CURATING A COVID-19 DATA REPOSITORY AND FORECASTING COUNTY-LEVEL DEATH COUNTS IN THE UNITED STATES

Nick Altieri[1], Rebecca L Barter[1], James Duncan[4], Raaz Dwivedi[2], Karl Kumbier[6],
Xiao Li[1], Robert Netzorg[2], Briton Park[1], Chandan Singh[2, *], Yan Shuo Tan[1],
Tiffany Tang[1], Yu Wang[1], Chao Zhang[3], Bin Yu[1, 2, 4, 5, 7, *]

[1] Department of Statistics, [2] Department of EECS, [3] Department of IEOR
[4] Division of Biostatistics, [5] Center for Computational Biology
University of California, Berkeley

[6] Department of Pharmaceutical Chemistry
University of California, San Francisco

[7] Chan Zuckerberg Biohub, San Francisco

ABSTRACT. As the COVID-19 outbreak evolves, accurate forecasting continues to play an extremely important role in informing policy decisions. In this paper, we present our continuous curation of a large data repository containing COVID-19 information from a range of sources. We use this data to develop predictions and corresponding prediction intervals for the short-term trajectory of COVID-19 cumulative death counts at the county-level in the United States up to two weeks ahead. Using data from January 22 to June 20, 2020, we develop and combine multiple forecasts using ensembling techniques, resulting in an ensemble we refer to as Combined Linear and Exponential Predictors (CLEP). Our individual predictors include county-specific exponential and linear predictors, a shared exponential predictor that pools data together across counties, an expanded shared exponential predictor that uses data from neighboring counties, and a demographics-based shared exponential predictor. We use prediction errors from the past five days to assess the uncertainty of our death predictions, resulting in generally-applicable prediction intervals, Maximum (absolute) Error Prediction Intervals (MEPI). MEPI achieves a coverage rate of more than 94% when averaged across counties for predicting cumulative recorded death counts two weeks in the future. Our forecasts are currently being used by the nonprofit organization, Response4Life, to determine the medical supply need for individual hospitals and have directly contributed to the distribution of medical supplies across the country. We hope that our forecasts and data repository at https://covidseverity.com can help guide necessary county-specific decision-making and help counties prepare for their continued fight against COVID-19.

Keywords: COVID-19, data-repository, time-series forecasting, ensemble methods, prediction intervals

Authors ordered alphabetically. Corresponding authors' emails: *{chandan_singh, binyu}@berkeley.edu

Media Summary

Accurate short-term forecasts for COVID-19 fatalities (e.g., over the next two weeks) are critical for making immediate policy decisions such as whether or not counties should re-open. This paper presents: (i) a large publicly available data repository that continuously scrapes, combines, and updates data from a range of different public sources, and (ii) a predictive algorithm CLEP along with a prediction interval MEPI, for forecasting short-term county-level COVID-19 mortality in the US. By combining different trends in the death count data, our county-level CLEP forecasts for cumulative deaths due to COVID-19 are accurate for 7-day into the future and decent for 14-day into the future. The MEPI prediction intervals exhibit high coverage for both 7-day and 14-day forecasts. Our approach was the first to develop forecasts for individual counties (rather than for entire countries or states). Our predictions, along with data and code, are open-source at https://covidseverity.com. They are currently being used by the non-profit organization, Response4Life, to determine the medical supply need for individual hospitals, and have directly contributed to the distribution of medical supplies across the country.

Contents

List of Figures

List of Tables

1. Introduction

In recent months, the COVID-19 pandemic has dramatically changed the shape of our global society and economy to an extent modern civilization has never experienced. Unfortunately, the vast majority of countries, the United States included, were thoroughly unprepared for the situation we now find ourselves in. There are currently many new efforts aimed at understanding and managing this evolving global pandemic. This paper, together with the data we have collated (and are collating continuously), represents one such effort.

Our goals are to provide access to a large data repository combining data from a range of different sources and to forecast short-term (up to two weeks) COVID-19 mortality at the county level in the United States. We also provide uncertainty assessments of our forecasts in the form of prediction intervals based on conformal inference [48].

Predicting the short-term impact (e.g., over the next week) of the virus in terms of the number of deaths is critical for many reasons. Not only can it help elucidate the overall impacts of the virus, but it can also help guide difficult policy decisions, such as whether or not to impose or ease lock-downs (or whether to re-open). While many other studies focus on predicting the long-term trajectory of COVID-19, these approaches are currently difficult to verify due to a lack of long-term COVID-19 data. On the other hand, predictions for immediate short-term trajectories are much easier to verify and are likely to be much more accurate than long-term forecasts due to comparatively fewer uncertainties involved, e.g., due to policy changes, or behavioral changes in the society. Short-term predictions are also necessary for PPE distribution planning and policy decisions such as safe re-opening of the counties and states. So far, a vast majority of predictive efforts have focused on modeling COVID-19 case-counts or death-counts at the national or state-level [18], rather than the more fine-grained county-level that we consider in this paper. To the best of our knowledge, ours was the first work on county-level forecasts.[1]

The predictions we produce in this paper focus on recorded cumulative death counts, rather than recorded cases since recorded cases fail to accurately capture the true prevalence of the virus due to previously limited testing availability. Moreover, comparing different counties based on *the number of* recorded cases is difficult since some counties have performed many more tests than others: the number of positive tests does not equal the number of actual cases. While the *proportion of positive tests* is more comparable across different counties, our modeling approach focuses on recorded death counts rather than proportions. The original motivation to predict death counts was to provide

---

[1]By the time of our first submission to arXiv on May 16, 2020, we were not aware of any concurrent work on county-level forecasts. See Section 6 for discussion on recent work [12] on county-level forecasts (time stamp of June 8) which we became aware of in mid-June while revising the manuscript.

a proxy for the severe case counts, where individuals would need intense care in a hospital (see Section 7 for further discussion). We note that the recorded death count is also likely to be an under-count of the number of true COVID-19 deaths (since it seems as though in many cases only deaths occurring in hospitals are being counted).[2] However, more recently, several efforts are being made to obtain better recorded (not using any algorithms or models) COVID-19 death counts, e.g., by including probable deaths and deaths occurring at home.[3] Nonetheless, the recorded death count is generally believed to be more reliable than the recorded case count.[4]

In Section 2, we introduce our data repository and summarize the data sources contained within. This data repository is being updated continuously (as of July 2020) and includes a wide variety of COVID-19 related information in addition to the county-level case-counts and death-counts (see Tables 1–4 for an overview). Given the rapidly evolving and dynamic nature of COVID-19, several biases arise in the COVID-19 infection data. We provide a detailed discussion on these biases in the context of our forecasts in Section 2.2.

In Section 3, we introduce our predictive approach, wherein we fit a range of different exponential and linear predictor models using our curated data. Each predictor captures a different aspect of the behaviors exhibited by COVID-19, both spatially and temporally, i.e., across regions and time. The predictions generated by the different methods are combined using an ensembling technique by Schueller et al. [39], which we refer to as Combined Linear and Exponential Predictors (CLEP). Additional predictive approaches, including those using social distancing information, are presented in Appendix A (even though they do not outperform the CLEP predictors in the main text based on the COVID-19 case and death counts in the past and neighboring counties).

In Section 4, we develop uncertainty estimates for our predictors in the form of prediction intervals, which we call Maximum (absolute) Error Prediction Intervals (MEPI). The ideas behind these intervals come from conformal inference [48] where the prediction interval coverage is well defined by observing the empirical proportion of time that the observed (cumulative) death counts fall into the prediction intervals over a period of time. Moreover, their guarantees rely on an exchangeability property of the prediction errors in the past several days, which we also examine in the context our prediction tasks.

Section 5 details the evaluation of the predictors and the prediction intervals for the 3, 5, 7, and 14-days-ahead forecasts. We use the data from January 22, 2020 (the day of the first COVID-19 death in the US[5]) and report the prediction performance over the period March 22, 2020 to June 20, 2020. Overall, we find that CLEP predictions are adaptive to the exponential and sub-exponential nature of COVID-19 outbreak (with about 15% error for 7-day-ahead and 30% error for 14-day-ahead predictions; e.g., see Table 6). We also provide detailed results for our prediction intervals MEPI from April 11, 2020 to June 20, 2020. And, we observe that MEPIs are reasonably narrow

---

[2]https://www.nytimes.com/interactive/2020/04/28/us/coronavirus-death-toll-total.html

[3]https://covidtracking.com/blog/confirmed-and-probable-covid-19-deaths-counted-two-ways

[4]https://coronavirus.jhu.edu/data/cumulative-cases

[5]https://www.cdc.gov/mmwr/volumes/69/wr/mm6924e2.htm
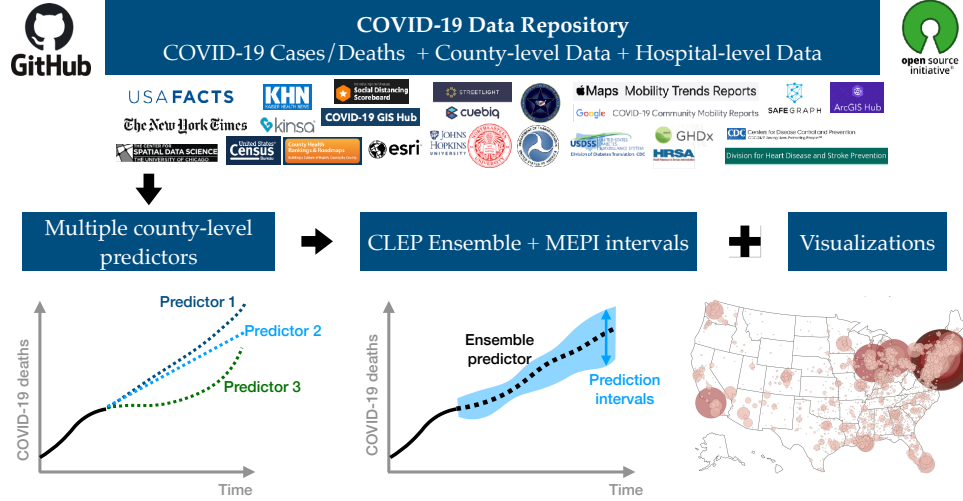
Figure 1. An overview of the paper. We curate an extensive data repository combining data from multiple data sources. We then build several predictors for county-level predictions of cumulative COVID-19 death counts, and develop an ensembling procedure (CLEP) and a prediction interval scheme (MEPI) for these predictions. Both CLEP and MEPI are generic machine learning methods and can be of independent interest (see Sections 3.6 and 4.1 respectively). All the data, and predictions are publicly available at GitHub repo (link in footnote). Visualizations are available at https://covidseverity.com/ and https://geodacenter.github.io/covid/map.html, in collaboration with the Center for Spatial Data Science at the University of Chicago.

and cover the recorded number of deaths for more than 90% of days for most of the counties in the US (e.g., see Figures 11 and 12).

Finally, we describe related work by other authors in Section 6, discuss the impact of our work in distributing medical supplies across the country in Section 7, and conclude in Section 8.

Making both data and the predictive algorithms used in this paper accessible to others is key to ensuring the usefulness of these resources. Thus the data, code, and predictors we discuss in this paper are open-source on GitHub (https://github.com/Yu-Group/covid19-severity-prediction) and are also updated daily with several visualizations at https://covidseverity.com. The results in this paper contain case and death information at county level in the U.S. from January 22, 2020 to June 20, 2020 but the data, forecasts, and visualizations in the GitHub repository and at website are updated daily. See Figure 1 for a high-level summary of the contributions made in this work.

2. COVID-19 DATA REPOSITORY

One of our primary contributions is the curation of a COVID-19 data repository that we have made publicly available on GitHub. It is updated daily with new information. Specifically, we have compiled and cleaned a large corpus of hospital-level and county-level data from 20+ public sources to aid data science efforts to combat COVID-19.

2.1. **Overview of the datasets on June 20, 2020.** At the *hospital-level*, our dataset covers over 7000 US hospitals and over 30 features including the hospital's CMS certification number (a unique ID of each hospital used by Centers for Medicare and Medicaid Services), the hospital's location, the number of ICU beds, the hospital type (e.g., short-term acute care, critical access), and numerous other hospital statistics.

There are more than 3,100 counties in the US. At the *county-level,* our repository includes data on

(i) daily recorded COVID-19-related case count and (recorded) death count by NY Times [43] and USA Facts [47];

(ii) demographic features such as age-wise population, and population density;

(iii) socioeconomic factors including poverty levels, unemployment, education, and social vulnerability measures;

(iv) health resource availability such as the number of hospitals, ICU beds, and medical staff;

(v) health risk indicators including heart disease, chronic respiratory disease, smoking, obesity, and diabetes prevalence;

(vi) social mobility measures such as the percent change in mobility from a pre-COVID-19 baseline; and

(vii) other relevant information such as county-level presidential election results from 2000 to 2016, county-level commute data that includes the number of workers in the commuting flow, and airline ticket survey data that includes origin, destination, and other itinerary details.

In total, there are over 8000 features in the county-level dataset. We provide a feature-level snapshot of the different types of data available in our repository, highlighting features in the county-level datasets in Table 1 and the hospital-level datasets in Table 2. Alternatively, in Tables 3 and 4, we provide an overview of the county-level and hospital-level data sources in our repository, respectively, organized by the dataset.

The full corpus of data, along with further details and extensive documentation, are available on GitHub. In particular, we have created a comprehensive data dictionary with the available data features, their descriptions, and source dataset for ease of navigation on our github. We have also provided a quick-start guide for accessing the unabridged county-level and hospital-level datasets with a single Python code line.

Datasets used by our predictors: In this paper, we focus on predicting the number of recorded COVID-19-related cumulative death counts in each county. For our analysis, we primarily use the county-level case and death reports provided by USAFacts from January 22, 2020 to June 20, 2020 (pulled on June 21, 2020) along with some county-level demographics and health data. We have marked these datasets with an asterisk (*) in Table 3. We discuss our prediction algorithms in

detail in Sections 3, 4 and 5.

Other potential use-cases for our repository: The original intent of our data repository was indeed to facilitate our work with Response4Life and aid medical supply allocation efforts. However, with time the data repository has grown to encompass a much larger audience and now supports investigations into a wide range of COVID-19 related problems. For instance, using the breadth of travel information in our repository, including (aggregated) air travel, work commutes, and social mobility data, researchers can investigate the impact of both local and between-city travel patterns on the spread of COVID-19. Our data repository also includes data on the prevalence of various COVID-19 health risk factors, including diabetes, heart disease, and chronic respiratory disease, which can be used to stratify counties. Furthermore, one can also potentially leverage socioeconomic and demographic information, as well as health resource data (e.g., number of ICU beds, medical staff) to gain a better understanding of the severity of the pandemic in a county. Stratification using these covariates is particularly crucial for assessing the COVID-19 status of rural communities, which are not directly comparable, both in terms of people and resources, to larger cities and counties that have received the most attention.

Comparison with the repository collated by Killeen et al. [29] at Johns Hopkins University: Note that similar but complementary county-level data was recently aggregated and released in another study [29]. Both our county-level repository and the repository in [29] include data on COVID-19 cases and deaths, demographics, socioeconomic information, education, and social mobility, albeit some are from different sources. For example, the repository [29] uses COVID-19 cases and deaths data from the John Hopkins University CSSE COVID-19 dashboard by Dong et al. [15] whereas our data is pulled from USAFacts [47] and the New York Times [43]. The main difference, however, between the two repositories is that our data repository also includes data on COVID-19 health risk factors. Furthermore, while the repository in [29] provides additional datasets at the state-level, we provide additional datasets at the hospital-level (given our initial goal of helping the allocation of medical supplies to hospitals, in partnership with the non-profit Response4Life). While their data repository contains both overlapping and complementary information to our repository, a thorough dataset-by-dataset comparison is beyond the scope of this work for two reasons: (i) We learned about this repository towards the completion of our work, and (ii) we were unable to find detailed documentation of how the datasets in their repository were cleaned.

2.2. **Data quality and bias.** Before introducing our prediction algorithms, it is vital to discuss the quality and limitations of the available COVID-19 data. Many downstream analyses, including ours, rely on accurate COVID-19 infection data, including accurate case and death counts. In this subsection, we focus our discussion and evaluation on the data quality of the county-level COVID-19 case and death count data. We also conduct some preliminary exploratory data analysis to shed light on the scale of bias and the possible directions of the biases in the data.

Description of County-level Features	Data Source(s)
COVID-19 Cases/Deaths	
Daily # of COVID-19-related recorded cases by US county	USAFacts [47]; The New York Times [43]
Daily # of COVID-19-related deaths by US county	USAFacts [47]; The New York Times [43]
Demographics	
Population estimate by county (2018)	Health Resources and Services Administration [23] (Area Health Resources Files)
Census population by county (2010)	Health Resources and Services Administration [23] (Area Health Resources Files)
Age 65+ population estimate by county (2017)	Health Resources and Services Administration [23] (Area Health Resources Files)
Median age by county (2010)	Health Resources and Services Administration [23] (Area Health Resources Files)
Population density per square mile by county (2010)	Health Resources and Services Administration [23] (Area Health Resources Files)
Socioeconomic Factors	
% uninsured by county (2017)	County Health Rankings & Roadmaps [13]
High school graduation rate by county (2016-17)	County Health Rankings & Roadmaps [13]
Unemployment rate by county (2018)	County Health Rankings & Roadmaps [13]
% with severe housing problems in each county (2012-16)	County Health Rankings & Roadmaps [13]
Poverty rate by county (2018)	United States Department of Agriculture, Economic Research Service [45]
Median household income by county (2018)	United States Department of Agriculture, Economic Research Service [45]
Social vulnerability index for each county	[7] (Social Vulnerability Index)
Health Resources Availability	
# of hospitals in each county	Kaiser Health News [28]
# of ICU beds in each county	Kaiser Health News [28]
# of full-time hospital employees in each county (2017)	Health Resources and Services Administration [23] (Area Health Resources Files)
# of MDs in each county (2017)	Health Resources and Services Administration [23] (Area Health Resources Files)
Health Risk Factors	
Heart disease mortality rate by county (2014-16)	Centers for Disease Control and Prevention [6] (Interactive Atlas of Heart Disease and Stroke)
Stroke mortality rate by county (2014-16)	Centers for Disease Control and Prevention [6] (Interactive Atlas of Heart Disease and Stroke)
Diabetes prevalence by county (2016)	Centers for Disease Control and Prevention [8] (Diagnosed Diabetes Atlas)
Chronic respiratory disease mortality rate by county (2014)	Institute for Health Metrics and Evaluation [27]
% of smokers by county (2017)	County Health Rankings & Roadmaps [13]
% of adults with obesity by county (2016)	County Health Rankings & Roadmaps [13]
Crude mortality rate by county (2012-16)	United States Department of Health and Human Services [46]
Social Mobility	
Start date of stay at home order by county	Killeen et al. [29]
% change in mobility at parks, workplaces, transits, groceries/pharmacies, residential, and retail/recreational areas	Google LLC [20]

Table 1. A list of select relevant features from across all county-level datasets contained in our COVID-19 repository grouped by feature topic. See Table 3 for an overview of each of the individual county-level datasets.

Description of Hospital-level Features	Data Source(s)
CMS certification number	Centers for Medicares & Medicaid Services [10] (Case Mix Index File)
Case Mix Index	Centers for Medicares & Medicaid Services [10] (Case Mix Index File); [11] (Teaching Hospitals)
Hospital location (latitude and longitude)	Homeland Infrastructure Foundation-Level Data [25]; Definitive Healthcare [14]
# of ICU/staffed/licensed beds and beds utilization rate	Definitive Healthcare [14]
Hospital type	Homeland Infrastructure Foundation-Level Data [25]; Definitive Healthcare [14]
Trauma Center Level	Homeland Infrastructure Foundation-Level Data [25]
Hospital website and telephone number	Homeland Infrastructure Foundation-Level Data [25]

Table 2. A list of select relevant features from across all hospital-level datasets contained in our COVID-19 repository. See Table 4 for an overview of each hospital-level dataset.

Though discussions on data quality issues and their possible consequences are relatively sparse in the existing literature, Angelopoulos et al. [1] discuss a variety of possible data biases in the context of estimating the case fatality ratio. They proposed a method that can theoretically account for two biases: time lag and imperfect reporting of deaths and recoveries. Unfortunately, it is hard to evaluate their method's performance since the actual death counts due to COVID-19 remain unknown. Moreover, some data biases (e.g., under-ascertainment of mild cases) for estimating the case fatality ratio do not affect estimation of future death counts. Nonetheless, many of the ideas we explore here in uncovering possible biases in the data are inspired by the work [1].

