.

Google. 2020. COVID-19 Community Mobility Report. URL https://www.google.com/covid19/mobility?hl=en.

JHU. 2020. JHU CSSE COVID-19 Dashboard. URL https://coronavirus.jhu.edu/map.html.

Miller, A. C.; et al. 2018. A smartphone-driven thermometer application for real-time population-and individual-level influenza surveillance. *Clinical Infectious Diseases* 67(3): 388–397.

Reich, N. G.; et al. 2019. A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the United States. *Proceedings of the National Academy of Sciences* 201812594.