HDSR
HARVARD DATA SCIENCE REVIEW
Special Issue 1: COVID-19: Unprecedented Challenges and Chances
DOI: /10.1162/99608f92.1d4e0dae
ISSN: 2644-2353

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 article, 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 2 weeks ahead. Using data from January 23 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 5 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 2 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 2 weeks) are critical for making immediate policy decisions such as whether or not counties should reopen. This article 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, combined linear and exponential predictors (CLEP), along with a prediction interval, maximum (absolute) error prediction intervals (MEPI), for forecasting short-term county-level COVID-19 mortality in the United States. By combining different trends in the death count data, our county-level CLEP forecasts accurately predicted cumulative deaths due to COVID-19 up to 14 days into the future. The MEPI prediction intervals exhibit high coverage for the forecasts. Our approach, originally developed in May 2020, 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 and available at https://covidseverity.com. Our predictions are currently being used by a 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.

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 article, together with the data we have collated (and continue to update), 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 2 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 (Vovk et al., 2005).

Predicting the short-term impact (e.g., 1 or 2 weeks in the future) of the virus in terms of the number of deaths is critical for many reasons. Not only can it help elucidate the overall fallout of the virus, but it can also help guide difficult policy decisions, such as whether or not to impose or ease lockdowns, and where to send much-needed personal protective equipment (PPE). While many other studies focus on predicting the long-term (several months or a year) trajectory of COVID-19, these approaches are currently difficult to verify due to a lack of long-term COVID-19 data.[1] On the other hand, predictions for immediate short-term trajectories are much easier to verify and are more likely to be accurate than long-term forecasts since there are comparatively fewer uncertainties involved, for example, due to policy change or behavioral changes in society. So far, the vast majority of predictive efforts have focused on modeling COVID-19 case counts or death counts at the national or state level (Ferguson et al., 2020), rather than the more fine-grained

[1]In the time since the first version of this article **in May 2020**, such longer-term predictions are likely now more verifiable.

county level that we consider in this article. To the best of our knowledge, ours was the first work on county-level forecasts.[2]

The predictions we produce in this article 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* recorded cases is difficult since some counties conducted more tests than others: the number of positive tests does not equal the number of actual cases. While the *proportion* of tests that are positive is more comparable across different counties, our modeling approach focuses on recorded death counts rather than proportions, since these are not influenced by testing biases. It is worth noting, however, that the recorded death count is likely to be an undercount of the number of true number COVID-19 deaths, since evidence implies that many deaths that occurred outside of hospitals were often not counted.[3] Nonetheless, the recorded death count is generally believed to be more reliable than the recorded case count, and recent efforts have been made to ensure that COVID-19 death counts are more accurately recorded, for example, by including probable deaths and deaths occurring at home (Nguyen & Schechtman, 2020).

In Section 2, we introduce our data repository and summarize the data sources contained within, as well as discussing any sources of bias in the data. This data repository is being updated continuously (as of October 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.

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, that is, across regions and time. The predictions generated by the different methods are combined using an ensembling technique by Schuller et al. (2002), and we refer to the ensemble model as the combined linear and exponential predictors (CLEP). Additional predictive approaches, including those using social distancing information are presented in Appendix A, although they do not outperform the original CLEP predictors that are presented in the main body of this article.

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 (Vovk et al., 2005) where the prediction interval coverage is well defined as the empirical proportion of days when the observed cumulative death counts fall inside the prediction intervals.

Section 5 details the evaluation of the predictors and the prediction intervals for the forecasts 3, 5, 7, and 14 days into the future. We use the data from January 23, 2020, the day after the first COVID-19 confirmed case (on January 22) in the United States (Stokes et al., 2020), 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 subexponential nature of COVID-19 outbreak, with errors of around 15% for 7-day-ahead predictions, and errors of around 30% 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

[2]At 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 related work where we discuss the county-level forecasts by Chiang et al. (2020) that was published after our original work, in early June.

[3]For the period up to June 21, 2020, considered in this article, for example, see Katz et al. (2020)

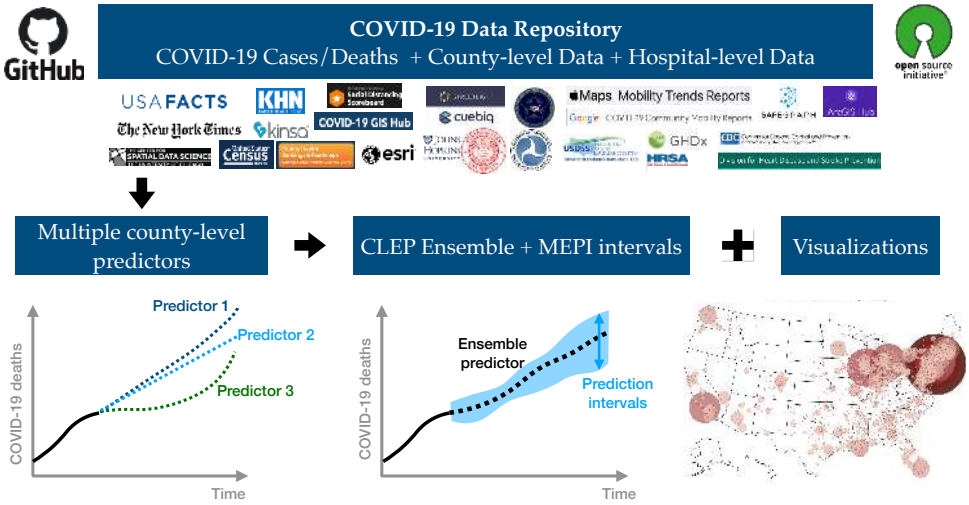


Figure 1. An overview of the article. We curate an extensive data repository combining data from multiple data sources. We then build several predictors for county-level predictions of cumulative COVID-19 death counts, and develop an ensembling procedure combined linear and exponential predictors (CLEP) and a prediction interval scheme maximum (absolute) error prediction intervals (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 (https://github.com/Yu-Group/covid19-severity-prediction). 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.