Imperfect reporting and attribution of deaths due to COVID-19: Numerous news articles have suggested that the official US COVID-19 death count is an underestimate [50]. According to The New York Times[6], on April 5, the Council of State and Territorial Epidemiologists advised states to include both the confirmed cases based on laboratory testing, and probable cases—using specific criteria for symptoms and exposure. The Centers for Disease Control adopted these definitions, and national CDC data began including confirmed and probable cases on April 14. The infection data included in our data repository (USAFacts and NY Times) contains both the probable death and the confirmed deaths beginning April 14. Although the probable death counts address imperfect reporting and attribution, it is unclear to what extent the problem is mitigated. Going forward, we use the term recorded death counts and recorded case counts to reflect that the recorded counts are based on both confirmed and probable deaths and cases.

Inconsistency across different data sources: There exist multiple sources of COVID-19 death counts in the US. In our data repository, we include data from USAFacts [47] and data from the New York Times [43]. According to USAFacts and the NY Times websites, they both collect data from state

---

[6] https://www.nytimes.com/interactive/2020/06/19/us/us-coronavirus-covid-death-toll.html

COUNTY-LEVEL DATASET	DESCRIPTION
COVID-19 Cases/Deaths Data	
USAFacts [47]*†	Daily cumulative number of reported COVID-19-related death and case counts by US county, dating back to Jan. 22, 2020
The New York Times [43]†	Similar to the USAFacts dataset, but includes aggregated death counts in New York City without county breakdowns
Demographics and Socioeconomic Factors	
Health Resources and Services Administration [23] (Area Health Resources Files)*	Includes data on health facilities, professions, resource scarcity, economic activity, and socioeconomic factors (2018-2019)
County Health Rankings & Roadmaps [13]*	Estimates of various health behaviors and socioeconomic factors (e.g., unemployment, education)
Centers for Disease Control and Prevention [7] (Social Vulnerability Index)	Reports the CDC's measure of social vulnerability from 2018
United States Department of Agriculture, Economic Research Service [45]	Poverty estimates and median household income for each county
Health Resources Availability	
Health Resources and Services Administration [23] (Area Health Resources Files)*	Includes data on health facilities, professions, resource scarcity, economic activity, and socioeconomic factors (2018-2019)
Health Resources and Services Administration [24] (Health Professional Shortage Areas)	Provides data on areas having shortages of primary care, as designated by the Health Resources & Services Administration
Kaiser Health News [28]*	# of hospitals, hospital employees, and ICU beds in each county
Health Risk Factors	
County Health Rankings & Roadmaps [13]*	Estimates of various socioeconomic factors and health behaviors (e.g., % of adult smokers, % of adults with obesity)
Centers for Disease Control and Prevention [6] (Interactive Atlas of Heart Disease and Stroke)*	Estimated heart disease and stroke death rate per 100,000 (all ages, all races/ethnicities, both genders, 2014-2016)
Centers for Disease Control and Prevention [8] (Diagnosed Diabetes Atlas)*	Estimated percentage of people who have been diagnosed with diabetes per county (2016)
Institute for Health Metrics and Evaluation [27]*	Estimated mortality rates of chronic respiratory diseases (1980-2014)
Institute for Health Metrics and Evaluation [9] (Chronic Conditions)	Prevalence of 21 chronic conditions based upon CMS administrative enrollment and claims data for Medicare beneficiaries
United States Department of Health and Human Services [46] Overall mortality rates (2012-2016) for each county from the Na- man Services [46]	Overall mortality rates (2012-2016) for each county from the National Center for Health Statistics
Social Mobility	
Killeen et al. [29] (JHU Date of Interventions)	Dates that counties (or states governing them) took measures to mitigate the spread by restricting gatherings
Google LLC, [20] (Google Community Mobility Reports)†	Reports relative movement trends over time by geography and across different categories of places (e.g., retail/recreation, groceries/pharmacies)
Apple Inc. [2] (Apple Mobility Trends)†	Uses Apple maps data to report relative (to Jan. 13, 2020) volume of directions requests per country/region, sub-region or city
Miscellaneous	
United States Census Bureau [44] (County Adjacency File)*	Lists each US county and its neighboring counties; from the US Census
Bureau of Transportation Statistics [5] (Airline Origin and Destination Survey)	Survey data with origin, destination, and itinerary details from a 10% sample of airline tickets in 2019
MIT Election Data and Science Lab [32] (County Presidential Data)	County-level returns for presidential elections from 2000 to 2016 according to official state election data records

Table 3. A list of county-level datasets contained within in our COVID-19 repository grouped by data category. Datasets marked with a † are updated daily while all other sources are static. Datasets marked with an asterisk (*) were used in the predictors discussed in this work. Several datasets are relevant to multiple categories and are thus listed multiple times. See Table 1 for an overview of select features from these county-level datasets.

Hospital-level Dataset	Description
Homeland Infrastructure Foundation-Level Data [25]	Includes number of ICU beds, and location for US hospitals
Definitive Healthcare [14]	Provides data on number of licensed beds, staffed beds, ICU beds, and the bed utilization rate for hospitals in the US
Centers for Medicares & Medicaid Services [10] (Case Mix Index File)	Reports the Case Mix Index (CMI) for each hospital
Centers for Medicares & Medicaid Services [11] (Teaching Hospitals)	Lists teaching hospitals along with address (2020)

Table 4. A list of hospital-level datasets contained within in our COVID-19 repository. Currently, all hospital-level sources are static. See Table 2 for an overview of select features from these hospital-level datasets.

and local agencies or health departments and manually curate the data. However, these websites do not scrape data from those sources at the same time. While USAFacts states that "they mostly collect data in the evening (Pacific Time)", NY Times mentions they update data throughout the day. Furthermore, while there are some discussions on how they collect and process the data on their websites, the specific data curation rules are not shared publicly. Possibly due to different scrapping times and curation rules, there are a few discrepancies in their case and death counts. In Figure 2(a), we plot the absolute difference in death counts from the two datasets for each county. In Figure 2(b), we plot the number of counties whose recorded COVID-19 deaths on a given day differ by more than 5. The proportion of counties with observably different death-counts (difference > 5) is in general small (< 1%), although sometimes the differences are quite significant (>100). Since two datasets are curated under different rules (which are unknown to us), it is not obvious how to combine them or assess their validity. For our analysis, we choose to use the USAFacts COVID-19 deaths data as they provide county-level death counts for New York City while the NY Times data aggregates the death counts over the five boroughs in that region.

Weekday patterns: The recorded case counts and death counts have a significant weekly pattern in both the USAFacts and NY Times data; such a pattern can possibly be attributed to the reporting delays as discussed in [1]. We show the total number of deaths recorded on each day of the week in the USAFacts data in Figure 3(a). The total number of deaths on Monday and Sunday is significantly lower than that for any other day. We try to account for these weekly patterns in our prediction methods later in Appendix A.

Historical data revision: We observed that some of the historical infection data was revised after initially being recorded. According to USAFacts, these revisions are typically due to earlier mistakes from local agencies which revised their previously recorded death counts. Note that these data revisions are not related to the probable deaths as we discussed earlier and therefore we regard this phenomenon as a distinct source of bias. This kind of revision is not common: until June 21, we
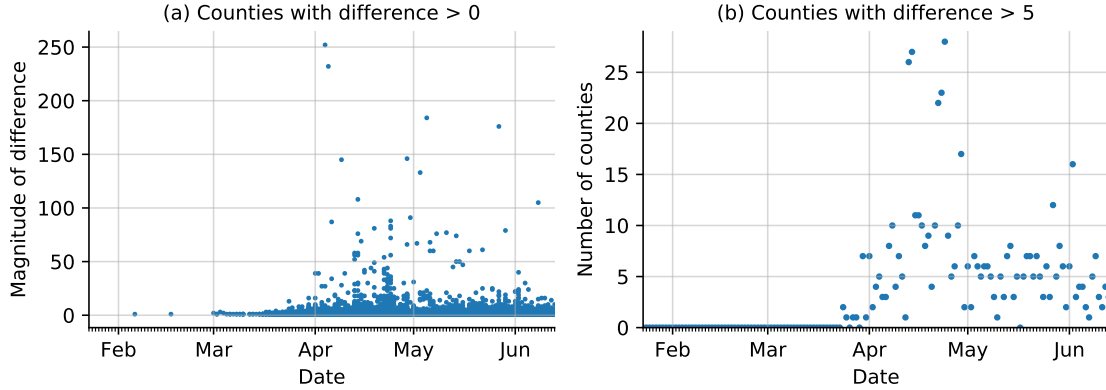
Figure 2. Plots illustrating the differences in county-wise recorded death counts between the USAFacts and NY Times datasets. In panel **(a)**, we plot the magnitude of difference in death counts for each county as a function of time, where one dot represents a particular county on a particular day. We notice that while there are counties where the discrepancy can be larger than 100 deaths recorded in a day, the majority of the discrepancies are not large. The large discrepancies are possibly due to different data curation protocols used by USAFacts and NY Times. In panel **(b)**, we plot the number of counties that have a discrepancy of more than 5 on any given day between the two datasets. We notice that on any given day, no more than 30 counties, i.e., < 1% of the more than 3,100 counties, have a difference larger than 5 between the two datasets.
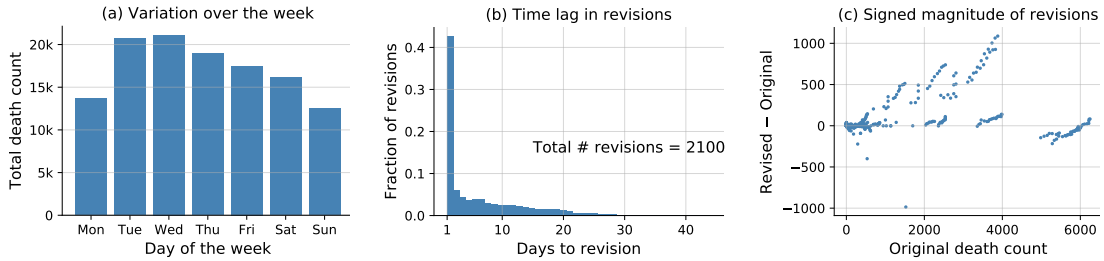
Figure 3. EDA plots for identifying potential biases in USAFacts data. **(a)** Variation of recorded death counts over different days of the week. We observe that the total death counts (for March-June) shows a trend across different weekdays, and the total count is significantly smaller for Sunday and Monday. In panel **(b)**, we plot the histogram of the number of days after which the count for a given county on a given day is revised. In panel **(c)**, we present the signed magnitude of the change, i.e., the "revised count minus the original count" against the original count for each revision.

observe that only 2.1% of counties across the U.S. had one or more historical revisions. Figure 3(b) shows a histogram of the amount of time from the initial record to the revision. It can be seen that almost half of the changes happen the day after the data was initially recorded. Figure 3(c) shows the signed magnitude of the change in death count that results from the revisions versus the

initial recorded death counts. Note that there are a few stripes (consecutive upward revisions) in the plot. Each stripe corresponds to the revision of a particular county on different dates. Since the reported data is cumulative death counts, when the deaths from a few days ago get revised, all the data after that day until the day when the revision is made also get revised accordingly, thereby explaining the short stripy trends.

However, only 582 out of 2100 revisions (around 27%) have absolute magnitude > 2 deaths, and 354 of these 582 revisions (around 67%) are in the positive direction (i.e., more deaths than initially recorded). Furthermore, amongst the 61 revisions with an absolute magnitude larger than 200, almost all of them (57/61) lead to an increase in the number of recorded deaths. The four most significant downward revisions, i.e., the points with large negative "revised-original count" in Figure 3(c), correspond to counties in the Washington State. This finding can be corroborated by the media news that Washington State admitted errors in reporting the death counts, and subsequently lowered these counts in the revisions.[7] It is natural for our predictions to vary if the training data (for a fixed period) varies with time, i.e., when the COVID-19 counts are adjusted for a backdate. Most of these revisions are minor, in which case the general performance of our predictors does not change significantly. However, when the revisions are a significant uptick, the predictions can become unstable for a few days (depending on the uptick, and the prediction-horizon). See Section 5.1 and 5.2 for further discussion on these biases. In this paper, we use the initial infection data available on June 21, 2020 to evaluate our algorithm performance, i.e., we do not use data that was revised after June 21. Nonetheless, we caution the reader to keep the following fact in mind while interpreting results from our work as well as other related COVID-19 studies: The recorded death counts themselves are an under-estimate and the consequent bias is hard to adjust for due to the lack of ground truth.

3. Predictors for forecasting short-term death counts

Figure 4 provides a visualization of the COVID-19 outbreak across the United States. We plot (a) the cumulative recorded death counts due to COVID-19 up to June 20, and (b) the new death counts from June 1 to June 20, 2020. Each bubble denotes a county-level count, a darker and larger bubble denotes a higher death count, and the absence of a bubble denotes that the count is zero. Panel (a) captures the extent of the outbreak in a region, while (b) captures the recent trends in the outbreak. The color scale differs between the two plots to better illustrate the respective counts in each plot, but the size scales are held constant between the two plots to help provide a comparison between the extent and recent trends of COVID-19. Overall, Figure 4 clearly shows that the COVID-19 outbreak in the United States is incredibly dynamic both in time and across different regions. The worst-affected regions include the states of New York, New Jersey, Massachusetts,

---

[7]See https://www.clarkcountytoday.com/news/washington-department-of-health-clarifies-covid-19-death-numbers/ and https://www.king5.com/article/news/health/coronavirus/washington-coronavirus-testing-death-toll-mistake, last accessed on July 24, 2020.

Michigan, Illinois, Florida, Louisiana, Georgia, Washington, and California. Moreover, most of these areas continue to face a substantial COVID-19 burden in the first two-thirds of June.

We develop several different statistical or machine learning prediction algorithms to capture the dynamic behavior of COVID-19 death counts. Since each prediction algorithm captures slightly different trends in the data, we also develop various weighted combinations of these prediction algorithms. The five prediction algorithms or predictors for cumulative recorded death counts that we devise in this paper are as follows:

(1) **A separate-county exponential predictor (the "separate" predictors)**: a series of predictors built for predicting cumulative death counts for each county using only past death counts from that county.

(2) **A separate-county linear predictor (the "linear" predictor)**: a predictor similar to the separate county exponential predictors, but uses a simple linear format, rather than the exponential format.

(3) **A shared-county exponential predictor (the "shared" predictor)**: a single predictor built using death counts from all counties, used to predict death counts for individual counties.

(4) **An expanded shared-county exponential predictor (the "expanded shared" predictor)**: a predictor similar to the shared-county exponential predictor, which also includes COVID-19 case numbers and neighboring county cases and deaths as predictive features.

(5) **A demographics shared-county exponential predictor (the "demographics shared" predictor)**: a predictor also similar to the shared-county exponential predictor, but which also includes various county demographic and health-related predictive features.

An overview of these predictors is presented in Table 5. We use the python package statsmodels [40] to train all the five predictors with the same Poisson log-likelihood loss function (but the set of features for each predictor is different).[8] To combine the different trends captured by each of these predictors, we also fit various combinations of them, which we refer to as Combined Linear and Exponential Predictors (CLEP). CLEP produces a weighted average of the predictions from the

---

[8]We use the default parameters in the Python statsmodels package (version 0.11.1) while training our predictors. For predictors (1) and (2), *glm.fit* was used which always converged. For predictors (3)-(5), we first tried *glm.fit_regularized* since we experimented with the use of explicit ℓ_1 and ℓ_2 regularization in the beginning. Although, eventually the predictors (3)-(5) reported in this paper did not use any explicit ℓ_1 or ℓ_2 regularization, it turns out that default settings (algorithm and stopping criterion) in the functions *glm.fit* and *glm.fit_regularized* are different leading to different implicit regularizations, and consequently different performance. We still chose to use *glm.fit_regularized* for fitting predictors (3)-(5) since it led to better performance for predictor (4) (which forms the basis of our best performing predictor CLEP). We also note that the function *glm.fit_regularized*, in fact, calls another function *fit_elasticnet* which uses block coordinate descent (BCD) by default to solve a generalized linear model. The default value for the maximum number of iterations of BCD (max_iter) is set to be 50, which, in some cases, resulted in early stopping (a form of implicit regularization) before the iterative algorithm converges.
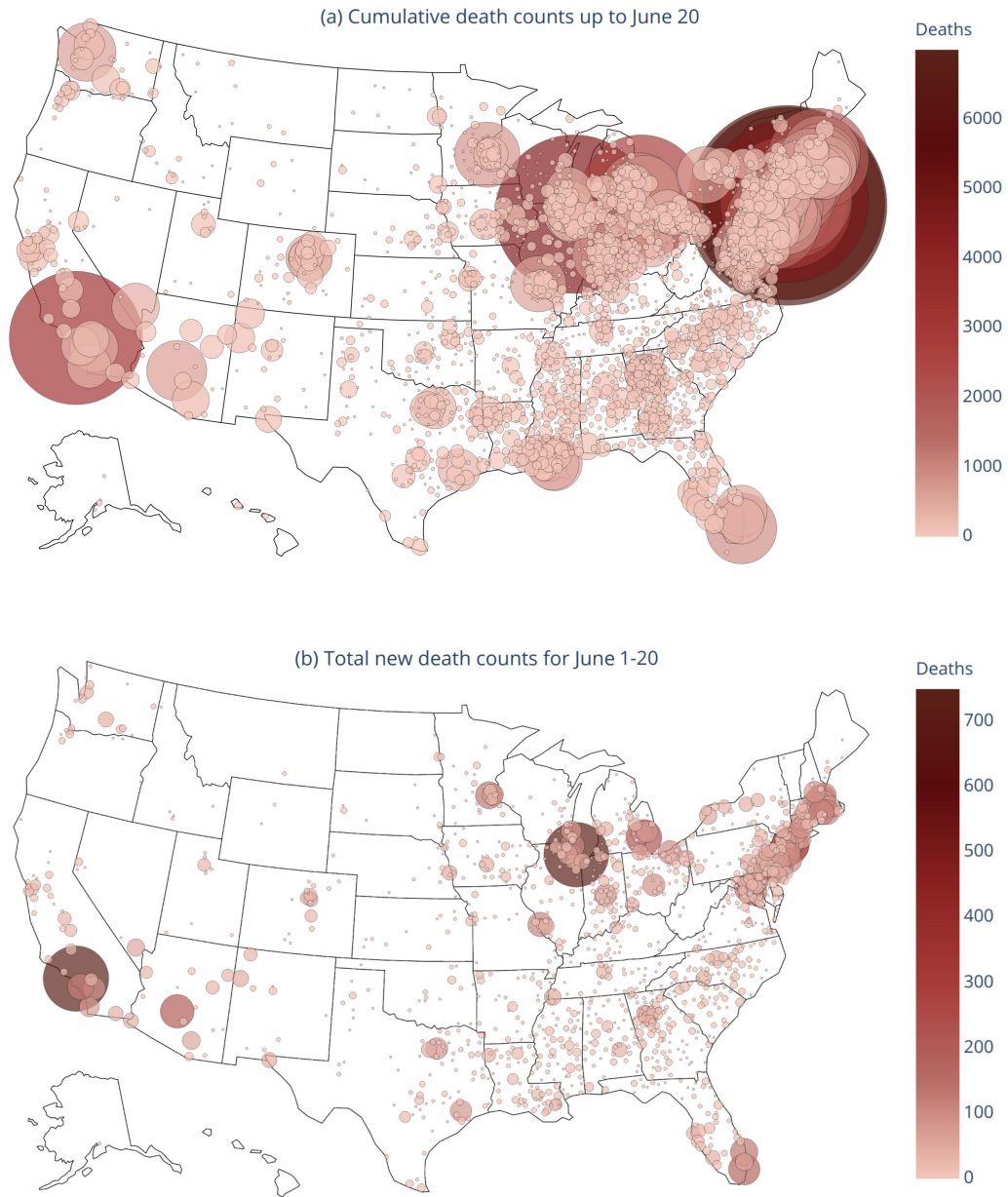
Figure 4. Visualization of the COVID-19 outbreak in the US. We depict the cumulative recorded death counts up to June 20 in panel **(a)** and newly recorded death counts for the period June 1-20 in panel **(b)**. Each bubble denotes the death count for a county (the absence of a bubble denotes a zero count). The bubble size (area) is proportional to the death counts in the region. The two panels' bubble sizes are on the same scale, but the color scale is different as shown respectively on each plot.

individual predictors, where we borrow the weighting scheme from prior work [39]. In this weighting scheme, a higher weight is given to those predictors with more accurate predictions, especially on recent time points. We find that the CLEP that combines only the *linear predictor* and the *expanded shared predictor* consistently has the best predictive performance when compared to the individual predictors and the CLEP that combines all five predictors. (We did not try all possible combinations to avoid over-fitting; also see Table 6). For the rest of this section, we expand upon the individual predictor models and the weighting procedure for the CLEP ensembles.

In addition, Appendix A contains results on variants of the two best single predictors (linear and expanded shared), which include features for social-distancing and features that account for the under-reporting of deaths on Sunday and Monday (as observed in Figure 3(a)). These additional features did not lead to better performance.

We note that although in this paper, we discuss our algorithms for predicting cumulative recorded death counts, the methods can be more generally applied to predict other quantities of interest, e.g., case counts, or new death counts for each day. Moreover, the combination scheme used for combining different predictors can be of independent interest in developing ensembling schemes with generic machine learning methods.

Predictor name	Type	Fit separately to each county?	Fit jointly to all counties?	Use neighboring counties?	Use demographics?
Separate	Exponential	✓			
Linear	Linear	✓			
Shared	Exponential		✓		
Expanded shared	Exponential		✓	✓	
Demographics shared	Exponential		✓		✓

Table 5. Overview of the 5 predictors used here. The best model is a combination of the linear predictor and the expanded shared predictor (see Section 3.6).

3.1. The separate-county exponential predictors (the "separate" predictors). The separate-county exponential predictor aims to capture the observed exponential growth of COVID-19 deaths [34]. We approximate an exponential curve for death count separately for each county using the most recent 5 days of data from that county. These predictors have the following form:

$$(3.1) \qquad \widehat{E}[\text{deaths}^c_{t+1}|t] = \exp\left(\beta_0^c + \beta_1^c(t+1)\right),$$

where $\widehat{E}[\text{deaths}^c_{t+1}|t]$ denotes the (fitted) cumulative death count by the end of day $t+1$ for county c, and it is trained on the data until day t, and computed on the morning of day $t + 1$. Note that we use $t + 1$ on the RHS of equation 3.1 just for notational exposition, and in practice we just use $\beta_0^c + \beta_1^c t$ in the exponent in our code.

Here we fit a separate predictor for each county, and the coefficients β_0^c and β_1^c for each county c are fit using maximum likelihood estimation under a Poisson generalized linear model (GLM) with t as the independent variable and deaths$_t$ as the observed variable. In simple words, on the morning of day $t+1$, the coefficients are estimated using the cumulative recorded death counts for day $\{t, t-1, t-2, t-3, t-4\}$. And to predict k-days-ahead cumulative death count on the morning of day $t+1$—denoted by $\widehat{\mathrm{E}}[\mathrm{deaths}_{t+k}^c|t]$—we simply replace $t+1$ with $t+k$ on the RHS of equation 3.1. Note that although the prediction $\widehat{\mathrm{E}}(\mathrm{deaths}_{t+1}^c|t)$ is being made on day $t+1$, we call it 1-day-ahead prediction since it is made in the morning of day $t+1$ using the data for up to day t. Moreover, the recorded count deaths$_{t+1}^c$ is reported only late in the night of day $t+1$ or early morning of the next day $t+2$.

If the first death in a county occurred less than 5 days prior to fitting the predictor, only the days from the first death were used for the fit. If there is less than three days' worth of data or the cumulative deaths remain constant in the past days, we simply use the most recent deaths as the predicted future value. We also fit exponential predictors to the full time-series (as opposed to just the most recent 5 days) of available data for each county. However, due to the rapidly shifting trends, these performed worse than our 5-day predictors. We also found that predictors fit using 6 days of data yielded similar results to predictors fit using 5 days of data, and using 4 days of data performed slightly worse.

To handle possible over-dispersion of data (when the variance is larger than the mean), we also explored estimating $\{\beta_0^c, \beta_1^c\}$ by fitting a negative binomial regression model (in place of Poisson GLM) with inverse-scale parameter taking values in $\{0.05, 0.15, 1\}$. However, we found that this approach yields a larger mean absolute error than the Poisson GLM for counties with more than 10 deaths.

3.2. **The separate-county linear predictor (the "separate linear" predictor).** The separate linear predictor aims to capture linear growth, based on the most recent 4 days of data in each county. In the early stages of tuning, we tried using 5 and 7 days of data, and obtained worse performance (see Appendix 14.). The motivation for the linear model is that some counties are exhibiting sub-exponential growth. For these counties, the exponential predictors introduced in the previous section may not be a good fit to the data. The separate linear predictors are given by

$$(3.2) \qquad \widehat{\mathrm{E}}[\mathrm{deaths}_{t+1}^c|t] = \beta_0^c + \beta_1^c(t+1),$$

where we fit the coefficients β_0^c and β_1^c via ordinary least squares using the cumulative death count for county c for most recent 4 days. Like equation 3.1, we use $t+1$ on the RHS simply for notational exposition. Put simply, on the morning of day $t+1$, the coefficients $\{\beta_0^c, \beta_1^c\}$ are estimated using the death counts for day $t, t-1, t-2, t-3$. To predict k-days-ahead, i.e., predict cumulative death counts by the end of day $t+k$ on the morning of day $t+1$ (in our notation, $\widehat{\mathrm{E}}[\mathrm{deaths}_{t+k}^c|t]$), we simply replace $t+1$ by $t+k$ on the RHS of equation 3.2.

3.3. **The shared-county exponential predictor (the "shared" predictor).** To incorporate additional data into our predictions, we fit a predictor that combines data across different counties. Rather than producing a separate predictor model for each county (as in the separate predictor approach above), we instead produce a single shared predictor that pools information from counties across the nation. The shared predictor is then used to predict future deaths in the individual counties. These changes allows us to leverage the early-stage trends from counties that are now much further along in the pandemic trajectory to inform the predictions for other current earlier-stage counties.

The data underlying the shared predictor is slightly different from the separate county predictors. Instead of only including the most recent 5 days of data from each county, we include all days after the third death in each county. (In the earlier stages of tuning, we also tried including the counties after first and fifth death, and then selected the choice of third death due to better performance.) Thus the data from many of the counties extend substantially further back than 5 days, and for each county, $t = 0$ is the day on which the third death occurred. Instead of basing the exponential predictor prediction on time $t + 1$ (as was the case for the separate predictors above), we base the prediction on the logarithm of the previous day's death count. This choice makes the counties comparable since the outbreaks began at different time points in each county. The shared predictor is given as follows:

$$(3.3) \qquad \widehat{\mathrm{E}}[\mathrm{deaths}_{t+1}^c | t] = \exp\left(\beta_0 + \beta_1 \log(\mathrm{deaths}_t^c + 1) \right),$$

where $\widehat{\mathrm{E}}[\mathrm{deaths}_{t+1}^c | t]$ denotes the (fitted) cumulative death count by the end of day $t + 1$ for a county c, and deaths_t^c denotes the recorded cumulative death count for that county by the end of day t. The coefficients β_0 and β_1 are shared across all counties and fitted by maximizing the log-likelihood corresponding to Poisson GLM (like that in the separate county predictor given by equation 3.1). We normalize the feature matrix to have zero mean and unit variance before fitting the coefficients. To predict k-days-ahead cumulative death count $\widehat{\mathrm{E}}[\mathrm{deaths}_{t+k}^c | t]$, we first obtain the estimate $\widehat{\mathrm{E}}[\mathrm{deaths}_{t+1}^c | t]$ using equation 3.3. Next, we plug-in $\log(\widehat{\mathrm{E}}[\mathrm{deaths}_{t+j}^c | t] + 1)$ on the RHS of equation 3.3 to compute $\widehat{\mathrm{E}}[\mathrm{deaths}_{t+j+1}^c | t]$ in a sequential manner for $j = 1, \ldots, k - 1$, and finally obtain $\widehat{\mathrm{E}}[\mathrm{deaths}_{t+k}^c | t]$ (k-day-ahead prediction computed on the morning of day $t + 1$).

3.4. **The expanded shared exponential predictor (the "expanded shared" predictor).** Next, we expand the shared county exponential predictor to include other COVID-19 dynamic (time-series) features. In particular, we include the number of recorded *cases* in the county, as this may give an additional indication to the severity of an outbreak. We also include the total sum of cumulative death (and case) counts in the *neighboring* counties. Let cases_t^c, $\mathrm{neigh\_deaths}_t^c$, $\mathrm{neigh\_cases}_t^c$ respectively denote the (recorded) cumulative case count in the county c at the end of day t, the total sum of cumulative death counts across all its neighboring counties at the end of day t, and the total sum of cumulative recorded case counts across all its neighboring counties at the end of day t. Then our (expanded) predictor to predict the number of recorded cumulative

deaths k days into the future is given by

$$\widehat{E}[\text{deaths}_{t+1}^c | t] = \exp\Bigg(\beta_0 + \beta_1 \log(\text{deaths}_t^c + 1) + \beta_2 \log(\text{cases}_{t-k+1}^c + 1)$$

$$(3.4) \qquad\qquad\qquad + \beta_3 \log(\text{neigh\_deaths}_{t-k+1}^c + 1) + \beta_4 \log(\text{neigh\_cases}_{t-k+1}^c + 1)\Bigg),$$

where the coefficients $\{\beta_i\}_{i=0}^4$ are shared across all counties and are fitted using the Poisson GLM after normalization of each feature (in the exponent) to have zero mean and unit variance. When *fitting* the predictor on the morning of day $t + 1$, we use the death counts for the county up to the end of day t. However, we only use the new features (cases in the current county, cases in neighboring counties, and deaths in neighboring counties) up to the end of day $t - k + 1$. Moreover, we normalize the feature matrix to have zero mean and unit variance before fitting the predictor. While *predicting* the death count for a given county k days into the future (i.e, the cumulative death count by the end of day $t + k$), we iteratively use the daily sequential predictions for the death counts for that county, and use the information for the other features only up to time t (the time up to which we have data available). More precisely, first we estimate $\widehat{E}[\text{deaths}_{t+1}^c | t]$ by plugging in the normalized features $\log(\text{deaths}_t^c)$, $\log(\text{cases}_{t-k+1}^c)$, $\log(\text{neigh\_deaths}_{t-k+1}^c)$, and $\log(\text{neigh\_cases}_{t-k+1}^c)$ in equation 3.4, where the normalization is done across counties so that each feature has zero mean and unit variance. Then, for $j = 1, 2, \ldots, k - 1$, we recursively plug-in $\log(\widehat{E}[\text{deaths}_{t+j}^c | t])$, $\log(\text{cases}_{t-k+j+1}^c)$, $\log(\text{neigh\_deaths}_{t-k+j+1}^c)$, $\log(\text{neigh\_cases}_{t-k+j+1}^c))$ in equation 3.4 (again after normalizing each of these features) to compute $\widehat{E}[\text{deaths}_{t+j+1}^c | t]$, and finally obtain the compute $\widehat{E}[\text{deaths}_{t+k}^c | t]$ for k-day-ahead prediction made with data until day t. It may be possible to jointly predict the new features along with the number of deaths, but we leave building such a predictor to future work. As before, the predictor is fitted by including all days after the third death in each county.

3.5. The demographics shared exponential predictor (the "demographics shared" predictor).
The demographics shared county exponential predictor is again very similar to the shared predictor. However, it includes several static county demographic and healthcare-related features to address the fact that some counties will be affected more severely than others, for instance, due to (a) their population makeup, e.g., older populations are likely to experience a higher death rate than younger populations, (b) their hospital preparedness, e.g., if a county has very few ICU beds relative to their population, they might experience a higher death rate since the number of ICU beds is correlated strongly (0.96) with the number of ventilators [38], and (c) their population health, e.g., age, smoking history, diabetes, and cardiovascular disease are all considered to be likely risk factors for acute COVID-19 infection [21, 37, 22, 19, 51].

For a county c, given a set of demographic and healthcare-related features $d_1^c, \ldots, d_m^c$ (such as median age, population density, or number of ICU beds), the demographics shared predictor is

given by

$$(3.5) \qquad \widehat{\mathrm{E}}[\mathrm{deaths}_{t+1}^c | t] = \exp\left(\beta_0 + \beta_1 \log(\mathrm{deaths}_t^c + 1) + \beta_{d_1} d_1^c + \cdots + \beta_{d_m} d_m^c\right).$$

Here the coefficients $\{\beta_0, \beta_1, \beta_{d_1}, \ldots, \beta_{d_m}\}$ are shared across all counties, and are fitted by maximizing the log-likelihood of the corresponding Poisson generalized linear model, where we include all the observations since the third death in each county. Moreover, we also normalize the feature matrix to have zero mean and unit variance before fitting the coefficients. The features we choose fall into three categories:

(1) County density and size: population density per square mile (2010), population estimate (2018)
(2) County healthcare resources: number of hospitals (2018-2019), number of ICU beds (2018-2019)
(3) County health demographics: median age (2010), percentage of the population who are smokers (2017), percentage of the population with diabetes (2016), deaths due to heart diseases per 100,000 (2014-2016).

The k-day-ahead predictions for this predictor are obtained in a very similar manner to the shared predictor (3.3): We first obtain the estimate $\widehat{\mathrm{E}}[\mathrm{deaths}_{t+1}^c | t]$ using equation 3.5 and then, sequentially plug-in $\log(\widehat{\mathrm{E}}[\mathrm{deaths}_{t+j}^c | t] + 1)$ on the RHS of the equation 3.5 (after normalization to obtain zero mean and unit variance) to compute $\widehat{\mathrm{E}}[\mathrm{deaths}_{t+j+1}^c | t]$ in a sequential manner for $j = 1, \ldots, k-1$.

3.6. **The combined predictors: CLEP.** Finally, we consider various combinations of the five predictors we have introduced above using an ensemble approach similar to that described in [39]. Specifically, we use the recent predictive performance (e.g., over the last week) of different predictors to guide an adaptive tuning of the corresponding weights in the ensemble. To simplify notation, let us denote the predictions for cumulative death count by the end of day $t+k$—where the prediction is made on the morning of day $t+1$—by $\{\widehat{y}_{t+k}^m\}$ with $m = 1, \ldots, M$ denoting the index of various linear and exponential predictors.[9] Then, their Combined Linear and Exponential Predictor (CLEP) is given by

$$(3.6) \qquad \widehat{y}_{t+k}^{\mathrm{CLEP}} = \sum_{m=1}^{M} w_{t+1}^m \widehat{y}_{t+k}^m.$$

---

[9]Our predictions are released around 11:30 AM Pacific Time each day, both on GitHub and our website (https://covidseverity.com). The released predictions on day $t+1$ include the county-wise predictions for cumulative death counts by the end of day $t+1$ itself. To summarize, 1-day-ahead prediction for day $t+1$ is denoted by $\widehat{\mathrm{E}}[\mathrm{deaths}_{t+1}^c | t]$ earlier is now written simply as $\widehat{y}_{t+1}$. Similarly, the k-day-ahead prediction $\widehat{\mathrm{E}}[\mathrm{deaths}_{t+k}^c | t]$ is denoted by $\widehat{y}_{t+k}$ in the simplified (and slightly abused) notation.

Here the weight, w_{t+1}^m—used for combining the predictions made on the morning of day $t+1$—for predictor m, is computed according to the recent performance as follows:

$$(3.7) \qquad w_{t+1}^m \propto \exp\left(-0.5 \sum_{i=t-6}^{t} (0.5)^{t-i} \left| \sqrt{\widehat{y}_i^m} - \sqrt{y_i} \right| \right),$$

where $\widehat{y}_i^m$ is the 3-day-ahead prediction from the predictor m trained on data up to time $i-3$ (and computed on the morning of day $i-2$). In addition, the weights are normalized so that $\sum_{m=1}^{M} w_{t+1}^m = 1$ for each $t+1$. The weights $\{w_{t+1}^m, m = 1, \ldots, M\}$ are computed separately for each county. We now turn to our discussion on the general combination scheme that leads to the equation 3.7 with a certain choice of hyperparameters (and how those hyperparameters were chosen).

The weights in equation 3.7 are based on the general ensemble weighting format introduced in [39]. This general format is given by

$$(3.8) \qquad w_{t+1}^m \propto \exp\left(-c(1-\mu) \sum_{i=t_0}^{t} \mu^{t-i} \ell(\widehat{y}_i^m, y_i) \right),$$

where $\mu \in (0, 1)$ and $c > 0$ are tuning parameters, t_0 represents some past time point, and the weights are computed on the morning of day $t+1$. Since $\mu < 1$, the μ^{t-i} term represents the greater influence given to more recent predictive performance. For a given day i and predictor m, we measure the predictive performance of the predictor via the term $\ell(\widehat{y}_i^m, y_i)$, which denotes the loss incurred due to the discrepancy between its predicted number of deaths $\widehat{y}_i^m$ and the recorded death counts y_i. The hyperparameter c controls the relative importance of predictors depending on their recent predictive performance. Given the same recent predictive performance and μ, a larger c gives a higher weight to the better predictors. The hyperparameter t_0 denotes the number of recent days used for evaluating the predictor performance to influence the weight (3.8).

Choice of hyper-parameters: Equation 3.7 corresponds to equation 3.8 with appropriate hyperparameters, c, μ, t_0, and a specific loss format, ℓ. In [39], the authors used the loss function $\ell(\widehat{y}_i^m, y_i) = |\widehat{y}_i^m - y_i|$, since their errors roughly had a Laplacian distribution. In our case, we found that this loss function led to vanishing weights due to our error distribution's heavy-tailed nature. To help address this, we apply a square root to the predictions and the true values, and define $\ell(\widehat{y}_i^m, y_i) = |\sqrt{\widehat{y}_i^m} - \sqrt{y_i}|$. We found that this transformation improved performance in practice. We also considered a logarithmic transform instead of a square root (i.e., $\ell(\widehat{y}_i^m, y_i) = |\log(1 + \widehat{y}_i^m) - \log(1 + y_i)|$), but we found that using the logarithm yielded worse performance than using the square root transformation.[10]

---

[10]In our first submission on May 16, 2020 to arXiv, we had presented results for March 22 to May 10, 2020. During the preparation of manuscript, we had updated the transform to be the square-root transform in our code, but we did not update the CLEP equation in the paper, and erroneously reported that our CLEP weights used a logarithmic transform.

To generate our predictions, we use the default value of c in [39] which is 1. However, we change the value of μ from the default of 0.9 to 0.5 for two reasons: (i) we found $\mu = 0.5$ yielded better empirical performance, and (ii) it ensured that performance more than a week ago had little influence over the predictor. We chose $t_0 = t - 6$ (i.e., we aggregate the predictions of the past week into the weight term), since we found that performance did not improve by extending further back than 7 days. Moreover, the information from more than a week effectively has a vanishing effect due to our choice of μ.

Finally, we found that for computing the weights in (3.7), using 3-day-ahead predictions in the loss terms $\ell(\widehat{y}_i^m, y_i)$ led to best predictive performance; i.e., these weights are computed based on the 3-day-ahead predictions generated over the course of a week starting with the predictor built 11 days ago (for predicting counts 8 days ago) up to the predictor built 4 days ago (for predicting yesterday's counts). In principle, the five hyper-parameters—c, μ, t_0, ℓ, and the choice of the prediction horizon to use for evaluating the loss ℓ—can be tuned jointly via a grid or randomized search. Nevertheless, to keep the computations tractable and our choices interpretable, we selected them sequentially. Moreover, a dynamic tuning of these hyper-parameters (over time) is left for future work (see last paragraph of Section 6).