narrow and cover the recorded number of deaths for more than 90% of days for most of the counties in the United States.

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 the data and the predictive algorithms used in this article accessible to others is key to ensuring their usefulness. Thus the data, code, and predictors we discuss in this article 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. While the results in this article contain case and death information at county level in the United States from January 23, 2020, to June 20, 2020; the data, forecasts, and visualizations in the GitHub repository and on our website continue to be 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-plus public sources to aid data science efforts to combat COVID-19.

2.1. **Overview of the data sets on June 20, 2020.** At the *hospital level*, our data set covers over 7,000 U.S. 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 and critical access), and numerous other hospital statistics.

There are more than 3,100 counties in the United States. At the *county-level*, our repository includes:

(i) daily recorded COVID-19–related case count and (recorded) death count by *New York Times (2020)* (hereafter NYT, 2020) and USAFacts (2020);
(ii) demographic features such as population distribution by age 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) 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 8,000 features in the county-level data set. We provide a feature-level snapshot of the different types of data available in our repository, highlighting features in the county-level data sets in Table 1 and the hospital-level data sets 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 data set.

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 data set for ease of navigation on our github. We have also provided a quick-start guide for accessing the unabridged county-level and hospital-level data sets with a single Python code line.

data sets used by our predictors: In this article, 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 23, 2020, to June 20, 2020 (pulled on June 21, 2020) along with some county-level demographics and health data. The data sets that are used to compute the predictors in this article are 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 to facilitate our work with Response4Life and aid medical supply allocation efforts. However, over time the data repository has grown to encompass more resources and now supports investigations into a wider range of COVID-19–related problems. For instance, using the breadth of travel information in our repository, including (aggregated) air travel and work commute data, researchers can investigate the impact of both local and between-city travel patterns on the spread of COVID-19. Our 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, our data also enables an investigation into the connection between socioeconomic and demographic information with health resource data (e.g., number of ICU beds, medical staff) to gain a deeper understanding of the severity of the pandemic at the county level. 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 the larger cities and counties that have received more attention.

Comparison with the repository collated by Killeen et al. (2020) at Johns Hopkins University: Note that a similar but complementary county-level data set was recently aggregated and released in another study (Killeen et al., 2020). Both our county-level repository and the repository in Killeen et al. (2020) include data on COVID-19 cases and deaths, demographics, socioeconomic information, education, and mobility, however, many are from different sources. For example, the repository by Killeen et al. (2020) uses COVID-19 cases and deaths data from the John Hopkins University Center for Systems Science and Engineering COVID-19 dashboard by Dong et al. (2020), whereas our data is pulled from USAFacts (2020) and NYT (2020). 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 Killeen et al. (2020) provides additional data sets at the state level, we provide additional data sets at the hospital level (given our initial goal of helping the allocation of medical supplies to hospitals, in partnership with the nonprofit Response4Life). While their data repository contains both overlapping and complementary information to ours, a thorough data set-by-data set comparison is beyond the scope of this work for two reasons: (i) we learned about this repository toward the completion of our work, and (ii) we were unable to find detailed documentation of how the data sets in their repository were cleaned.

2.2. **Data quality and bias.** 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 exploratory data analysis (EDA) to shed light on the scale of bias and the possible directions of the biases in the data.

Though discussions on data-quality issues and their possible consequences are relatively sparse in the existing literature, Angelopoulos et al. (2020) discussed 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., underascertainment of mild cases) for estimating the case fatality ratio do not affect estimation of future death counts. Nonetheless, many of the ideas we present with our EDA here in uncovering possible biases in the data are inspired by Angelopoulos et al. (2020).

Imperfect reporting and attribution of deaths due to COVID-19: Numerous news articles have suggested that the official U.S. COVID-19 death count is an underestimate (Wu et al., 2020). According to *The New York Times* (Walker et al., 2020), on April 5, the Council of State and Territorial Epidemiologists advised states to include in their death counts both confirmed cases based on laboratory testing, and probable cases using specific criteria for symptoms and exposure. The Centers for Disease Control and Prevention (CDC) adopted these definitions, and national CDC COVID-19