3.7. **Ensuring monotonicity of predictions.** In this work, we predict county-wise cumulative death count, which is a non-decreasing sequence. However, the predictors discussed in the previous sections need not provide monotonous estimates for different prediction horizon, i.e., $\widehat{\mathrm{E}}[\mathrm{deaths}_{t+k}^c|t]$ may decrease as k increases for a fixed t. Moreover, the predictors may estimate a future count that is smaller than the last observed cumulative death count, i.e., $\widehat{\mathrm{E}}[\mathrm{deaths}_{t+k}^c|t] < \mathrm{deaths}_t^c$. In our setting, expanded shared predictor exhibited both these issues. To avoid these pitfalls, we use post-hoc maxima adjustments for all the predictors as follows. First, we replace the estimate $\widehat{\mathrm{E}}[\mathrm{deaths}_{t+1}^c|t]$ by $\max\{\widehat{\mathrm{E}}[\mathrm{deaths}_{t+1}^c|t], \mathrm{deaths}_t^c\}$ to make sure that the predicted counts in the future are at least as large as the latest observed cumulative death counts. Next, we iteratively replace the estimate $\widehat{\mathrm{E}}[\mathrm{deaths}_{t+j}^c|t]$ by $\max\{\widehat{\mathrm{E}}[\mathrm{deaths}_{t+j}^c|t], \widehat{\mathrm{E}}[\mathrm{deaths}_{t+j-1}^c|t]\}$ for $j = 2, 3, \ldots 21$. Imposing these constraints for the individual predictors also ensures the monotonicity of predictions by the CLEP. Note that we use these monotonous predictions (after the maxima calculations) to determine the weights in equation 3.7.[11]

Note that even after imposing the previous monotonicity corrections, it is still possible that $\widehat{\mathrm{E}}[\mathrm{deaths}_{t+k}^c|t] > \widehat{\mathrm{E}}[\mathrm{deaths}_{t+k+1}^c|t + 1]$ since the predictors are re-fitted over time. Hence, when plotted over time, k-day-ahead predictions, need not be monotonous with respect to t. For example, see the plots of 7-day-ahead predictions in Figure 9 and 14-day-ahead predictions in Figure 10.

---

[11]We report partial results up to 21-day-ahead predictions, and detailed results up to 14-day-ahead predictions in Section 5. In the first arXiv submission of this work on May 16, 2020, we had not implemented monotonicity of predictions. The monotonicity implementation improved the overall results both for predictions and prediction intervals.

4. Prediction intervals via conformal inference

Accurate assessment of the uncertainty of forecasts is necessary to help determine how much emphasis to put on them, for instance, when making policy decisions. As such, the next goal of the paper is to quantify the uncertainty of our predictions by creating prediction intervals. A common method to do so involves constructing (probabilistic) model-based confidence intervals, which rely heavily on the probabilistic assumptions made about the data. However, due to the highly dynamic nature of COVID-19, assumptions on the distribution of death and case rate are challenging to check. Moreover, such prediction intervals based on probability models are likely to be invalid when the underlying probability model does not hold to the desired extent. For instance, a recent study [31] reported that the 95% uncertainty credible intervals for state-level daily mortality predicted by the initial IHME model [42], had a coverage of a mere 27% to 51% of recorded death counts over March 29 to April 2. The authors of the IHME model noted this behavior, and have since updated their uncertainty intervals so that they now provide more than 95% coverage (where coverage is defined below in equation 4.4a). However, while the previous releases of the intervals were based on asymptotic confidence intervals, the IHME authors have not precisely described the methodology for their more recent intervals. In this section, we construct prediction intervals that attempt to avoid these pitfalls by taking into account the recent observed performance of our predictors; and later in Section 5.3, we show that these intervals obtain high empirical coverage while maintaining reasonable width.

4.1. **Maximum-absolute-Error Prediction Interval (MEPI).** We now introduce a generic method to construct prediction intervals for sequential or time-series data. In particular, we build on the ideas from conformal inference [48] and make use of the past errors made by a predictor to estimate the uncertainty for its future predictions.

To construct prediction intervals for county-level cumulative death counts caused by COVID-19, we calculate the largest (normalized absolute) error for the death count predictions generated over the past 5 days for the county of interest and use this value (the "maximum absolute error") to create an interval surrounding the future (e.g., tomorrow's) prediction. We call this interval the Maximum absolute Error Prediction Interval (MEPI).

Let y_t be the actual recorded cumulative deaths by the end of day t, and $\widehat{y}_t$ denote the estimate for y_t made k days earlier (in our case on the morning of day $t - k + 1$) by a prediction algorithm. We call $\widehat{y}_t$ the k-day-ahead prediction for day t. (Note that we suppress the dependence on county and prediction horizon k for brevity and ease of exposition; and the notation here is slightly abused version of that used in Section 3.6. In particular, we have $\widehat{y}_t = \widehat{\mathrm{E}}[\mathrm{deaths}_t^c | t - k]$ for some fixed county c.) We define the normalized absolute error, Δ_t, of the prediction, $\widehat{y}_t$, to be

$$(4.1) \qquad \Delta_t := \left| \frac{y_t}{\max\{\widehat{y}_t, 1\}} - 1 \right|.$$

We use the normalization so that y_t (when non-zero) is equal to either $\widehat{y}_t(1 - \Delta_t)$ or $\widehat{y}_t(1 + \Delta_t)$. This normalization addresses the fact that the <mark>counts are increasing over time</mark>, and thus the un-normalized errors, $|y_t - \widehat{y}_t|$, also tend to be increasing over time. The normalization ensures that the errors across time are comparable in magnitude, which is essential for the exchangeability of the errors (see Section 4.3).

To compute the k-day-ahead prediction interval for day $t + k$ on the morning of day $t + 1$, we first compute the k-day-ahead prediction $\widehat{y}_{t+k}$ $(= \widehat{\mathrm{E}}[\mathrm{deaths}^c_{t+k}|t])$ using a CLEP. Next, we compute the normalized errors for the k-day-ahead predictions for the most recent 5 days $\Delta_t, \Delta_{t-1}, ..., \Delta_{t-4}$ (5 days was chosen to balance the trade-off between coverage and length, see Appendix B.2 for more details). The largest of these normalized errors is then used to define the *maximum absolute error prediction intervals* (MEPI) for the k-day-ahead prediction as follows:

$$(4.2a) \qquad \widehat{\mathrm{PI}}_{t+k} := \left[\max\left\{\widehat{y}_{t+k}(1 - \Delta_{\max}), y_t\right\}, \; \widehat{y}_{t+k}(1 + \Delta_{\max})\right],$$

$$(4.2b) \qquad \text{where} \quad \Delta_{\max} := \max_{0 \le j \le 4} \Delta_{t-j};$$

where the lower bound for the interval includes a maxima calculation to account for the fact that y_t is a cumulative count, and thereby non-decreasing. This maxima calculation ensures that the lower bound for the interval is not smaller than the last observed value.

For a general setting beyond increasing time-series, this maxima calculation can be dropped, and the MEPIs can be defined simply as

$$(4.3) \qquad \boxed{[\widehat{y}_{t+k}(1 - \Delta_{\max}), \; \widehat{y}_{t+k}(1 + \Delta_{\max})].}$$

In our case, we construct the MEPIs (4.2a) separately for each county for the cumulative death counts. We remind the reader that when constructing k-day-ahead MEPIs, the Δ_t defined in equation 4.1 is computed using k-day-ahead predictions (our notation does not highlight this fact), so that the maximum error $\Delta_{\max}$ would be typically different, say, for 7-day-ahead and 14-day-ahead predictions.

4.2. **Evaluation metrics.** For any time-series setting, stationary or otherwise, the quality of a prediction interval can be assessed in terms of the percentage of time—over a not too short period—that the prediction interval covers the observed value of the target of interest (e.g., recorded cumulative death counts as in this paper). A good prediction interval should both contain the true value most of the time, i.e., <mark>have a good coverage, and have a reasonable width or length</mark>.[12] Indeed, one can trivially create very wide prediction intervals that would always contain the target of interest. We thus consider two metrics to measure the performance of prediction intervals: <mark>*coverage*</mark> and <mark>*normalized length*</mark>.

Let y_t denote a positive real-valued time-series of interest, which in this case is the target variable: COVID-19 deaths (t denotes the time index). Let $\{\widehat{\mathrm{PI}}_t = [a_t, b_t]\}$ denote the sequence of prediction intervals produced by an algorithm. The coverage of this prediction interval, $\mathrm{Coverage}(\mathcal{T})$, over

---

[12]We use the terms width and length for an interval interchangeably in this paper.

a specified period, $\mathcal{T}$, corresponds to the fraction of days in this period for which the prediction interval contained the observed cumulative death counts. This notion of *coverage* for streaming data has been used extensively in prior works on conformal inference [48] and can be calculated for a given evaluation period $\mathcal{T}$ (which we set to be from April 11 to June 20) as follows:

$$(4.4a) \qquad \text{Coverage}(\mathcal{T}) = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \mathbb{I}(y_t \in \widehat{\text{PI}}_t),$$

where $\mathbb{I}(y_t \in \widehat{\text{PI}}_t)$ takes value 1 if y_t belongs to the interval $\widehat{\text{PI}}_t$ and 0 otherwise. The average *normalized length* of the prediction intervals, $\text{NL}(\mathcal{T})$, is calculated as follows:

$$(4.4b) \qquad \text{NL}(\mathcal{T}) = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \frac{b_t - a_t}{y_t}.$$

In practice, we replace the denominator on the RHS of equation 4.4b with $\max\{1, y_t\}$ to avoid possible division by 0. We use normalized length to address the fact that the death counts across different counties can differ by orders of magnitude.

Importantly, the definitions of coverage (equation 4.4a) and the average length (equation 4.4b) are entirely data-driven and do not rely on any probabilistic or generative modeling assumptions.

4.3. **Exchangeability of the normalized prediction errors.** While the ideas from MEPI are a special case of conformal prediction intervals [48, 41], there are some key differences. While conformal inference uses the raw errors in predictions, MEPI uses the normalized errors, and while conformal inference uses a percentile (e.g., the 95th percentile) of the errors, MEPI uses the maximum. Furthermore, we only make use of the previous five days instead of the full sequence of errors. The reason behind these alternate choices is because the validity of prediction intervals constructed in this manner relies crucially on the assumption that the sequence of errors is exchangeable. Our choices are designed to make this assumption more reasonable. Due to the dynamic nature of COVID-19, considering a longer period (e.g., substantially longer than five days) would mean that it is less likely that the errors across the different days are exchangeable. Meanwhile, the normalization of the errors eliminates a potential source of non-exchangeability by removing the sequential growth of the errors resulting from the increasing nature of the counts themselves. Since we only use five time points to construct the interval, the 95th percentile can be approximated by the maximum.

We now provide some empirical evidence that the exchangeability of the past 5 *normalized* errors for CLEP is indeed reasonable for both 7-day-ahead and 14-day-ahead predictions, Figures 5 and 18 in the Appendix, respectively. For k-day-ahead prediction, we rank the errors $\{\Delta_{t+k}, \Delta_t, \Delta_{t-1}, \ldots, \Delta_{t-4}\}$ in increasing order so that the largest error has a rank of 6. (Interested readers may first refer to Section 4.4 for how exchangeability of these 6 errors is useful for establishing theoretical guarantees for MEPI.) For a given $j \in \{0, ..., 4\}$, Δ_{t-j} denotes the error in k-day-ahead prediction for day $t - j$, where the prediction was made on the morning of day $t - j - k + 1$, but the error can be computed only by the end of day $t - j$ (or the morning of day

$t - j + 1$). If the errors were exchangeable, then for each of them, the rank has a uniform distribution on $\{1, 2, 3, 4, 5, 6\}$, and in particular has a mean of 3.5. To approximate this numerically for 7-day-ahead predictions, we measure the rank of the errors Δ_{t+7}, and $\Delta_{t-j}, j = 0, \ldots, 4$, for each day t between March 26 to June 13, and take an average. Figure 5 plots the results for each of the 6 worst hit counties as well as 6 randomly-selected counties for these errors (see Section 5.2 for further discussion on these counties.). Corresponding results for 14-day-ahead predictions, where we rank the errors Δ_{t+14} and $\Delta_{t-j}, j = 0, \ldots, 4$ is presented in Figure 18 in the Appendix. In both figures, we see that across most counties, the average rank for almost all the errors is around 3.5 as would be expected if the errors were exchangeable. Thus, the observations from Figures 5 and 18 provide a heuristic justification for the construction of MEPI, albeit not a formal proof since average rank being close to 3.5 is not sufficient to claim exchangeability of the six errors. Moreover, we refer the interested reader to Appendix B.2 for further discussion on MEPIs, where we provide more evidence on why we chose only past 5 errors and normalization to define MEPI (see Figure 17). Next, we derive a theoretical result for coverage with MEPI under the exchangeability of errors.

4.4. **Theoretical guarantees for MEPI coverage.** In order to obtain a rough baseline coverage for the MEPIs, we now reproduce some of the theoretical computations from the conformal literature. For a given county and a fixed time t, and a parameter k, if the six errors in the set $\{\Delta_{t+k}, \Delta_t, \Delta_{t-1}, \Delta_{t-2}, \Delta_{t-3}, \Delta_{t-4}\}$ are exchangeable, then we have

$$(4.5) \qquad \mathbb{P}\left(y_{t+k} \in \widehat{\mathrm{PI}}_{t+k}\right) = \mathbb{P}\left(\Delta_{t+k} < \Delta_{\max}\right) = 1 - \mathbb{P}\left(\Delta_{t+k} = \Delta_{\max}\right) = \frac{5}{6} \approx 0.83.$$

Recall the definition (equation 4.4a) for Coverage($\mathcal{T}$) for a given period of days $\mathcal{T}$. Given equation 4.5, we may believe that the Coverage($\mathcal{T}$) $\approx 83\%$ holds for large $|\mathcal{T}|$, where the coverage was defined in equation 4.4a. However, we now elaborate that a few challenges remain to take claim (4.5) as a proof for the stronger claim that MEPI achieves 83% coverage as defined by equation 4.4a.

On the one hand, the probability in equation 4.5 is taken over the randomness in the errors, and the time-index $t + k$ remains fixed. This observation, in conjunction with the law of large numbers, implies the following: Over multiple independent runs of the time-series, for a given county and a given time $t + k$, the fraction of runs for which the MEPI $\widehat{\mathrm{PI}}_{t+k}$ contains the observed value y_{t+k} converges to 5/6 as the number of runs goes to infinity. However, analyzing such a fraction over several different independent runs of the COVID-19 outbreak is not relevant for our work.

On the other hand, the evaluation metric we consider is the average coverage of the MEPI over a single run of the time-series, c.f., the definition (equation 4.4a) for Coverage($\mathcal{T}$). Thus, we require an online version of the law of large numbers in order to guarantee that Coverage($\mathcal{T}$) $\to 83\%$ as $|\mathcal{T}| \to \infty$. Such a law of large numbers, established in prior works [41], has been crucial for establishing theoretical guarantees in conformal inference. In our case, this law—stated as Proposition 1 in Section 3.4 in their paper [41]—guarantees that, when the entire sequence of errors $\{\Delta_t, t \in \mathcal{T}\}$ for a given county is exchangeable, the corresponding Coverage($\mathcal{T}$) $\approx 83\%$, when the

period $\mathcal{T}$ is large. Unfortunately, such an assumption is both hard to check and unlikely to hold for the prediction errors obtained from CLEP for the COVID-19 cumulative death counts.

Despite the challenges listed above, later, we show in Section 5.3 that MEPIs with CLEP achieved good coverage with narrow widths for COVID-19 cumulative death count predictions.
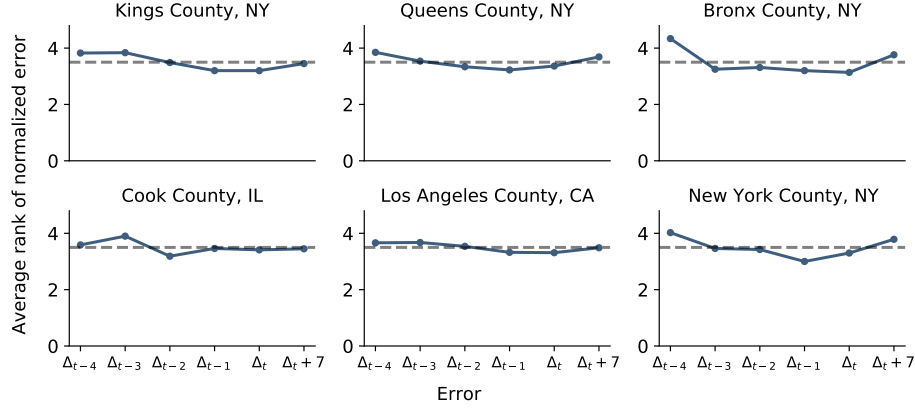
(a) Six worst-affected counties
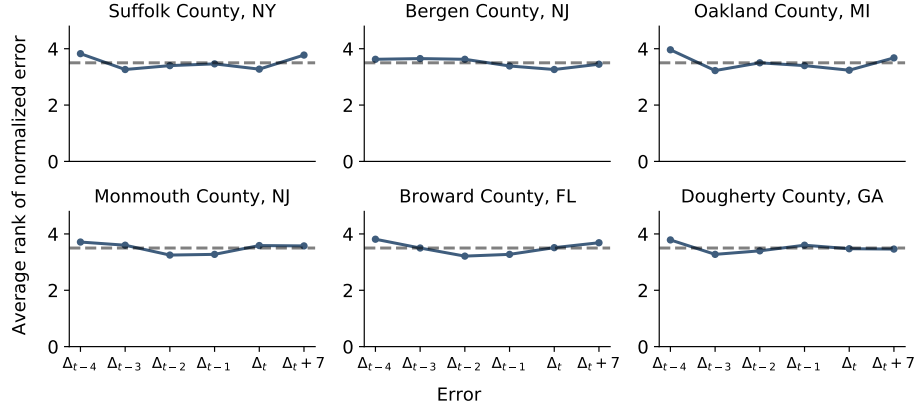
(b) Six randomly-selected counties

Figure 5. EDA plot for investigating exchangeability of normalized errors of *7-day-ahead* CLEP predictions with its last 5 errors made at time t, over the period $t =$ March 26, ..., Jun 13. We plot the average rank of the six errors $\{\Delta_{t+7}, \Delta_t, \Delta_{t-1}, \ldots, \Delta_{t-4}\}$ of our CLEP (with the expanded shared and linear predictors) for **(a)** the six worst affected counties, and **(b)** six random counties. We rank the errors $\{\Delta_{t+7}, \Delta_t, \Delta_{t-1}, \ldots, \Delta_{t-4}\}$ in increasing order so that the largest error has a rank of 6. If $\{\Delta_{t+7}, \Delta_t, \Delta_{t-1}, \ldots, \Delta_{t-4}\}$ are exchangeable for any day t, then the expected average rank for each of the six errors would be 3.5 (dashed black line).

5. Prediction results for March 22 to June 20

In this paper, we focus on predictive accuracy for up to 14 days. In this section, we first present and compare the results of our various predictors, and then give further examinations of the best performing predictor: the CLEP ensemble predictor that combines the expanded shared exponential predictor and the linear predictor (the best two among the individual predictors). Finally, we report the performance of the coverage and length of the MEPIs for this CLEP. (We note that CLEP that combined all five predictors performed worse since even bad performing predictors get non-zero weight (3.7) in the ensemble and can adversely affect the prediction performance.) A Python script to reproduce all the results in this section is made available on Github at `https://github.com/Yu-Group/covid19-severity-prediction/tree/master/modeling`.

5.1. **Empirical performance of the single predictors and CLEP.** Table 6 summarizes the Mean Absolute Errors (MAEs) of our predictions for cumulative recorded deaths on raw, square-root and logarithm scale. We now explain how these errors are computed.

First, on the morning of day $t+1$, we compute $\mathcal{C}_t$—the collection of counties in the US that have at least 10 cumulative recorded deaths by the end of day t. Let $\widehat{y}_t^c$ and y_t^c respectively denote the predicted and recorded cumulative death count of county $c \in \mathcal{C}_t$ by the end of day t. We note that while the set of counties $\mathcal{C}_t$ varies with time, it is computable on the day the error is computed (i.e., $\mathcal{C}_t$ does not depend on future information). We define the set of counties in this manner, to ensure that only the counties with non-trivial cumulative death counts are included in our evaluation on a given day. Moreover this definition satisfies the condition $\mathcal{C}_t \subseteq \mathcal{C}_{t+1}$, that is, only new counties can be added in the set $\mathcal{C}_t$ as t progresses (and a county once included is never removed).