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

Description of County-level Features	Data Source(s)
COVID-19 Cases/Deaths	
Daily # of COVID-19-related recorded cases by US county	USAFacts, 2020; New York Times, 2020
Daily # of COVID-19-related deaths by US county	USAFacts, 2020; New York Times, 2020
Demographics	
Population estimate by county (2018)	Health Resources and Services Administration, 2019 (Area Health Resources Files)
Census population by county (2010)	Health Resources and Services Administration, 2019 (Area Health Resources Files)
Age 65+ population estimate by county (2017)	Health Resources and Services Administration, 2019 (Area Health Resources Files)
Median age by county (2010)	Health Resources and Services Administration, 2019 (Area Health Resources Files)
Population density per square mile by county (2010)	Health Resources and Services Administration, 2019 (Area Health Resources Files)
Socioeconomic Factors	
% uninsured by county (2017)	County Health Rankings & Roadmaps, 2020
High school graduation rate by county (2016-17)	County Health Rankings & Roadmaps, 2020
Unemployment rate by county (2018)	County Health Rankings & Roadmaps, 2020
% with severe housing problems in each county (2012-16)	County Health Rankings & Roadmaps, 2020
Poverty rate by county (2018)	United States Department of Agriculture, Economic Research Service, 2018
Median household income by county (2018)	United States Department of Agriculture, Economic Research Service, 2018
Social vulnerability index for each county	Centers for Disease Control and Prevention et al., 2018 (Social Vulnerability Index)
Health Resources Availability	
# of hospitals in each county	Kaiser Health News, 2020
# of ICU beds in each county	Kaiser Health News, 2020
# of full-time hospital employees in each county (2017)	Health Resources and Services Administration, 2019 (Area Health Resources Files)
# of MDs in each county (2017)	Health Resources and Services Administration, 2019 (Area Health Resources Files)
Health Risk Factors	
Heart disease mortality rate by county (2014-16)	Centers for Disease Control and Prevention, 2018a (Interactive Atlas of Heart Disease and Stroke)
Stroke mortality rate by county (2014-16)	Centers for Disease Control and Prevention, 2018a (Interactive Atlas of Heart Disease and Stroke)
Diabetes prevalence by county (2016)	Centers for Disease Control and Prevention et al., 2016 (Diagnosed Diabetes Atlas)
Chronic respiratory disease mortality rate by county (2014)	Institute for Health Metrics and Evaluation, 2017
% of smokers by county (2017)	County Health Rankings & Roadmaps, 2020
% of adults with obesity by county (2016)	County Health Rankings & Roadmaps, 2020
Crude mortality rate by county (2012-16)	United States Department of Health and Human Services et al., 2017
Mobility	
Start date of stay at home order by county	Killeen et al., 2020
% change in mobility at parks, workplaces, transits, groceries/pharmacies, residential, and retail/recreational areas	Google LLC, 2020

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

Description of Hospital-level Features	Data Source(s)
CMS certification number	Centers for Medicares & Medicaid Services, 2018 (Case Mix Index File)
Case Mix Index	Centers for Medicares & Medicaid Services, 2018 (Case Mix Index File); Centers for Medicares & Medicaid Services, 2020 (Teaching Hospitals)
Hospital location (latitude and longitude)	Homeland Infrastructure Foundation-Level Data, 2020; Definitive Healthcare, 2020
# of ICU/staffed/licensed beds and beds utilization rate	Definitive Healthcare, 2020
Hospital type	Homeland Infrastructure Foundation-Level Data, 2020; Definitive Healthcare, 2020
Trauma Center Level	Homeland Infrastructure Foundation-Level Data, 2020
Hospital website and telephone number	Homeland Infrastructure Foundation-Level Data, 2020

death counts began including confirmed and probable cases on April 14. The data included in our repository (from USAFacts and *NYT*) contains both the probable death and the confirmed COVID-19 deaths beginning April 14. Although the probable death counts address imperfect reporting and attribution, there is still the possibility of underreporting in some counties. However, the magnitude of the possible underreporting and the counties where underreporting occurred is unclear. Going forward, we use the terms '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 United States. In our data repository, we include data from USAFacts (2020) and NYT (2020). According to USAFacts and *NYT* websites, they both collect data from state 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)," *NYT* claims to repeatedly update data throughout the day. Furthermore, while there are some discussions on how they collect and process the data on their respective websites, the specific data curation rules are not shared publicly.

Possibly due to these differences in scraping times and data curation, there are a small percentage of discrepancies in the case and death counts between USAFacts and *NYT* data sets. Across the 3,222 counties and 150 days under study, we plot a histogram of differences in death counts between the two data sets for each county-day combination in Figure 2(a). Here, 475,533 out of all 483,300 county-day combinations (98.4%) of the COVID-19 case and death counts in USAFacts and *NYT* data sets are identical. Since Figure 2 (a) is dominated by the number of 0 counts, in Figure 2(b), we truncate the histogram at 200 to provide details on the discrepancies between the two data sets. We note that 7,207 (1.5%) entries have an absolute difference between 1 and 5; 560 (0.1%) entries have an absolute difference larger than 5. Moreover, out of the 3,222 counties, 2,193 (68.1%) counties show no discrepancies between USAFacts and *NYT* data sets on any day throughout the entire period under study. Although the overall differences between USAFacts and *NYT* case- and death-count data appear to be small, it is unclear how to combine or validate

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

County-level data set	Description
COVID-19 Cases/Deaths Data	
USAFacts, 2020*†	Daily cumulative number of reported COVID-19-related death and case counts by US county, dating back to Jan. 23, 2020
New York Times, 2020†	Similar to the USAFacts data set, but includes aggregated death counts in New York City without county breakdowns
Demographics and Socioeconomic Factors	
Health Resources and Services Administration, 2019 (Area Health Resources Files)*	Includes data on health facilities, professions, resource scarcity, economic activity, and socioeconomic factors (2018-2019)
County Health Rankings & Roadmaps, 2020*	Estimates of various health behaviors and socioeconomic factors (e.g., unemployment, education)
Centers for Disease Control and Prevention et al., 2018 (Social Vulnerability Index)	Reports the CDC's measure of social vulnerability from 2018
United States Department of Agriculture, Economic Research Service, 2018	Poverty estimates and median household income for each county
Health Resources Availability	
Health Resources and Services Administration, 2019 (Area Health Resources Files)*	Includes data on health facilities, professions, resource scarcity, economic activity, and socioeconomic factors (2018-2019)
Health Resources and Services Administration, 2020 (Health Professional Shortage Areas)	Provides data on areas having shortages of primary care, as designated by the Health Resources & Services Administration
Kaiser Health News, 2020*	# of hospitals, hospital employees, and ICU beds in each county
Health Risk Factors	
County Health Rankings & Roadmaps, 2020*	Estimates of various socioeconomic factors and health behaviors (e.g., % of adult smokers, % of adults with obesity)
Centers for Disease Control and Prevention, 2018a (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 et al., 2016 (Diagnosed Diabetes Atlas)*	Estimated percentage of people who have been diagnosed with diabetes per county (2016)
Institute for Health Metrics and Evaluation, 2017*	Estimated mortality rates of chronic respiratory diseases (1980-2014)
Centers for Medicare & Medicaid Services, 2017 (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 et al., 2017	Overall mortality rates (2012-2016) for each county from the National Center for Health Statistics
Mobility	
Killeen et al., 2020 (JHU Date of Interventions)	Dates that counties (or states governing them) took measures to mitigate the spread by restricting gatherings
Google LLC, 2020 (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, 2020 (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, 2018 (County Adjacency File)*	Lists each US county and its neighboring counties; from the US Census
Bureau of Transportation Statistics, 2020 (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, 2018 (County Presidential Data)	County-level returns for presidential elections from 2000 to 2016 according to official state election data records

Table 4. A list of hospital-level data sets 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 data sets.

Hospital-level data set	Description
Homeland Infrastructure Foundation-Level Data, 2020	Includes number of ICU beds, and location for US hospitals
Definitive Healthcare, 2020	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, 2018 (Case Mix Index File)	Reports the Case Mix Index (CMI) for each hospital
Centers for Medicares & Medicaid Services, 2020 (Teaching Hospitals)	Lists teaching hospitals along with address (2020)

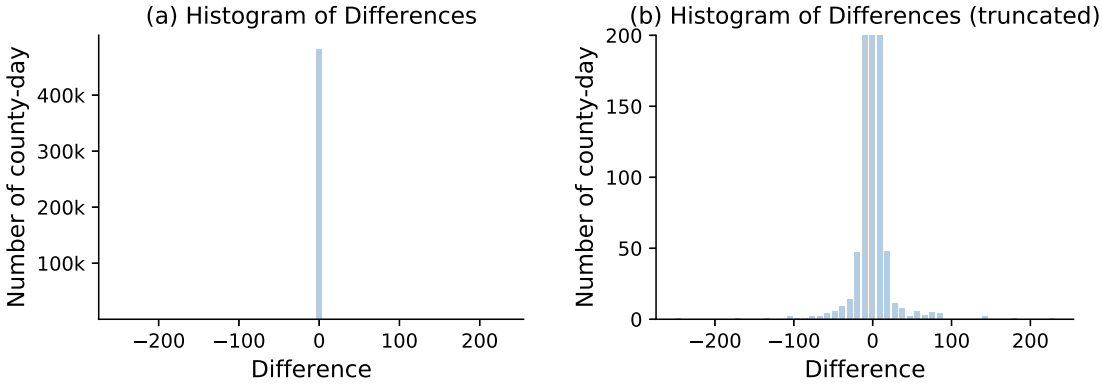


Figure 2. Histograms illustrating the distribution of the differences in county-level recorded daily death counts between USAFacts and *NYT* data sets for all 3,222 counties and 150 days (January 23 to June 20) in our study. In panel **(a)**, we plot the entire histogram of differences between these two data sources, where the vertical axis denotes the number of county-day combinations for a given difference (maximum possible $3{,}222 \times 150 = 483{,}300$). In panel **(b)**, we truncate the vertical axis of the histogram at 200 to show the pattern of nonzero differences.

the two data sources given the different curation rules (which are unknown to us). We choose to proceed with the USAFacts COVID-19 deaths data for our analysis as they provide county-level death counts for New York City while *NYT* aggregates the death counts over the NYC counties.