Given the set $\mathcal{C}_t$, the mean absolute percentage error (MAPE), the raw-scale MAE, and the square-root-scale MAE for day t are given by

$$\text{(5.1a)} \qquad \text{MAPE}_t(\% \text{ error}) = 100 \times \frac{1}{|\mathcal{C}_t|} \sum_{c \in \mathcal{C}_t} \frac{|\widehat{y}_t^c - y_t^c|}{y_t^c}, \quad \text{and}$$

$$\text{(5.1b)} \qquad \text{Raw-scale MAE}_t = \frac{1}{|\mathcal{C}_t|} \sum_{c \in \mathcal{C}_t} |\widehat{y}_t^c - y_t^c|,$$

$$\text{(5.1c)} \qquad \text{Sqrt-scale MAE}_t = \frac{1}{|\mathcal{C}_t|} \sum_{c \in \mathcal{C}_t} \left| \sqrt{\widehat{y}_t^c} - \sqrt{y_t^c} \right|.$$

The percentage error (MAPE) captures the relative errors of the predictors without attention to the scale of the counts while the raw-scale MAE would be heavily affected by the counties with large death counts. Both of these MAEs are commonly used to report the prediction performance for regression tasks with machine learning methods. We report the MAE at square-root-scale to be consistent with the square-root transform used in the CLEP weighting scheme (3.7). Each row in Table 6 corresponds to a single predictor, and we report different statistics of errors made for

k-day-ahead predictions over the period $t \in \{$March 22, ..., June 20$\}$ for $k \in \{3, 5, 7, 14\}$.[13] For any given predictor, we compute these errors for each day and report the 10th percentile (p10), 50th percentile (median), and 90th percentile (p90) values over the evaluation periods mentioned above. From Table 6, we find that the CLEP ensemble that combines the expanded shared exponential predictor and the separate county linear predictors has the best overall performance; with median MAPE of 8.18%, 12.21%, 15.14% and 26.45% for 3-, 5-, 7-, and 14-day-ahead predictions.[14] We note that the (p90) MAEs for the separate (exponential) and demographics shared predictors are too large, especially for larger horizons. For separate (exponential) predictors, we can attribute these large errors directly to the fact that exponential fit 3.1 for large horizons is very likely to over-predict. On the other hand, the demographics shared predictor has large errors potentially due to over-fitting, and the recursive plug-in to obtain longer horizon estimates in the exponential fit 3.5. In the early stages of this project (and the COVID-19 outbreak in the US), these predictors had provided a reasonable fit for short-term (3- and 5-day-ahead) predictions in late-March to mid-April.

In Figure 6, all three errors from the display 5.1 as a function of time over the past 3 months for the expanded shared exponential predictor, the separate county linear predictor, and the CLEP that combines the two. We found that the MAE of the CLEP is often similar to, and usually slightly smaller than the smaller MAE of the two single predictors.

Next, in Figure 7, we plot the performance of this CLEP for longer horizons. In particular, we plot the three MAEs, raw-scale, percentage-scale, and square-root-scale for k-day-ahead predictions for $k = 7$, 10, and 14. Notice that the 7-day-ahead CLEP predictor has the lowest MAE and that the MAE increases as the prediction horizon increases. (Recall that precise statistics for 7-day-ahead and 14-day-ahead MAEs are listed in Table 6.) The increases in MAE in mid-late April was caused due to the state of New York adding thousands of deaths (3,778) that were previously reported as "probable" to their counts on a single day, April 14. This change led the CLEP to greatly over-predict deaths in New York in mid-late April—(i) 7 days later on April 21 for the 7-day-ahead CLEP, (ii) 10 days later on April 24 for the 10-day-ahead CLEP, and (iii) 14 days later on April 28 for the 14-day-ahead CLEP. As further evidence that this is indeed the case, when we manually removed this uptick in the death counts in New York, the raw-scale MAE for the

---

[13]As the expanded shared predictor is trained on counties with at least 3 deaths, there was not enough data to train 14-day-ahead CLEP that predicts recorded deaths before March 29. Hence for $t \in \{$March 22, ..., March 28$\}$, we use the 14-day-ahead predictions of the linear predictor to impute the 14-day-ahead predictions of the CLEP.

[14]In the first version of this paper submitted on May 16, 2020 on arXiv, we reported results for the period March 22 to May 10, 2020 for 3-, 5-, 7-day ahead predictions. We made an error while computing aggregate statistics for 5-day and 7-day ahead predictions (reported in Table 3 of that version) and reported errors that were smaller than the actual errors made by CLEP. Correcting our aggregation code revealed that the actual performance of median of raw-scale MAE of CLEP for 5-day and 7-day predictions was worse by 16% and 29% respectively when compared to the reported errors.

(a) MAPE   (b) Raw-scale MAE   (c) Square-root-scale MAE
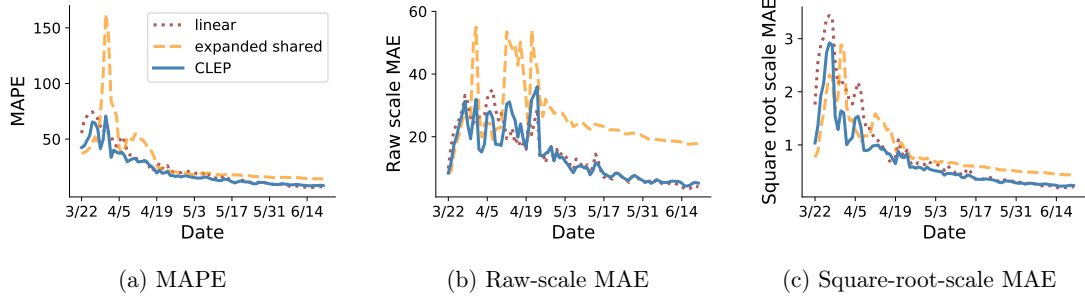
Figure 6. Plots of mean absolute error (MAE) of different predictors for *7-day-ahead* predictions from March 22 to June 20. We plot the **(a)** mean absolute percentage error (MAPE), **(b)** raw-scale MAE, and **(c)** square-root-scale MAE versus time. Results are shown for expanded shared exponential predictor (orange dashed line), the separate county linear predictor (red dotted line), and the CLEP that combines the two predictors (solid blue line).

(a) MAPE   (b) Raw-scale MAE   (c) Square-root-scale MAE
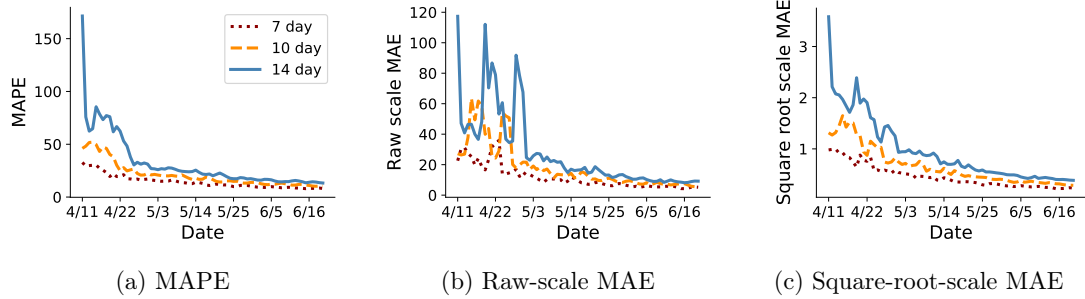
Figure 7. Plots of mean absolute error (MAE) of CLEP with different prediction horizons from April 11, 2020 to June 20, 2020. We plot the **(a)** mean absolute percentage error (MAPE), **(b)** raw-scale MAE, and **(c)** square-root-scale MAE versus time, for k-day-ahead predictions with $k \in \{7, 10, 14\}$.

14-day-ahead CLEP on April 28 was 29.5, which is much smaller than the original raw-scale MAE on April 28, which was 91.9 in Figure 7(b).

We further evaluate the performance of CLEP for longer prediction horizons in Figure 8, where all predictions were made over the period April 11-June 20. In panels (a-c) of Figure 8, we show the box plots of the different MAEs for up to 14-day prediction horizon. From these plots, and the panels (d-f), we observe that the MAEs degrade roughly linearly with the horizon for up to 21 days.

Putting together the results from Table 6, Figures 6, 7 and 8, we find that the adaptive combination used for building our ensemble predictor CLEP is able to leverage the advantages of linear and exponential predictors, and—by improving upon the MAE of single predictors—it is able to provide very good predictive performance for up to 14 days in future.

	3-day-ahead			5-day-ahead			7-day-ahead			14-day-ahead		
	p10	median	p90	p10	median	p90	p10	median	p90	p10	median	p90
separate	3.80	13.16	59.63	6.26	22.56	114.07	9.95	39.56	300.53	30.37	226.26	>1000
shared	7.05	12.55	25.99	11.68	19.77	37.73	16.59	28.65	55.01	36.55	62.45	224.75
demographics	17.82	25.70	30.90	30.30	41.02	50.62	47.77	62.26	117.11	260.48	551.78	>1000
expanded shared	6.86	9.59	35.55	11.17	14.54	44.28	15.09	18.52	52.13	23.13	31.18	>1000
linear	**3.39**	9.37	29.67	**5.27**	14.25	40.26	**7.18**	18.60	56.10	15.58	33.16	**87.21**
CLEP	4.34	**8.18**	**22.60**	6.59	**12.21**	**31.99**	8.79	**15.14**	**42.47**	**14.61**	**26.45**	93.03

(A) Summary statistics of mean absolute percentage error (MAPE)

	3-day-ahead			5-day-ahead			7-day-ahead			14-day-ahead		
	p10	median	p90	p10	median	p90	p10	median	p90	p10	median	p90
separate	2.35	8.10	25.13	**3.67**	13.94	57.03	5.33	24.30	124.61	14.58	105.64	>1000
shared	7.54	12.04	19.43	13.12	19.93	36.74	18.81	28.09	72.74	33.69	69.35	325.50
demographics	17.47	48.35	54.54	35.41	108.47	119.71	59.29	217.64	243.56	675.96	>1000	>1000
expanded shared	8.07	10.69	14.32	13.10	16.68	23.02	18.24	22.95	42.84	29.90	36.56	329.21
linear	**2.15**	**5.93**	13.81	3.67	9.49	20.02	**4.91**	12.05	**26.89**	10.24	25.47	**56.73**
CLEP	2.76	5.98	**11.93**	4.09	**8.64**	**18.67**	5.42	**10.64**	27.29	**9.18**	**22.50**	81.77

(B) Summary statistics of raw-scale MAE

	3-day-ahead			5-day-ahead			7-day-ahead			14-day-ahead		
	p10	median	p90	p10	median	p90	p10	median	p90	p10	median	p90
separate	0.11	0.39	1.66	0.19	0.63	2.91	0.25	1.03	3.83	0.67	3.40	18.26
shared	0.22	0.40	0.76	0.36	0.63	1.22	0.50	0.90	1.82	1.09	2.06	5.70
demographics	0.81	1.11	1.38	1.27	2.04	2.48	2.02	3.25	3.90	6.61	13.45	71.60
expanded shared	0.22	0.34	0.75	0.35	0.52	1.15	0.47	0.67	1.59	0.73	1.12	10.62
linear	**0.11**	0.29	0.99	**0.17**	0.46	1.45	**0.23**	0.58	2.15	0.50	1.10	4.62
ensemble	0.13	**0.26**	**0.66**	0.19	**0.37**	**0.93**	0.26	**0.47**	**1.51**	**0.43**	**0.92**	**4.13**

(C) Summary statistics of sqrt-scale MAE

Table 6. Summary statistics of Mean Absolute Errors (equation 5.1) based on **(A)** the mean absolute percentage error (MAPE), **(B)** the raw-scale MAE, and **(C)** the square-root-scale MAE. The results are presented for the 3, 5, 7, and 14-days-ahead forecasts for each of the predictors considered in this paper, and the CLEP that combines the expanded shared and separate linear predictors. The evaluation period is March 22, 2020 to June 20, 2020 (91 days). "p10", "median", and "p90" denote the 10th-percentile, median, and 90th-percentile of the 91 mean absolute errors computed daily in the evaluation period. The smallest error in each column is displayed in bold.

5.2. **Performance of CLEP and MEPI at the county-level.** Having examined the overall performance of our predictors, we now take a closer look at how our predictors are performing at the county level. In this section, we focus on the performance of our CLEP predictor (based on the best-performing CLEP of the expanded shared and separate linear predictor models) for the period
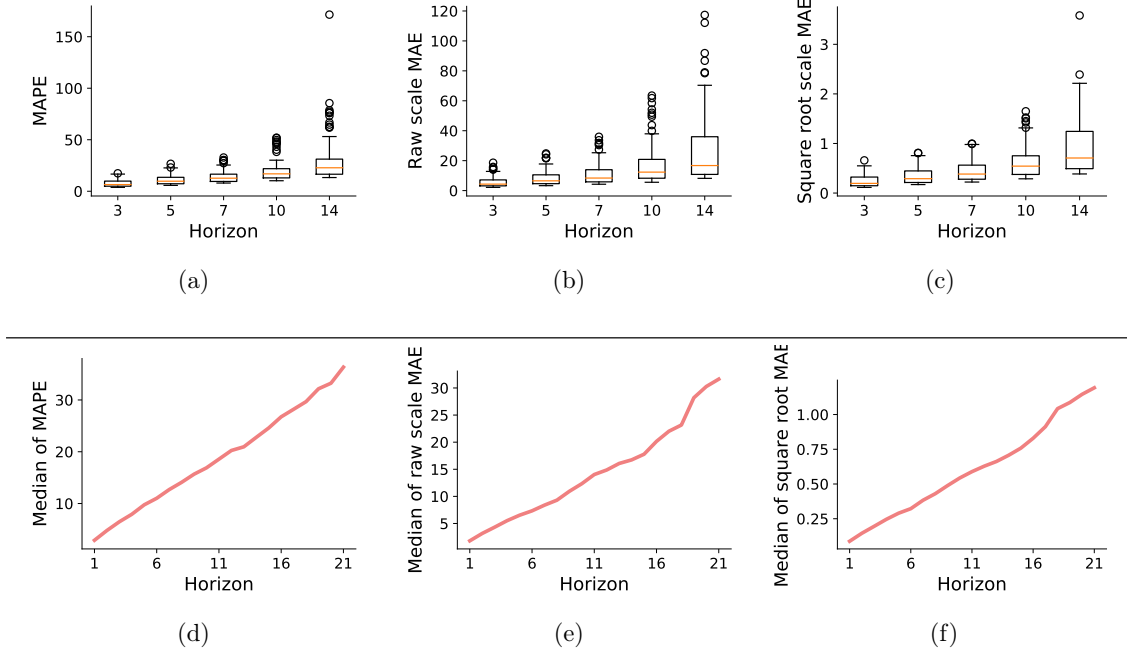
Figure 8. Plots of MAE distribution of CLEP for different prediction horizon over the period April 11, 2020 to June 20, 2020. In panels **(a)**, **(b)** and **(c)**, we show the box-plots (over time) of the mean absolute percentage error (MAPE), raw-scale MAE, and square-root-scale MAE for k-day ahead predictions $k = \{3, 5, 7, 10, 14\}$. In panels **(d)**, **(e)** and **(f)**, we plot the median value of the different MAEs for even longer horizons, up to 21 days.

April 11 to June 20 for 7-day and 14-day-ahead predictions (Figures 9 and 10 respectively). Since there are over 3,000 counties in the United States, we present results for two sets of counties:

- The six worst-affected counties on June 20: Kings County, NY; Queens County, NY; Bronx County, NY; Cook County, IL; Los Angeles County, CA; and New York County, NY.
- Six randomly-selected counties: Suffolk County, NY; Bergen County, NY; Oakland County, MI; Monmouth County, NJ; Broward County, FL; Dougherty County, GA.

In Figure 9, we present the 7-day-ahead prediction results for the worst-affected counties in the top panel (a), and for the randomly-selected counties in the bottom panel (b). The solid black line denotes the recorded death counts, dashed blue line denotes the CLEP 7-day-ahead predictions, and shaded blue region denotes the corresponding MEPI (prediction interval). The predictions and prediction intervals for a given day t ($t =$ April 11, ..., June 20) are based on data up to day $t - 7$. Corresponding results for 14-day-ahead predictions are plotted in Figure 10. Although the recorded cumulative death counts are monotonically increasing, our predictions for it (blue dashed lines in Figures 9 and 10) need not be since the predictions are updated daily. (Recall from Section 3.7 that we did impose monotonicity constraints for predictions with respect to k for k-day-ahead predictions made on a given day t.)

From Figure 9(a), we observe that, among the worst-affected counties, the CLEP appears to fit the data very well for Cook County, IL, and Los Angeles County, CA. After initially over-predicting the number of deaths in the four NY counties in mid-late April, our predictor also performs very well on these NY counties. Moreover, the MEPIs—plotted as blue shaded region—have reasonable width and appear to cover the recorded values for the majority of the days (detailed results on MEPIs are presented in Section 5.3).

The rapid increase in our predictions and MEPI interval widths observed in mid-late April in Figure 9(a) is caused due to a major upward revision of cumulative death counts by the New York state for Kings, Queens, Bronx, and New York counties on a single day, April 14. The effect of this sudden change in recorded counts on CLEP performance was also discussed in the previous section in the context of Figure 7. Our predictors quickly corrected themselves after the recorded counts stabilized.

From Figure 9(b), we find that our predictors and MEPI perform well for each of our six randomly-selected counties (Broward County, FL, Dougherty County, GA, Monmouth County, NJ, Bergen County, NJ, and Oakland County, MI). However, notice again that for Suffolk County, NY, there is a sudden uptick in cumulative deaths on May 5, leading to a fluctuation in the predictions shortly thereafter, just like in the other NY counties in mid-April.

In both panels, our predictions have higher uncertainty at the beginning of the examination period when recorded death numbers are low, which is reflected in the slightly wider MEPIs in the bottom left of each plot.

Next, in Figure 10, we plot the visualizations for 14-day-ahead CLEP and MEPI predictions for the two sets of counties discussed above. On the one hand, overall, the CLEP predictions appear to track the cumulative death counts and MEPIs cover the observed death counts most of the times. On the other hand, we find that the MEPIs are pretty wide for all the counties in the beginning of the prediction period. This phenomenon is caused due to the fact that till mid-April, predictions made 14 days ago were trained on data with very few death counts, which led the expanded shared predictor to significantly underestimate the death counts in the beginning period (leading to larger normalized errors and thereby wider MEPIs). Furthermore, we also observe that for the counties in the NY State, CLEP greatly overestimates (sharp peak around April 22) the cumulative recorded death counts towards the end of April. Like in Figure 9, this overestimation is caused due to the major upward revision of the death counts in the NY State on April 14. It is worth noting that these sharp peaks for 14-day-ahead CLEP predictions in Figure 10 occurs after 7 days of the sharp peaks observed for 7-day-ahead CLEP predictions in Figure 9.