Weekday patterns: The recorded case counts and death counts have a significant weekly pattern in both USAFacts and *NYT* data sets; such a pattern can possibly be attributed to the reporting delays as discussed in Angelopoulos et al. (2020). We show the total number of deaths recorded on each day of the week in the USAFacts data in Figure 3a. 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 observe that some of the historical infection data was revised after initially being recorded. According to USAFacts, these revisions are typically due to earlier mistakes

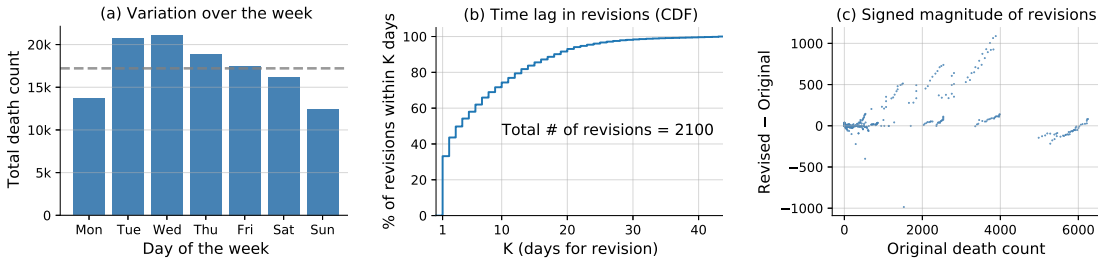


Figure 3. Exploratory data analysis (EDA) plots with USAFacts data set for identifying potential biases over the period January 23 to June 20, 2020. Panel **(a)** shows the recorded death counts for each day of the week, where the horizontal dashed line represents the average across all days. In panel **(b)**, we plot the cumulative distribution for the fraction of revisions made versus the number of days for the revisions. In panel **(c)**, we present the signed magnitude of the change, that is, the "revised count minus the original count" against the original count for each revision.

from local agencies that 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 observe that only 2.1% of counties across the United States had one or more historical revisions. Figure 3b 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 within 2 days from the day data was initially recorded.

Figure 3c shows the signed magnitude of the change in death count that results from the revisions versus the initial recorded death count. Note that there are several trends displaying consecutive upward revisions in the plot. Each such trend 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 are revised, all the data after that day until the day when the revision is made also are revised accordingly, thereby explaining the observed trends. Note that only 582 out of 2,100 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, among 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, that is, the points with large negative "revised-original count" in Figure 3c, correspond to counties in Washington State. This finding can be corroborated by the media news that Washington State admitted to making errors in reporting the death counts, and subsequently lowered these counts in the revisions (Granneman, 2020). It is natural for our predictions to vary if the training data (for a fixed period) varies with time, that is, 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 larger, our predictions tend to be less accurate for a few days following the revision. See Section 5.1 and 5.2 for further discussion on these biases.

We thus 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 not perfect, and the subsequent bias is hard to adjust since we do not know the true counts.

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 death counts per 1,000 residents[4] for the same period. Each bubble denotes a county-level count, a darker and larger bubble denotes a higher count, and the absence of a bubble denotes that the count is zero. Panel (a) captures the absolute extent of the outbreak in a region, while (b) captures the impact relative to population sizes. Overall, Figure 4 clearly shows that up to June 20, the COVID-19 outbreak had taken a great toll throughout the United States. Large population centers were particularly affected in terms of total deaths, and the North- and Southeast in terms of deaths per capita. While Los Angeles County recorded 3,110 deaths up to June 20—the fifth highest in the country—it appears relatively unaffected in panel (b) due to the large population size (over 10 million). Of the 10 largest bubbles in panel (b), four counties experienced more than 1,000 total deaths up to June 20, all of which are associated with the severe NYC area outbreak early in the course of the pandemic. In order from largest to smallest total deaths per 1,000 residents, they are Bronx, New York (4,631 deaths among 1,385,108 estimated residents in 2018), Queens, New York (6,492 deaths among 2,230,722 estimated residents), Kings, New York (6,985 deaths among 2,504,700 estimated residents), and Essex, New Jersey (1,759 deaths among 783,969 estimated residents). In order to obtain a better understanding of the pandemic, it is important to study it from multiple perspectives using both normalized and absolute figures. We provide an interactive dashboard at https://covidseverity.com, which includes visualizations of the cumulative and new case and death counts for counties across the country, both with and without normalization by population size.

We develop several different statistical and 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 article 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.

[4]We display deaths per 1,000 residents rather than per 100,000 as is common to avoid misleading conclusions for smaller counties. For example, the county with the largest number of deaths per 1,000 up to June 20 was Hancock, Georgia, which had an estimated 8,348 residents in 2018 and 32 deaths from COVID-19 by June 20. If we were to use deaths per 100,000, the value for Hancock would be about 383, far exceeding the actual total number of deaths.

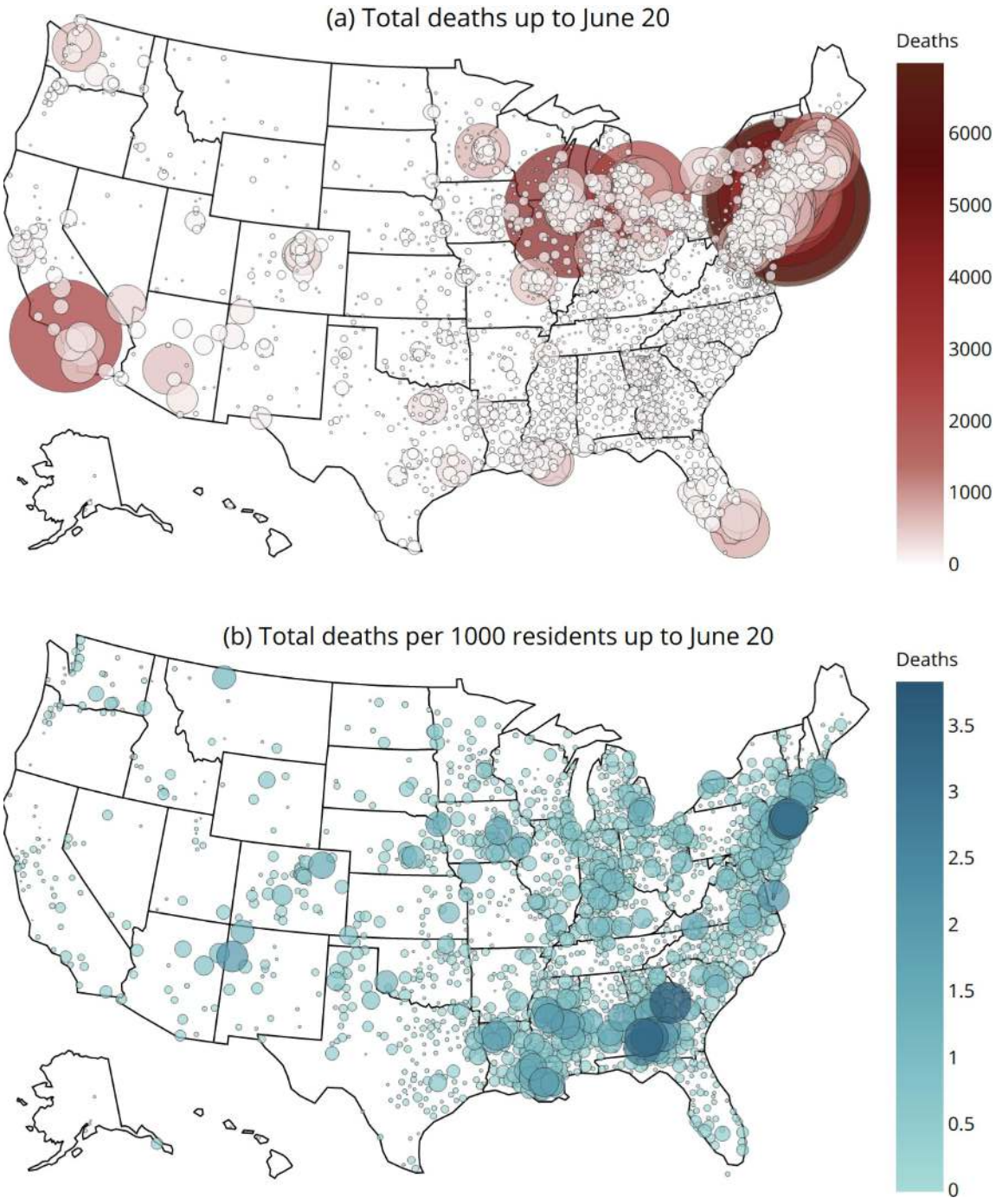


Figure 4. Visualizations of the COVID-19 outbreak in the United States. We depict the cumulative recorded death counts up to June 20 in panel **(a)** and death counts per 1,000 residents in panel **(b)**. Each bubble denotes the count for a county (the absence of a bubble denotes a zero count). The bubble size (area) is proportional to the counts in the county. Note that the two panels' bubble sizes are not on the same scale.

(5) **A demographics shared-county exponential predictor (the 'demographics shared' predictor)**: A predictor also similar to the shared-county exponential predictor, but that 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 (Seabold & Perktold, 2010) to train all the five predictors: ordinary least squares for predictor (2), and Poisson regression for predictors (1), (3), (4) and (5) (where the set of features for each predictor is different).[5] 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 individual predictors, where we borrow the weighting scheme from prior work (Schuller et al., 2002). 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 overfitting; 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 underreporting 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 article we discuss our algorithms for predicting cumulative recorded death counts, the methods can be more generally applied to predict other quantities of interest, for example, 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.

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 (Nebehay & Kelland, 2020). 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:

$$\widehat{\mathrm{E}}[\mathrm{deaths}^c_{t+1}|t] = \exp\left(\beta^c_0 + \beta^c_1(t+1)\right), \tag{3.1}$$

[5]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. For predictors (3)-(5), *glm.fit_ regularized* was used to fit ℓ_1 and ℓ_2 regularized predictive models. While we ended up using regularization for only predictor (5) and not for (3) and (4), 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 even in predictors (3) and (4), and consequently different performance. We still chose to use *glm.fit_ regularized* for fitting predictors (3) and (4) 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.

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).

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		✓		✓

where $\widehat{\mathrm{E}}[\mathrm{deaths}^c_{t+1}|t]$ 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 right-hand side of (3.1) just for notational exposition, and in practice we just use $\beta^c_0 + \beta^c_1 t$ in the exponent in our code.

Here we fit a separate predictor for each county, and the coefficients β^c_0 and β^c_1 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 response (or dependent) variable. On the morning of day $t+1$, the coefficients are estimated using the cumulative recorded death counts for day $\{t, t-1, t-2, t-3, t-4\}$. To predict k-days-ahead cumulative death count on the morning of day $t+1$—denoted by $\widehat{\mathrm{E}}[\mathrm{deaths}^c_{t+k}|t]$—we replace $t+1$ with $t+k$ on the RHS of (3.1). Note that although the prediction $\widehat{\mathrm{E}}(\mathrm{deaths}^c_{t+1}|t)$ 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^c_{t+1} 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 are used for the fit. If there is less than 3 days' worth of data or the cumulative deaths remain constant in the past days, we 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 that using 4 days of data performed slightly worse.