5.3. **Empirical performance of MEPI.** We now present the performance of our MEPI at the county level for cumulative death counts, with respect to both coverage (equation 4.4a) and average normalized length (equation 4.4b) (see Section 4.2). Since the average performance may change with time, we report the results for two time periods: {April 11, ..., May 10}, and {May 11, ..., June 20}. We choose these time periods for a simple reason: The first draft of this paper presented results
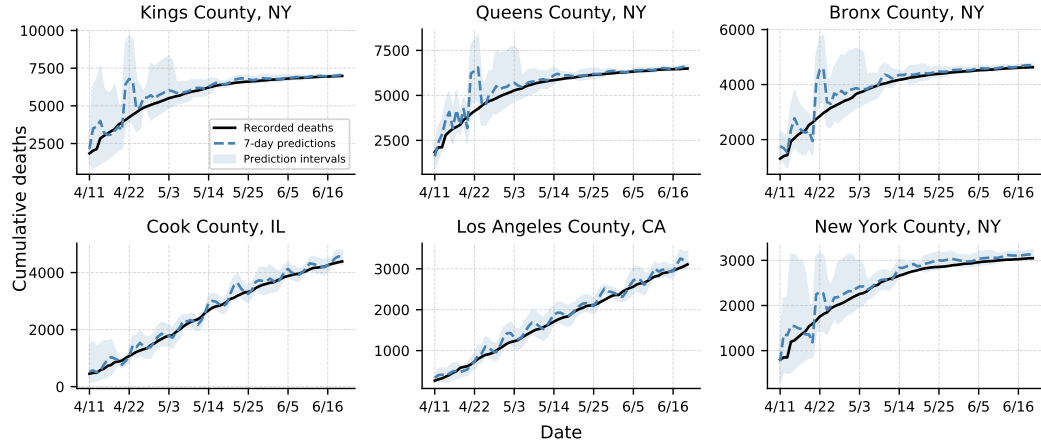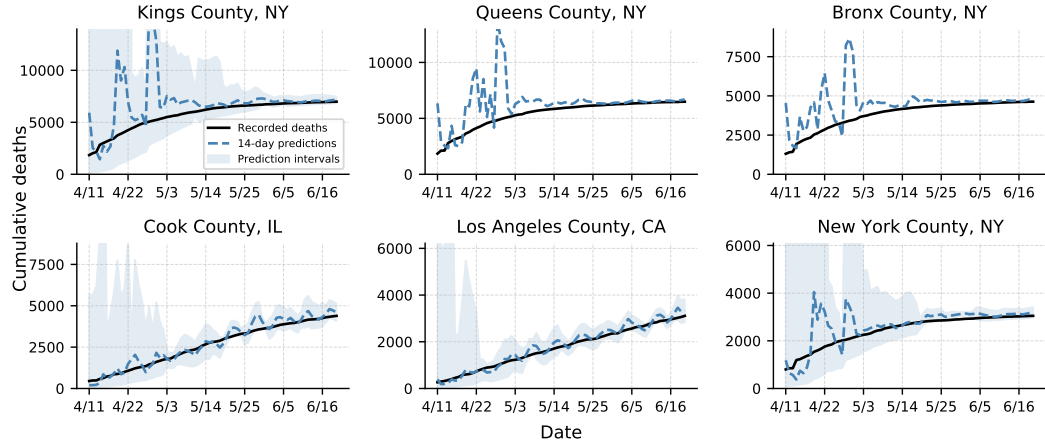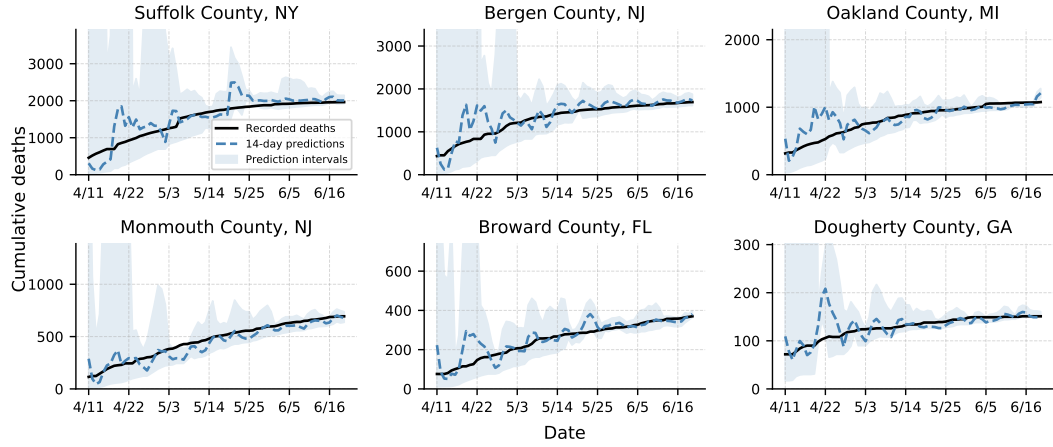
(a) Worst-affected counties on June 20
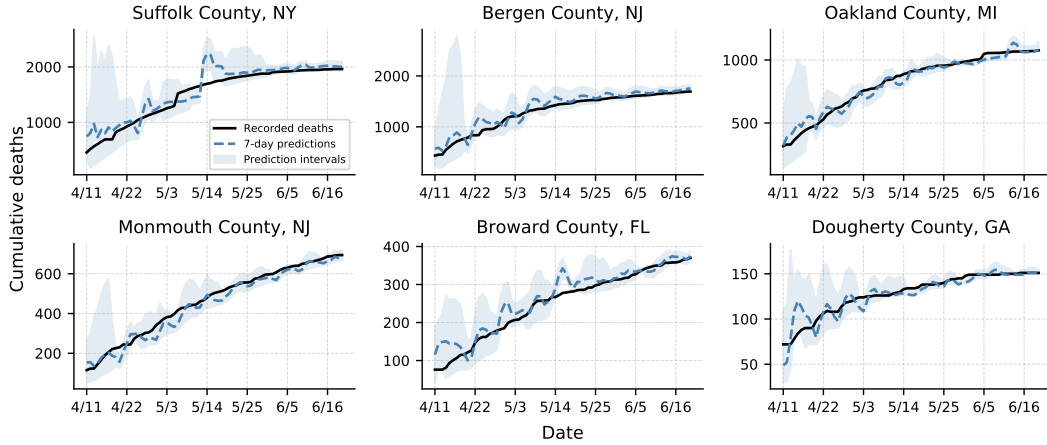
(b) Randomly-selected counties

Figure 9. A grid of line charts displaying the performance of *7-day-ahead* CLEP and MEPI for the cumulative death counts due to COVID-19 between April 11, 2020 and June 20, 2020. The observed data is shown in black, CLEP predictions are shown in the dashed blue, and the corresponding 7-day-ahead MEPIs are shown as shaded blue regions. The prediction and prediction intervals for day t (t = April 11, ..., June 20) are based on data up to day $t - 7$. In panel **(a)**, the MEPI coverage for the 6 counties are Kings (92%), Queens (80%), Bronx (90%), Cook (90%), Los Angeles (89%) and New York (86%). In panel **(b)**, Suffolk (85%), Bergen (96%), Oakland (87%), Monmouth (86%), Broward (89%) and Dougherty county (86%). (It is worth noting that the theoretical coverage guarantees under certain assumptions is 83%; see equation 4.5.)

(a) Worst-affected counties on June 20

(b) Randomly-selected counties

Figure 10. A grid of line charts displaying the performance of *14-day-ahead* CLEP and MEPI for the cumulative death counts due to COVID-19 between April 11, 2020 and June 20, 2020. The observed data is shown in black, CLEP predictions are shown in the dashed blue, and the corresponding 14-day-ahead MEPIs are shown as shaded blue regions. The prediction and prediction intervals for day t (t = April 11, . . . , June 20) are based on data up to day $t - 14$. In panel **(a)**, the MEPI coverage for the 6 counties are Kings (99%), Queens (99%), Bronx (99%), Cook (93%), Los Angeles (96%) and New York (89%). In panel **(b)**, Suffolk (92%), Bergen (99%), Oakland (89%), Monmouth (94%), Broward (97%) and Dougherty county (94%). Note that for these counties, the coverage of 14-day-ahead MEPIs is higher than that of 7-day-ahead MEPIs (shown in Figure 9) due to the wider intervals in the beginning of the period. (It is worth noting that the theoretical coverage guarantees under certain assumptions is 83%; see equation 4.5.)

until May 10, and thus the period May 11 to June 20 serves as additional out-of-sample validation. (Our methods were proposed and tuned prior to May 10, except for the change of transform from logarithm to square-root in equation 3.7.)

We evaluate the 7-day-ahead and 14-day-ahead MEPIs, i.e., $k = 7$ and 14 in equation 4.2a, designed with the CLEP that combines the expanded shared and separate linear predictors, and summarize the results in Figure 11 and Figure 12 respectively. We now discuss these results, first for 7-day-ahead and then for 14-day-ahead MEPIs.

Coverage: In panels (a) and (c) of Figure 11, we plot the histogram of observed coverage (equation 4.4a) of 7-day-ahead MEPIs across all counties in the US for April 11–May 10, and May 11–June 20 respectively. Panel (e) of Figure 11 shows the observed coverage of 7-day-ahead MEPIs for the 700 counties that had at least 10 deaths on June 11 (each such county has had significant counts for at least 10 days by the end of our evaluation period June 20). For each county, we include the days starting from either April 11, 2020 or the day of 10 deaths (if this day is after April 11) until June 20, 2020. On June 20, the median number of days since 10 deaths is 58. From these plots, we observe that for the majority of the counties, we achieve excellent coverage. Finally, Figure 12 shows the corresponding results for 14-day-ahead MEPIs, i.e., coverage for April 11–May 10 in panel (a), May 11–June 20 in panel (c), and over the county-specific period for 700 counties that had at least 10 deaths on June 11 in panel (e).

The coverage observed for the two periods in panels (a) and (c) of Figure 11 are very similar. For the earlier period—April 11 to May 10—in panel (a), the 7-day-ahead MEPIs have a median coverage of 100% and mean coverage of 95.6%. On the other hand, for the later period—May 11 to June 20—in panel (c), the 7-day-ahead MEPIs have a median coverage of 100% and mean coverage of 96.2%. However, the coverage is slightly decreased when restricting to counties with at least 10 deaths in panel (e), for which we observe a median coverage of 88.7%, and mean coverage of 87.9%. This observation is consistent with the fact that at the beginning of the pandemic, several counties had zero or very few deaths resulting in very good coverage with the prediction interval. On the other hand, note that smaller death counts would also imply a relatively larger normalized length for the MEPI intervals.

Figure 12(a), (c) and (e) show that, in general, our 14-day-ahead MEPIs achieve similar coverage as our 7-day-ahead MEPIs. For example, over the April 11–May 10 and May 11-June 20 periods, our prediction intervals have mean coverage of 95.0% and 97.0% (median is 100% for both periods). For panel (e)—counties with at least 10 deaths on June 11—the coverage has a median of 89.7% and a mean of 87.9%.

Overall, the statistics discussed above show that both 7-day-ahead and 14-day-ahead MEPIs achieve excellent coverage in practice. In fact, for the counties with poor coverage, we show in Appendix B.1 that there is usually a sharp uptick in the number of recorded deaths at some point in the evaluation period, possibly due to recording errors, or backlogs of counts. Modeling these

upticks and obtaining coverage for such events is beyond the scope of this paper.

Normalized length: Next, we discuss the other evaluation metric of the MEPIs, their normalized length (4.4b). In panels (b) and (d) of Figure 11, we plot the histogram of the observed average normalized length of 7-day-ahead MEPIs for the periods April 11–May 10, and May 11–June 20 respectively. Panel (f) covers the same counties as did panel (e): those with at least 10 deaths for at least 10 days in the period April 11 to June 20, 2020.

Recall that the normalized length is defined as the length of the MEPI over the recorded number of deaths (equation 4.4b). And, more than 70% of counties in the US recorded 2 or less COVID-19 deaths by May 1. For these counties, having a normalized length of 2 means the actual length of the prediction interval is 4 (or less). And thus, it is not surprising to see that the average normalized length of MEPI for a non-trivial fraction of counties is larger than 2 in panels (b) and (d). When considering counties with at least 10 deaths in panel (f), the average normalized length over these (county-specific) periods is much smaller; and the median is 0.470.

Turning to 14-day-ahead MEPIs in Figure 12, panels (b) and (d) show that that the normalized length for 14-day-ahead MEPIs can be quite wide for counties with a small number of deaths. Nevertheless, panel (f) shows that the 14-day-ahead MEPIs are reasonably narrow for counties with more than 10 deaths, with a median average normalized length of 1.027—which is roughly two times the median size of 0.470 for the 7-day-ahead MEPIs in Figure 11(f).

Overall, Figures 11 and 12 show that our MEPIs provide a reasonable balance between coverage and length, especially when the cumulative counts are not too small, for up to 14 days in future.

6. Related work

Several recent works have tried to predict the number of cases and deaths related to COVID-19. Even more recently, the Center for Disease Control and Prevention (CDC) has started aggregating results from several models.[15] But to the best of our knowledge, ours was the first work focusing on predictions at the county-level. Besides, directly comparing other models' forecasting results to our own can be difficult for several other reasons: (1) the predictors mostly make strong assumptions and typically do not involve data-fitting, (2) we do not have access to a direct implementation of their predictors (or results), and (3) their predictors focus on substantially longer time horizons. Keeping these points in mind, we now provide a brief summary of recent work on predictive modeling for COVID-19.
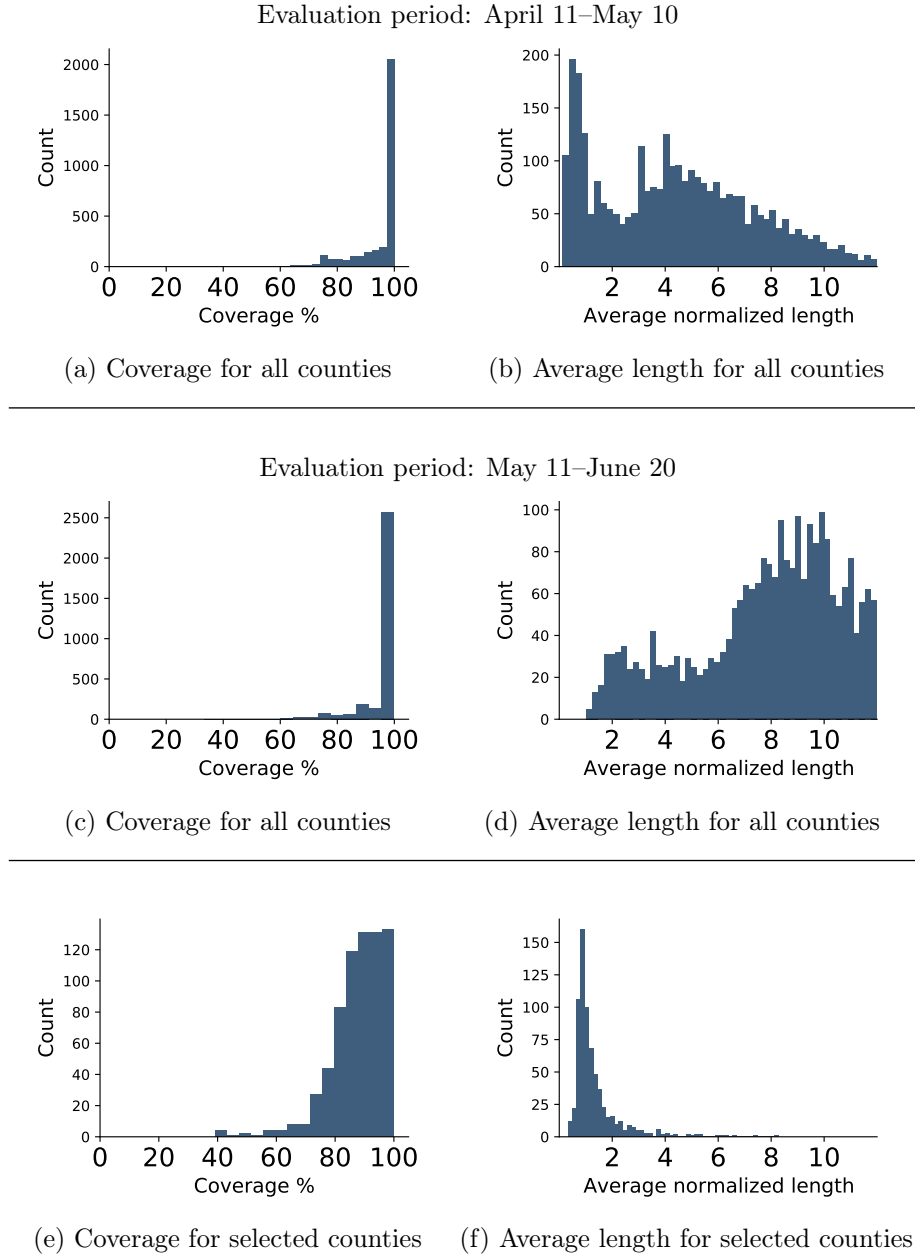
Two recent works [33, 18] have modeled the death counts at the state level in the US. The earlier versions of the model by Murray et al. (also referred to as the IHME model) was based on Farr's Law with feedback from Wuhan data. On the other hand, the Imperial College model [18] uses an individual-based simulation models with parameters chosen based on prior knowledge. On the topic

---

[15]Forecasts available at https://www.cdc.gov/coronavirus/2019-ncov/covid-data/forecasting-us.html

Evaluation period: April 11–May 10

(a) Coverage for all counties

(b) Average length for all counties

Evaluation period: May 11–June 20

(c) Coverage for all counties

(d) Average length for all counties

(e) Coverage for selected counties

(f) Average length for selected counties
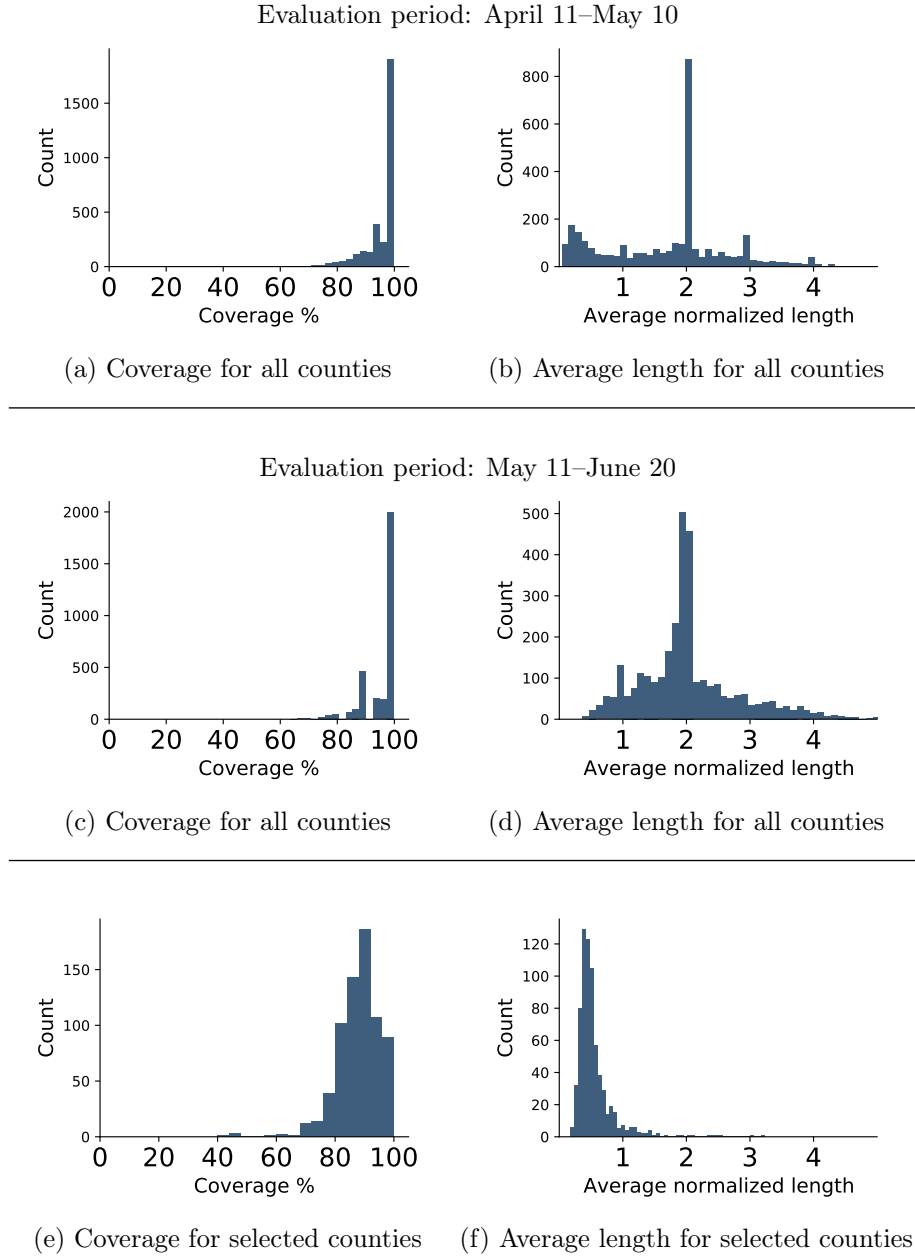
Figure 11. Histograms showing the performance of *7-day-ahead MEPI intervals* for county-level cumulative death counts. For the top panels **(a)** and **(b)**, we compute the histogram for all counties spanning April 11–May 10, 2020 and for the middle panels **(c)** and **(d)**, spanning May 11–June 20, 2020. For the bottom two panels **(e)** and **(f)**, we only include counties that had at least 10 cumulative deaths by June 11, and the histogram is based on the county-specific periods, over the days April 11 to June 20, which only includes days for which the county's cumulative deaths is at least 10. See Figure 12 for similar plots for 14-day-ahead MEPIs.

Evaluation period: April 11–May 10

(a) Coverage for all counties   (b) Average length for all counties

Evaluation period: May 11–June 20

(c) Coverage for all counties   (d) Average length for all counties

(e) Coverage for selected counties   (f) Average length for selected counties

Figure 12. Histograms showing the performance of the *14-day-ahead MEPI intervals* for county-level cumulative death counts. Panels **(a)** and **(b)** present histograms of coverage and average normalized length, respectively, for all counties over April 11–May 10. Panels **(c)** and **(d)** present histograms of coverage and average normalized length, respectively, for all counties over May 11–June 20. Panels **(e)** and **(f)** present histograms of coverage and average normalized length, respectively, only for counties that had at least 10 cumulative deaths by June 11, and the coverage and lengths are computed over the county-specific periods, over the days April 11 to June 20, for which the county's cumulative death count is at least 10. See Figure 11 for similar plots for 7-day-ahead MEPIs.

of Farr's Law, we note that Bregman [4] used Farr's Law to predict that the total cases from the AIDS pandemic would diminish by the mid-1990s and the total number of cases would be around 200,000 in the entire lifetime of the AIDS pandemic. It is now estimated that 32 million people have died from the disease so far. While the AIDS pandemic is very different from the COVID-19 pandemic, it is still useful to keep this historical performance in mind.

Another approach uses exponential smoothing from time-series predictors to estimate day-level COVID-19 cases [16]. In addition, several works use compartmental epidemiological predictors such as SIR, SEIR, and SIRD [17, 36, 3] to provide simulations at the national level. Other works [35, 26] simulate the effect of social distancing policies either in the future for the US, or in a retrospective manner for China. Finally, several papers estimate epidemiological parameters retrospectively based on data from China [49, 30].

During the revision of our paper, another work was published by Chiang et al. [12] that appeared in medRxiv on June 8, 2020 [16] after the submission of our paper to arXiv in mid-May. Chiang et al. use models based on Hawkes' process to provide county-level predictions for new daily cases as well as new death counts. Of note is that the authors also explore a CLEP with adaptive tuning of c and μ (whereas we used fixed values for these parameters). Such a tuning approach might present a promising improvement of CLEP performance in general and we plan to investigate adaptive tuning of various hyper-parameters in the CLEP in our own future work. Unfortunately, we were unable to properly reproduce the CLEP results provided in Chiang et al.'s work using their provided documentation. During a private email exchange the authors kindly provided further information regarding some of our questions about their methodology[17], but several of their choices make it difficult to compare their work to ours. For instance, their work focuses on daily counts, rather than cumulative counts as ours does. More importantly, their prediction numbers are not available on their GitHub repository (both for the period till May 20 analyzed in their paper and the days since then). The authors do not report the performance of their confidence intervals in the paper, and report the MAE performance metric only for the counties that fall in the top quantiles of cumulative counts at the end of the evaluation period. Such a quantile-based group of counties is not interpretable (since it is time-varying and not spatially meaningful) and does not allow for real-time use, since one must wait until the end of the evaluation period to calculate the performance. In addition, the authors compute their predictions in blocks of days, e.g., once a week for the 7-day-ahead predictions (rather than daily as in this paper). Thus, from our point of view, these decisions unfortunately make their work ill-suited to real-time usage for making fast-paced policy decisions related to COVID-19.

---

[16] Accessed at https://www.medrxiv.org/content/10.1101/2020.06.06.20124149v1.full.pdf

[17] In particular, via this email exchange we learned that: (i) they had implemented the adaptive tuning of CLEP; and (ii) they had computed the % error (in Table S1 of their paper) for total new counts over the entire k-day-block (for k-day-ahead predictions) summed over all the counties in a given quantile thereby explaining the (surprising at first) decrease in % error as the prediction horizon k increases.

7. Impact: a hospital-level severity index for distributing medical supplies

Together with the non-profit Response4Life[18], our models have been used to determine which hospitals are most urgently in need of medical supplies, and have subsequently been directly involved in the distribution of medical supplies across the country. To do this, we translate our forecasts into the COVID pandemic severity index, which is a simple measure of the COVID-19 outbreak severity for each hospital.

To generate this hospital-level severity index, we divided the total county-level deaths among all of the hospitals in the county proportional to their number of employees. Next, for each hospital, we computed its percentile among all US hospitals with respect to total deaths so far and also with respect to predicted new deaths in the next seven days. These two percentiles are then averaged to obtain a single score for each hospital. Finally, this score is quantized evenly into three categories: low, medium, and high severity. Evaluation and refinement of this index are ongoing, as more hospital-level data becomes available. The interested reader can find a daily-updated map of the COVID pandemic severity index and additional hospital-level data at our website https://covidseverity.com.

8. Conclusion

In this paper, we made three key contributions. We (1) introduced a data repository containing COVID-19-related information from a variety of public sources, (2) used this data to develop CLEP predictors for short-term forecasting at the county level (up to 14 days), and (3) introduced a novel yet simple method MEPI for producing prediction intervals for these predictors. By focusing on county-level predictions, our forecasts are at a finer geographic resolution than those from a majority of other relevant studies. By comparing our predictions to real observed data, we found that our predictions are accurate and that our prediction intervals are reasonably narrow and yet provide good coverage. We hope that these results will be useful for individuals, businesses, and policymakers to plan and cope with the COVID-19 pandemic and that our data repository and forecasting and interval methodology will be useful for academic purposes. Indeed, our results are already being used to determine the hospital-level need for medical supplies and have been directly influential in determining the distribution of these supplies.

Our data repository will be useful both for educational purposes, as well as for other teams interested in analyzing the data underlying the COVID-19 pandemic. Our CLEP ensembling techniques and MEPI methodology can be applied to other models for COVID-19 forecasting, as well as to online methods and time-series analysis more broadly. Our data, codes and models can be found at https://covidseverity.com.

Last but not the least, inspired by the recent work [12], we are beginning our investigation into adaptive tuning (over time) of μ, c and other hyperparameters for CLEP, in the hope to improve its performance.

[18]https://response4life.org/

Acknowledgements

Disclosure Statement. The authors have no conflicts of interest to declare.

Appendix A. Predictors with additional features

We now describe a few additional features that were considered to potentially improve our predictors (but did not lead to any significant improvements). We included these features after our first submission (on May 16, 2020), and hence tried the new features only in the context of the CLEP that combines the expanded shared and linear predictors.[19]

---

[19]We note that the expanded shared predictor in this appendix is implemented without the monotonicity adjustment (discussed in Section 3.7). The linear predictor does not need such adjustments in our setting. Since our attempts with new features considered in this appendix did not lead to any improvement, we did not re-do the investigation with the monotonicity adjustment. We leave any further investigation with these new features (or their variants) for future work.
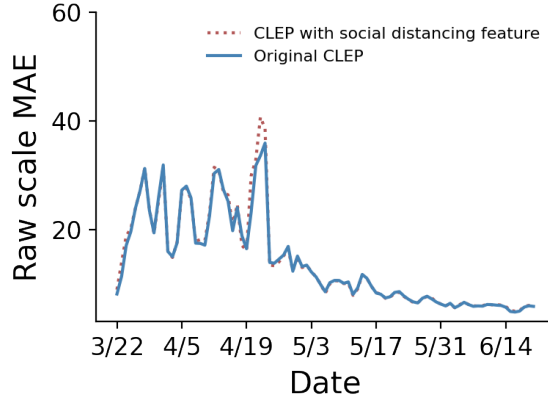
Figure 13. Plots of raw-scale MAE for *7-day-ahead* predictions of two variants of CLEP combining expanded shared and linear predictors: CLEP with a social distancing indicator feature for whether social distancing was in place in a county for more than two weeks or not, and the original CLEP considered in the main paper. The social distancing feature is included in the expanded shared model.

A.1. **Social-distancing feature.** Here we consider adding a social distancing feature to the expanded shared model (discussed in Section 3.4). We included an indicator feature in equation 3.4 for every county that takes value 1 on a day if at least two weeks have passed since social distancing was first instituted in a county, and 0 otherwise. We chose two weeks as the time lag to account for the two week progression time for the illness to the recovery of the COVID-19. We found it necessary to regularize this predictor since, without regularization, our 7-day-ahead predictions became infinite in some cases. We regularized this model with the elastic net and an equal penalty of .01 for both ℓ_1 and ℓ_2 regularization.

We now evaluate the performance of the two variants of the expanded shared model by ensembling each of them with the separate linear model as done for the original CLEP in the main paper. In particular, we compare CLEP with the social-distancing feature included in the expanded shared predictor, with the original CLEP from the main paper, for 7-day-ahead prediction of the recorded cumulative death counts. We found that the new variant (with the social distancing feature) performed slightly worse than our original CLEP. Over the period March 22 to June 20, the original CLEP has a mean (over time) raw-scale MAE (equation 5.1) of 13.95, while the social-distancing variant has an MAE of 14.2. In Figure 13, we plot the behavior of raw-scale MAEs with time for the evaluation period from March 22 to June 20. We observe that the performance of the new CLEP variant is similar to that of the original CLEP, with the exception of a couple of the peaks, where the new CLEP variant performs slightly worse.

A.2. **Weekday feature.** As illustrated in Section 2.2 and Figure 3(a), the COVID-19 death counts are under-reported on Sunday and Monday, which could potentially lead to increased errors for our prediction algorithm.
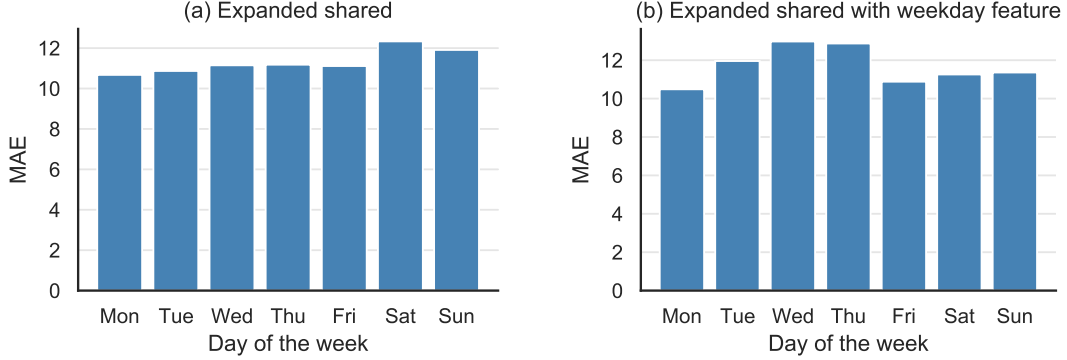
Figure 14. Mean raw-scale MAEs by weekdays for the expanded shared predictor for *3-day-ahead* predictions with and without the weekday feature. The MAE for a given day is the MAE for the 3-day-ahead prediction *computed on that day.* So the MAE for Wednesday is the MAE for predicting the cumulative deaths on Friday.

To address this, we first investigated our expanded shared predictor's (equation 3.4) performance for 3-day-ahead prediction on a per weekday basis and plotted them in Figure 14(a). We can observe the average raw-scale MAE is slightly higher for the days when the 3-day-ahead period included both Sunday and Monday. For example, 3-day-ahead predictions made on Saturday would require making predictions for Saturday, Sunday, and Monday, and that made on Sunday would require making predictions for Sunday, Monday, and Tuesday.

To help account for this bias, we introduced an additional indicator feature in equation 3.4 that takes a value of 1 when the day—for which the prediction is made—is either Sunday or Monday, and 0 otherwise. For instance, when we make 3-day-ahead predictions on Saturday, this feature would take value 0 while computing the prediction for Saturday, and 1 when we compute the prediction for Sunday and Monday. We plot the error distribution over days of this new variant in Figure 14(b). For the new variant, we find that the raw-scale MAEs for the days, when the 3-day-ahead period does not include both Sunday and Monday, typically have higher MAE. Overall when averaging across all days for March 22 to June 20, we find that the new variant of expanded shared predictor performed slightly worse than the original version. The raw-scale MAE (5.1b) for the new variant is 11.7, while the original variant had a raw-scale MAE of 11.5.

Next, we experiment with adding a weekday feature to the separate linear predictors (discussed in Section 3.2) for 3-day-ahead predictions, by adding a binary feature that takes value 1 if the day—for which the prediction is made—is either Sunday or Monday, and 0 otherwise. Thus, the new variant of the separate linear predictors is given by

$$(A.1) \qquad \qquad \widehat{\mathrm{E}}[\mathrm{deaths}_{t+1}^c|t] = \beta_0^c + \beta_1^c(t+1) + \beta_2^c v_{t+1},$$

where v_{t+1} indicates whether day $t+1$ is Sunday/Monday or not. For 3-day-ahead predictions ($\widehat{\mathrm{E}}[\mathrm{deaths}_{t+3}^c|t]$), we simply replace $t+1$ by $t+3$, and v_{t+1} by v_{t+3} on the RHS of equation A.1.
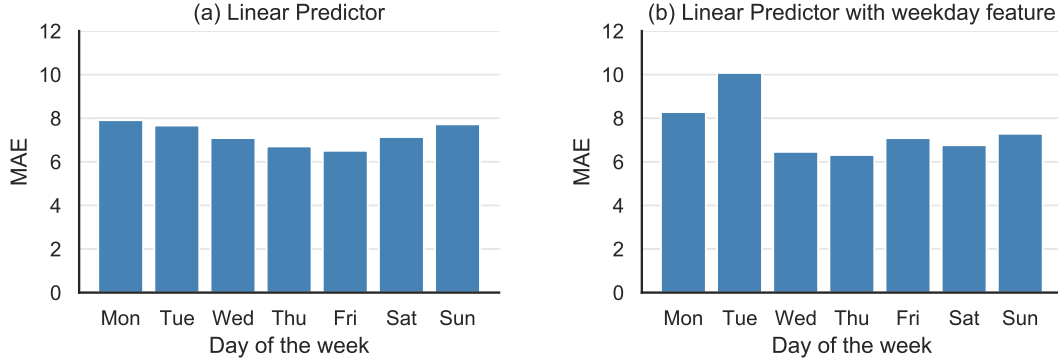
Figure 15. Mean raw-scale MAEs by weekdays for the separate linear predictors for *3-day-ahead* predictions with and without the weekday feature. The MAE for a given day is the MAE for the 3-day-ahead prediction *computed on that day.* So the MAE for Wednesday is the MAE for predicting the cumulative deaths on Friday.

Recall that the original separate linear predictors in Section 3.2 were fit only with the four most recent days data. For some choices of days, the new feature v_t takes only a single value 0 in the training data. For instance, when day $t+1$ is Saturday, we have $v_t = v_{t-1} = v_{t-1} = v_{t-2} = v_{t-3} = 0$, i.e., the new feature is identically zero in the training data. For such cases, the parameter β_2^c is not identifiable. To address this issue of non-identifiability, for these experiments, we use the 7 most recent days to fit the linear predictors. For comparison, using the 7 most recent days instead of the 4 most recent days for the original linear predictor increases the raw-scale MAE (5.1b) from 7.0 to 7.2. Adding the new indicator feature v_t increases this error further to 7.4. As with the expanded shared model, we plot the mean raw-scale MAE per day of the week for 3-day-ahead predictions in Figure 15. With the weekday feature (panel (b)), we see that errors for most days are lower than that without the weekday feature (panel (a)), but this gain is offset by a high error for predictions made on Tuesday.

Appendix B. Further discussion on MEPI

We now first shed light on why MEPI had slightly worse coverage for some of the counties. And then, we provide a further discussion on various choices made for designing MEPI.

B.1. **Counties with poor coverage.** While Figure 11(a) shows that MEPI intervals achieve higher than 83% coverage for the vast majority of counties over the April 11–May 10 period, there are also counties with coverage below the targeted level. We provide a brief investigation of counties where the coverage of MEPIs for cumulative death counts is below 0.8 in Figure 11(a). Among 198 such counties, Figure 16 shows the cumulative deaths from April 11 to May 10 of the worst-affected 24 counties. Many of these counties exhibit a sharp uptick in the number of recorded deaths similar to that which we encountered in New York, possibly due to reporting lag. For instance,

Philadelphia (top row, first column from left) only recorded 2 new deaths between April 28 and May 3, but recorded 201 new deaths on May 4, which brought the cumulative deaths on May 4 to 625.
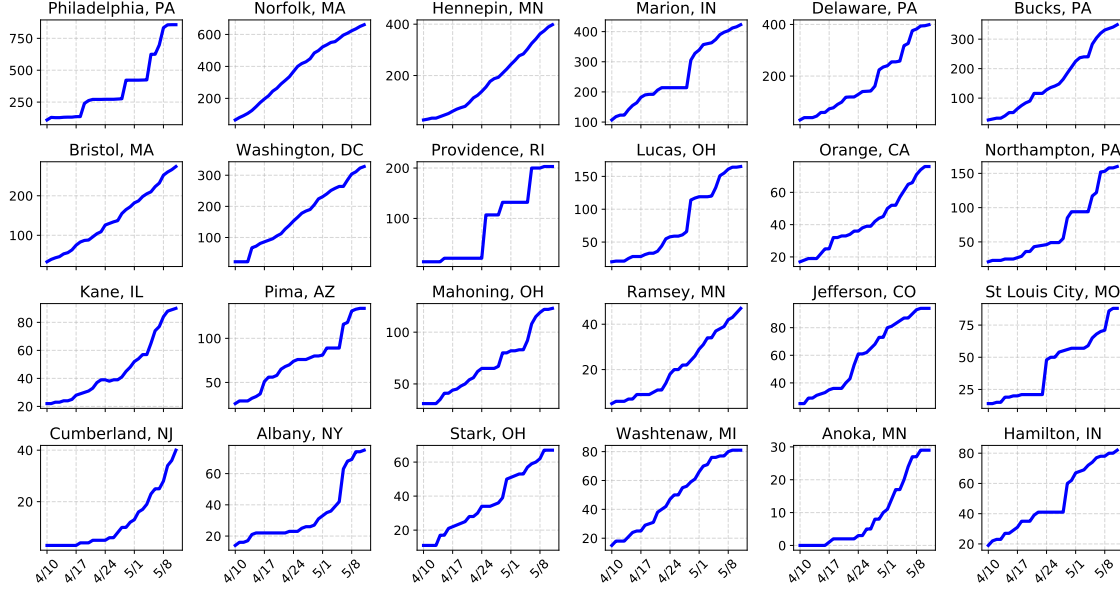
Figure 16. The cumulative death count data from the 24 worst affected counties where the coverage of the *7-day-ahead* MEPIs is below 0.8 (in Figure 11(a)).

B.2. **MEPI vs conformal interence.** Recall that the MEPI (equation 4.2a) can be viewed as a special case of conformal prediction interval [48, 41]. Here, we provide further discussion on this connection and discuss the assumptions under which the MEPI should achieve good coverage. A general recipe in conformal inference with streaming data is to compute the past several errors of the prediction model and use an s-percentile value for some suitable s (e.g., $s = 95$) to construct the prediction interval for the future observations. At a high-level, theoretical guarantees for conformal prediction intervals rely on the assumption that the sequence of errors is exchangeable. Roughly, the proof proceeds as follows: the exchangeability of the residuals ensures that the rankings of future residuals are uniformly distributed. Hence, the probability of the future residuals being in the top s-percentile is no larger than s, thereby obtaining the promised $s\%$-coverage. For more details, we refer the reader to the excellent tutorial [41] and the book [48].

Given the dynamic nature of COVID-19, it is unrealistic to assume that the prediction errors are exchangeable over a long period. As the cumulative death count grows, so too will the magnitude of the errors. Thus our MEPI scheme deviates from the general conformal recipe in two ways. We compute a *maximum error over the past 5 days*, and we *normalize* the errors. Each of these choices—of normalized errors and looking at only past 5 errors— is designed to make the errors

more exchangeable. Moreover, given that we take 5 data points to bound the future error, computing a maximum over them is a more conservative choice (e.g., when compared to taking median or a percentile-based cut-off).Furthermore, in order to compute a 95-percentile value, we need to consider errors for at least the past 20 days. Exchangeability is not likely to hold for such a long horizon.