To handle possible overdispersion of data (when the variance is larger than the mean), we also explored estimating $\{\beta^c_0, \beta^c_1\}$ by fitting a negative binomial regression model (in place of Poisson GLM) with inverse-scale parameter taking values in $\{0.05, 0.15, 1\}$.[6] 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

[6]These values were based on the typical permissible range of [0.02, 2] as per the documentation at https://www.statsmodels.org/stable/generated/statsmodels.genmod.families.family.NegativeBinomial.html

county. In the early stages of tuning, we tried using 5 and 7 days of data, and obtained worse performance (also see the discussion in Appendix A.2.2). The motivation for the linear model is that some counties are exhibiting subexponential 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

$$\widehat{\mathrm{E}}[\mathrm{deaths}^c_{t+1}|t] = \beta^c_0 + \beta^c_1(t+1), \tag{3.2}$$

where we fit the coefficients β^c_0 and β^c_1 via ordinary least squares using the cumulative death count for county c for most recent 4 days. Like (3.1), we use $t+1$ on the RHS simply for notational exposition. For instance, on the morning of day $t+1$, the coefficients $\{\beta^c_0, \beta^c_1\}$ are estimated using the death counts for day $t, t-1, t-2, t-3$. To predict k-days-ahead, that is, predict cumulative death counts by the end of day $t+k$ on the morning of day $t+1$ (in our notation, $\widehat{\mathrm{E}}[\mathrm{deaths}^c_{t+k}|t]$), we simply replace $t+1$ by $t+k$ on the RHS of (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 allow 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. For each county, instead of including only the most recent 5 days, we include all days after the third death in the 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 prediction from the exponential predictor 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:

$$\widehat{\mathrm{E}}[\mathrm{deaths}^c_{t+1}|t] = \exp\Big(\beta_0 + \beta_1 \log(\mathrm{deaths}^c_t + 1)\Big), \tag{3.3}$$

where $\widehat{\mathrm{E}}[\mathrm{deaths}^c_{t+1}|t]$ denotes the (fitted) cumulative death count by the end of day $t+1$ for a county c, and deaths^c_t 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 (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 $\widehat{\mathrm{E}}[\mathrm{deaths}^c_{t+k}|t]$, we first obtain the estimate $\widehat{\mathrm{E}}[\mathrm{deaths}^c_{t+1}|t]$ using (3.3). Next, we substitute $\log(\widehat{\mathrm{E}}[\mathrm{deaths}^c_{t+j}|t] + 1)$ (after normalizing across all counties) on the RHS of (3.3) to compute $\widehat{\mathrm{E}}[\mathrm{deaths}^c_{t+j+1}|t]$ in a sequential manner for $j = 1, \ldots, k-1$, and finally obtain $\widehat{\mathrm{E}}[\mathrm{deaths}^c_{t+k}|t]$ (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, neigh_deaths_t^c, 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

$$\begin{aligned}\widehat{\mathrm{E}}[\text{deaths}_{t+1}^c|t] = \exp\Big(&\beta_0 + \beta_1 \log(\text{deaths}_t^c + 1) + \beta_2 \log(\text{cases}_{t-k+1}^c + 1) \\ &+ \beta_3 \log(\text{neigh\_deaths}_{t-k+1}^c + 1) + \beta_4 \log(\text{neigh\_cases}_{t-k+1}^c + 1)\Big),\end{aligned} \tag{3.4}$$

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 use only the new features (cases in the current county, cases in neighboring counties, and deaths in neighboring counties) up to the end of day $t-k+1$, since when predicting $\widehat{\mathrm{E}}[\text{deaths}_{t+k}^c|t]$ these covariates would be available only up to day t, that is, k days before. 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 $\widehat{\mathrm{E}}[\text{deaths}_{t+1}^c|t]$ by substituting 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 (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 substitute $\log(\widehat{\mathrm{E}}[\text{deaths}_{t+j}^c|t])$, $\log(\text{cases}_{t-k+j+1}^c)$, $\log(\text{neigh\_deaths}_{t-k+j+1}^c)$, $\log(\text{neigh\_cases}_{t-k+j+1}^c)$) in (3.4) (again after normalizing each of these features) to compute $\widehat{\mathrm{E}}[\text{deaths}_{t+j+1}^c|t]$, and finally compute $\widehat{\mathrm{E}}[\text{deaths}_{t+k}^c|t]$ 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 for 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 exponential predictor. However, it includes several static county demographic and health care–related features to address the fact that some counties will be affected more severely than others, for instance, due to (a) their population makeup, for example, older populations are likely to experience a higher death rate than younger populations, (b) their hospital preparedness, for example, 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 (Rubinson et al., 2010), and (c) their population health, for example, age, smoking history,

diabetes, and cardiovascular disease are all considered to be likely risk factors for acute COVID-19 infection (Fei et al., 2020; Goh et al., 2020; Guan, Liang, et al., 2020; Guan, Ni, et al., 2020; Qi et al., 2020).