Normalized vs. unnormalized errors: We now provide some numerical evidence to support our choice of normalized errors to define the MEPI. Figure 17(a) shows the rank distribution of normalized errors of our 7-day-ahead CLEP predictions for the six worst-affected counties over over an earlier period (March 26–April 25), and Figure 17(b) shows the (unnormalized) ℓ_1 errors $|\widehat{y}_t - y_t|$ over the same period. We found that in Figure 17(b), the ℓ_1 errors on days $t-4, t-3, t-2, t-1, t$ and $t+7$ do not appear to be exchangeable. Recall that under exchangeability conditions, the expected average rank of each of these six ℓ_1 errors would be 3.5. However, for all six counties, the average rank of the absolute error on day $t+7$ is larger than 4. This indicates that the future absolute error tends to be higher than past errors, and using the ℓ_1 error $|\widehat{y}_t - y_t|$ in place of the normalized error Δ_t can lead to substantial underestimation of future prediction uncertainty.

Longer time window: In Figure 17(c), we show the rank distribution of normalized errors over a longer window of 10 days. We found that due to the highly dynamic nature of COVID-19, these errors appear to be even less exchangeable. Under exchangeability conditions, the expected average rank of each of these 11 errors would be 6. However, we found that the average rank substantially deviates from this expected value for many days in this longer window for all displayed counties.

Overall, we believe that putting together the observations from Figures 5 and 17 yield reasonable justification for the two choices we made to define MEPI (equation 4.2a), namely, the 5-day window (versus the entire past), and the choice of normalized errors (versus the unnormalized absolute errors).

REFERENCES

[1] A. N. Angelopoulos, R. Pathak, R. Varma, and M. I. Jordan. On identifying and mitigating bias in the estimation of the COVID-19 case fatality rate. *Harvard Data Science Review*, 6 2020.

[2] Apple Inc. Apple Mobility Trends Reports. 2020. Accessed on 05-15-2020 at https://www.apple.com/covid19/mobility.

[3] M. Becker and C. Chivers. Announcing CHIME, a tool for covid-19 capacity planning. 2020. Accessed on 04-02-2020 at http://predictivehealthcare.pennmedicine.org/2020/03/14/accouncing-chime.html.

[4] D. J. Bregman and A. D. Langmuir. Farr's Law Applied to AIDS Projections. *JAMA*, 263(11):1522–1525, 03 1990.

[5] Bureau of Transportation Statistics. Airline origin and destination survey (db1b). 2020. Accessed on 04-20-2020 at https://transtats.bts.gov/Databases.asp?Mode_ID=1&Mode_Desc=Aviation&Subject_ID2=0.

[6] Centers for Disease Control and Prevention. Interactive Atlas of Heart Disease and Stroke. 2018. Accessed on 04-02-2020 at http://nccd.cdc.gov/DHDSPAtlas.

[7] Centers for Disease Control and Prevention, Agency for Toxic Substances and Disease Registry, and Geospatial Research, Analysis, and Services Program. Social Vulnerability Index Database. 2018. Accessed on 04-03-2020 at https://svi.cdc.gov/data-and-tools-download.html.

[8] Centers for Disease Control and Prevention, Division of Diabetes Translation, and US Diabetes Surveillance System. Diagnosed diabetes atlas. 2016. Accessed on 04-02-2020 at https://www.cdc.gov/diabetes/data.

[9] Centers for Medicare & Medicaid Services. Chronic Conditions Prevalence State/County Level: All Beneficiaries by Age, 2007-2017. 2017. Accessed on 04-02-2020 at https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Chronic-Conditions/CC_Main.

[10] Centers for Medicares & Medicaid Services. Case mix index file. 2018. Accessed on 04-01-2020 at https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/FY2020-IPPS-Final-Rule-Home-Page-Items/FY2020-IPPS-Final-Rule-Data-Files.

[11] Centers for Medicares & Medicaid Services. 2020 reporting cycle: Teaching hospital list. 2020. Accessed on 04-01-2020 at https://www.cms.gov/OpenPayments/Downloads/2020-Reporting-Cycle-Teaching-Hospital-List-PDF-.pdf.

[12] W.-H. Chiang, X. Liu, and G. Mohler. Hawkes process modeling of COVID-19 with mobility leading indicators and spatial covariates. *medRxiv preprint 2020.06.06.20124149*, 2020.

[13] County Health Rankings & Roadmaps. County Health Rankings & Roadmaps 2020 Measures. 2020. Accessed on 04-02-2020 at https://www.countyhealthrankings.org/explore-health-rankings/measures-data-sources/2020-measures.

[14] Definitive Healthcare. Definitive Healthcare: USA Hospital Beds. 2020. Accessed on 04-01-2020 at https://coronavirus-resources.esri.com/datasets/definitivehc::definitive-healthcare-usa-hospital-beds.

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

[16] H. H. Elmousalami and A. E. Hassanien. Day level forecasting for Coronavirus disease (COVID-19) spread: Analysis, modeling and recommendations. *arXiv preprint arXiv:2003.07778*, 2020.

[17] D. Fanelli and F. Piazza. Analysis and forecast of COVID-19 spreading in China, Italy and France. *Chaos, Solitons & Fractals*, 134:109761, 2020.

[18] N. Ferguson, D. Laydon, G. Nedjati-Gilani, N. Imai, K. Ainslie, M. Baguelin, S. Bhatia, A. Boonyasiri, Z. Cucunubá, G. Cuomo-Dannenburg, et al. Impact of non-pharmaceutical interventions (NPIs) to reduce COVID19 mortality and healthcare demand, 2020. Accessed on 04-02-2020 at https://www.imperial.ac.uk/media/imperial-college/medicine/sph/ide/gida-fellowships/Imperial-College-COVID19-NPI-modelling-16-03-2020.pdf.

[19] K. J. Goh, S. Kalimuddin, and K. S. Chan. Rapid progression to acute respiratory distress syndrome: Review of current understanding of critical illness from COVID-19 infection. *Annals of the Academy of Medicine, Singapore*, 49(1):1, 2020.

[20] Google LLC. Google COVID-19 Community Mobility Reports. 2020. Accessed on 05-15-2020 at https://www.google.com/covid19/mobility/.

[21] W. Guan, W. Liang, Y. Zhao, H. Liang, Z. Chen, Y. Li, X. Liu, R. Chen, C. Tang, T. Wang, et al. Comorbidity and its impact on 1590 patients with COVID-19 in China: A nationwide analysis. *European Respiratory Journal*, 2020.

[22] W. Guan, Z. Ni, Y. Hu, W. Liang, C. Ou, J. He, L. Liu, H. Shan, C. Lei, D. S. Hui, et al. Clinical characteristics of Coronavirus disease 2019 in China. *New England Journal of Medicine*, 2020.

[23] Health Resources and Services Administration. Area Health Resources Files. 2019. Accessed on 04-02-2020 at https://data.hrsa.gov/data/download.

[24] Health Resources and Services Administration. Health Professional Shortage Areas - Primary Care. 2020. Accessed on 04-04-2020 at https://data.hrsa.gov/data/download.

[25] Homeland Infrastructure Foundation-Level Data. Hospitals. 2020. Accessed on 06-23-2020 at https://hifld-geoplatform.opendata.arcgis.com/datasets/6ac5e325468c4cb9b905f1728d6fbf0f_0.

[26] S. Hsiang, D. Allen, S. Annan-Phan, K. Bell, I. Bolliger, T. Chong, H. Druckenmiller, A. Hultgren, L. Y. Huang, E. Krasovich, P. Lau, J. Lee, E. Rolf, J. Tseng, and T. Wu. The effect of large-scale anti-contagion policies on the Coronavirus (COVID-19) pandemic. *medRxiv*, 2020.

[27] Institute for Health Metrics and Evaluation. United States Chronic Respiratory Disease Mortality Rates by County 1980-2014. 2017. Accessed on 04-02-2020 at http://ghdx.healthdata.org/record/ihme-data/united-states-chronic-respiratory-disease-mortality-rates-county-1980-2014.

[28] Kaiser Health News. ICU Beds by County. 2020. Accessed on 04-02-2020 at https://khn.org/news/as-coronavirus-spreads-widely-millions-of-older-americans-live-in-counties-with-no-icu-beds/.

[29] B. D. Killeen, J. Y. Wu, K. Shah, A. Zapaishchykova, P. Nikutta, A. Tamhane, S. Chakraborty, J. Wei, T. Gao, M. Thies, and M. Unberath. A county-level dataset for informing the United States' response to COVID-19. *arXiv preprint arXiv:2004.00756*, 2020.

[30] A. J. Kucharski, T. W. Russell, C. Diamond, Y. Liu, J. Edmunds, S. Funk, and R. M. Eggo. Early dynamics of transmission and control of COVID-19: A mathematical modelling study. *medRxiv*, 2020.

[31] R. Marchant, N. I. Samia, O. Rosen, M. A. Tanner, and S. Cripps. Learning as we go: An examination of the statistical accuracy of covid19 daily death count predictions. *arXiv preprint arXiv:2004.04734*, 2020.

[32] MIT Election Data and Science Lab. County Presidential Election Returns 2000-2016. 2018.

[33] C. J. Murray and I. H. M. E. COVID-19 health service utilization forecasting team. Forecasting COVID-19 impact on hospital bed-days, ICU-days, ventilator-days and deaths by US state in the next 4 months. *medRxiv*, 2020.

[34] S. Nebehay and K. Kelland. COVID-19 cases and deaths rising, debt relief needed for poorest nations: WHO. *Reuters*, 2020-04-01. Accessed on 04-01-2020 at https://www.reuters.com/article/us-health-coronavirus-who/covid-19-infections-growing-exponentially-deaths-nearing-50000-who-idUSKBN21J6IL?il=0.

[35] C. M. Peak, R. Kahn, Y. H. Grad, L. M. Childs, R. Li, M. Lipsitch, and C. O. Buckee. Modeling the comparative impact of individual quarantine vs. active monitoring of contacts for the mitigation of COVID-19. *medRxiv*, 2020.

[36] S. Pei and J. Shaman. Initial simulation of SARS-CoV2 spread and intervention effects in the continental US. *medRxiv*, 2020.

[37] D. Qi, X. Yan, X. Tang, J. Peng, Q. Yu, L. Feng, G. Yuan, A. Zhang, Y. Chen, J. Yuan, X. Huang, X. Zhang, P. Hu, Y. Song, C. Qian, Q. Sun, D. Wang, J. Tong, and J. Xiang. Epidemiological and clinical features of 2019-nCoV acute respiratory disease cases in Chongqing municipality, China: A retrospective, descriptive, multiple-center study. *medRxiv*, 2020.

[38] L. Rubinson, F. Vaughn, S. Nelson, S. Giordano, T. Kallstrom, T. Buckley, T. Burney, N. Hupert, R. Mutter, M. Handrigan, et al. Mechanical ventilators in US acute care hospitals. *Disaster medicine and public health preparedness*, 4(3):199–206, 2010.

[39] G. D. Schuller, B. Yu, D. Huang, and B. Edler. Perceptual audio coding using adaptive pre-and post-filters and lossless compression. *IEEE Transactions on Speech and Audio Processing*, 10(6):379–390, 2002.

[40] S. Seabold and J. Perktold. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010.

[41] G. Shafer and V. Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(Mar):371–421, 2008.

[42] The Institute for Health Metrics and Evaluation. COVID-19: WhatâĂŹs New for April 5, 2020, 2020. http://www.healthdata.org/sites/default/files/files/Projects/COVID/Estimation_update_040520_3.pdf, Last accessed on 2020-04-13.

[43] The New York Times. COVID-19 Data in the United States. https://github.com/nytimes/covid-19-data, 2020. Accessed on 04-01-2020 at https://github.com/nytimes/covid-19-data.

[44] United States Census Bureau. County Adjacency File. 2018. Accessed on 05-15-2020 at https://www.census.gov/geographies/reference-files/2010/geo/county-adjacency.html.

[45] United States Department of Agriculture, Economic Research Service. Poverty estimates for the U.S., states, and counties. 2018. Accessed on 04-24-2020 at https://www.ers.usda.gov/data-products/county-level-data-sets/download-data/.

[46] United States Department of Health and Human Services, Centers for Disease Control and Prevention , and National Center for Health Statistics. Compressed Mortality File (CMF) on CDC WONDER Online Database, 2012-2016. 2017. Accessed on 04-02-2020 at https://wonder.cdc.gov/cmf-icd10.html.

[47] USAFacts. COVID-19 Deaths Data. 2020. Accessed on 03-31-2020 at https://www.reuters.com/article/us-health-coronavirus-who/covid-19-spread-map.

[48] V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic learning in a random world*. Springer Science & Business Media, 2005.

[49] C. Wang, L. Liu, X. Hao, H. Guo, Q. Wang, J. Huang, N. He, H. Yu, X. Lin, A. Pan, S. Wei, and T. Wu. Evolving epidemiology and impact of non-pharmaceutical interventions on the outbreak of Coronavirus disease 2019 in Wuhan, China. *medRxiv*, 2020.

[50] J. Wu, A. McCann, J. Katz, and E. Peltier. 109,000 missing deaths: Tracking the true toll of the coronavirus outbreak. *The New York Times*, 2020.

[51] F. Zhou, T. Yu, R. Du, G. Fan, Y. Liu, Z. Liu, J. Xiang, Y. Wang, B. Song, X. Gu, et al. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: A retrospective cohort study. *The Lancet*, 2020.
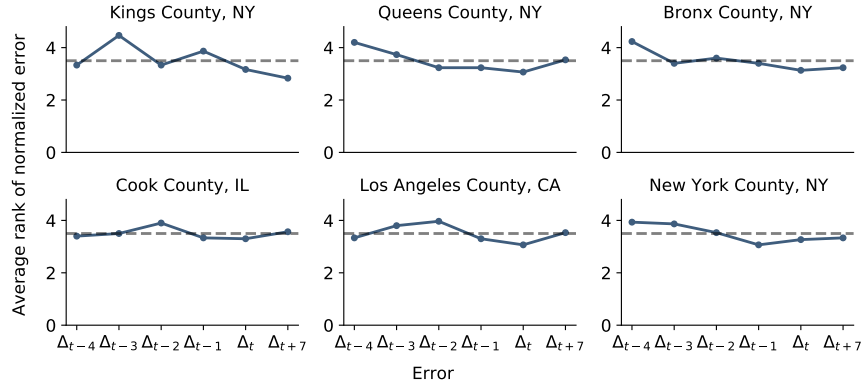
(a) Rank distribution of normalized errors over 5 days
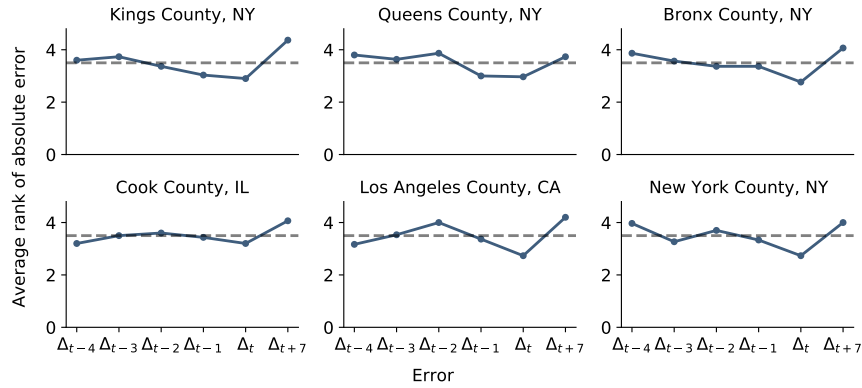
(b) Rank distribution of absolute unnormalized errors over 5 days
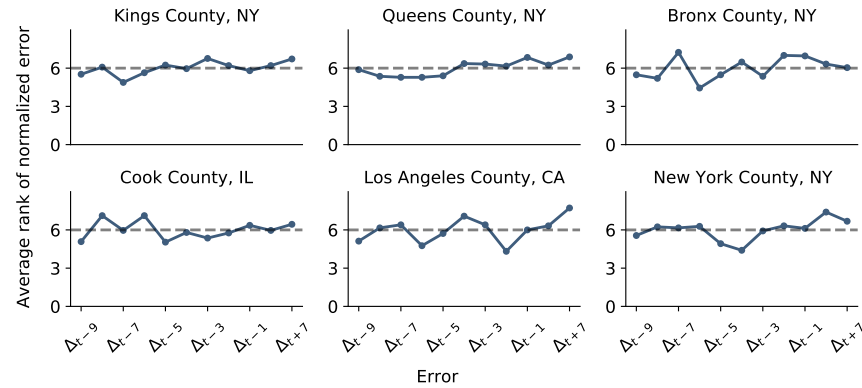
(c) Rank distribution of normalized errors over a longer window of 10 days

Figure 17. EDA plot with unnormalized and normalized errors for *7-day-ahead* predictions made by CLEP, computed over $t =$ March 26, ..., April 25. **(a)** The rank distribution of normalized errors of our CLEP (with the expanded shared and linear predictors) for the six worst affected counties; **(b)** the absolute unnormalized errors of our CLEP for the six worst affected counties and **(c)** the rank distribution of the normalized errors over a longer window.
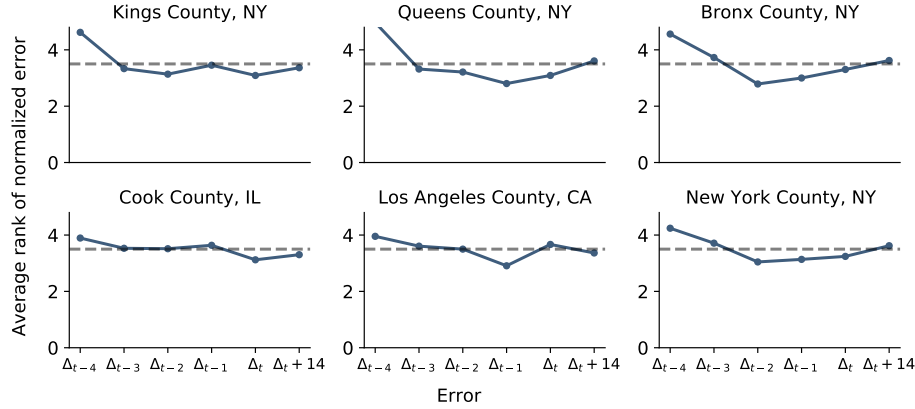
(a) Six worst-affected counties
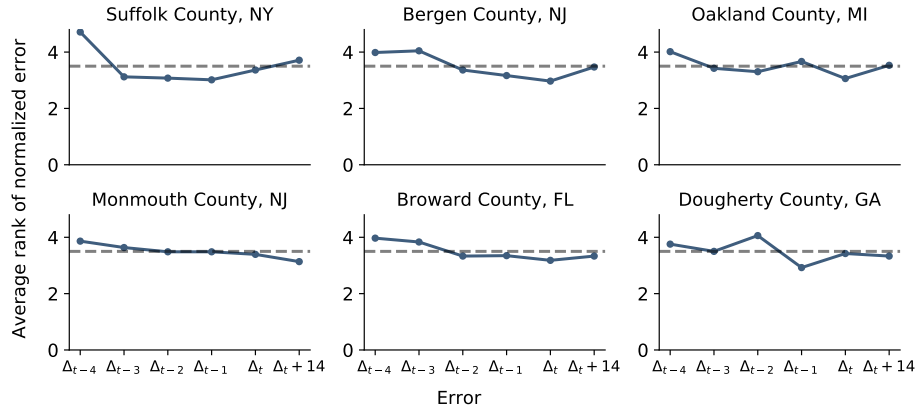
(b) Six randomly-selected counties

Figure 18. EDA plot for investigating exchangeability of normalized errors of *14-day-ahead* CLEP predictions with its last 5 errors made at time t, over the period $t =$ April 2, ..., Jun 6. We plot the average rank of the six errors $\{\Delta_{t+14}, \Delta_t, \Delta_{t-1}, \ldots, \Delta_{t-4}\}$ of our CLEP (with the expanded shared and linear predictors) for **(a)** the six worst affected counties, and **(b)** six random counties. We rank the errors $\{\Delta_{t+14}, \Delta_t, \Delta_{t-1}, \ldots, \Delta_{t-4}\}$ in increasing order so that the largest error has a rank of 6. If $\{\Delta_{t+14}, \Delta_t, \Delta_{t-1}, \ldots, \Delta_{t-4}\}$ are exchangeable for any day t, then the expected average rank for each of the six errors would be 3.5 (dashed black line).