For a county c, given a set of demographic and health care–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

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

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 health care 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 manner very similar to the shared predictor (3.3): We first obtain the estimate $\widehat{\mathrm{E}}[\mathrm{deaths}_{t+1}^c|t]$ using (3.5) and then, sequentially substitute $\log(\widehat{\mathrm{E}}[\mathrm{deaths}_{t+j}^c|t] + 1)$ on the RHS of the (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 $j = 1, \ldots, k-1$. We found that regularization was quite helpful in addressing overfitting in this predictor and found that ℓ_1-penalized Poisson regression with a penalty of 0.5 performed the best.

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 Schuller et al. (2002). 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.[7] Then, their CLEP is given by

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

[7]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 $t+1$, which was 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:

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

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.

The weights in (3.7) are based on the general ensemble weighting format introduced in Schuller et al. (2002). This general format is given by

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

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 hyperparameters: Formula (3.7) corresponds to (3.8) with appropriate hyper-parameters, c, μ, t_0, and a specific loss format, ℓ. In Schuller et al. (2002), the authors used the loss function $\ell(\widehat{y}_i^m, y_i) = |\widehat{y}_i^m - y_i|$, since their errors roughly followed 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 find that this transformation improves 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.[8]

To generate our predictions, we use the default value of c in Schuller et al. (2002), 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 choose $t_0 = t-6$ (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.

[8]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 formula in the article, and erroneously reported that our CLEP weights used a logarithmic transform.

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; that is, 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 hyperparameters—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 hyperparameters (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 nondecreasing sequence. However, the predictors discussed in the previous sections need not provide monotonic estimates for different prediction horizon, that is, $\widehat{\mathrm{E}}[\mathrm{deaths}^c_{t+k}|t]$ 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, that is, $\widehat{\mathrm{E}}[\mathrm{deaths}^c_{t+k}|t] < \mathrm{deaths}^c_t$. In our setting, the 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 $\widehat{\mathrm{E}}[\mathrm{deaths}^c_{t+1}|t]$ by $\max\{\widehat{\mathrm{E}}[\mathrm{deaths}^c_{t+1}|t], \mathrm{deaths}^c_t\}$ 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 $\widehat{\mathrm{E}}[\mathrm{deaths}^c_{t+j}|t]$ by $\max\{\widehat{\mathrm{E}}[\mathrm{deaths}^c_{t+j}|t], \widehat{\mathrm{E}}[\mathrm{deaths}^c_{t+j-1}|t]\}$ for $j = 2, 3, \ldots 21$. Imposing these constraints for the individual predictors also ensures that the predictions made by the CLEP are monotone. Note that we use these monotonic predictions (after the maxima calculations) to determine the weights in (3.7).[9]

Note that even after imposing the previous monotonicity corrections, it is still possible that $\widehat{\mathrm{E}}[\mathrm{deaths}^c_{t+k}|t] > \widehat{\mathrm{E}}[\mathrm{deaths}^c_{t+k+1}|t+1]$ since the predictors are refitted over time. Hence, when plotted over time, k-day-ahead predictions, need not be monotonic 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.

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 this article 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 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 (Marchant et al., 2020) reported that the 95% credible intervals for state-level daily mortality predicted by the initial model by the Institute for Health Metrics and Evaluation (IHME, 2020), 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

[9]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.

so that they now provide more than 95% coverage (where coverage is defined below in (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 (Vovk et al., 2005) 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 a slightly abused version of that used in Section 3.6. To be precise, we define $\widehat{y}_t = \widehat{\mathrm{E}}[\mathrm{deaths}_t^c | t-k]$ for some fixed county c.) We define the normalized absolute error, Δ_t, of the prediction, $\widehat{y}_t$, to be

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

We use the normalization so that y_t (when nonzero) is equal to either $\widehat{y}_t(1-\Delta_t)$ or $\widehat{y}_t(1+\Delta_t)$. This normalization addresses the fact that the counts are increasing over time, causing the unnormalized errors, $|y_t - \widehat{y}_t|$, to also increase over time, violating the exchangeability of errors condition (see Section 4.3). This normalization thus ensures that the errors across time are comparable in magnitude, which is essential for the exchangeability of errors.

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 $\widehat{y}_{t+k}$ $(= \widehat{\mathrm{E}}[\mathrm{deaths}_{t+k}^c | t])$ using a CLEP. Next, we compute the normalized errors for the k-day-ahead predictions for the most recent 5 days $\Delta_t, \Delta_{t-1}, ..., \Delta_{t-4}$ (a window of 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:

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

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

where the lower bound for the interval involves a maximum operator to account for the fact that y_t is a cumulative count, and thereby nondecreasing. 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

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

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 (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 sufficiently long 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, that is, have a good coverage, and have a reasonable width or length.[10] 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: *coverage* and *normalized length*.

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, Coverage($\mathcal{T}$), over a specified period, $\mathcal{T}$, corresponds to the fraction of days in this period for which the prediction interval contained the observed values of y_t (cumulative COVID-19 death counts in our case). This notion of *coverage* for streaming data has been used extensively in prior works on conformal inference (Vovk et al., 2005) and can be calculated for a given evaluation period $\mathcal{T}$ (which we set to be from April 11 to June 20) as follows:

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

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

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

For our case, we replace the denominator on the RHS of (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 (4.4a) and the average length (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 (Shafer & Vovk, 2008; Vovk et al., 2005), 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 make use of only 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 realistic. Due to the dynamic nature of COVID-19, considering a longer period (e.g., substantially longer than five

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

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 nonexchangeability by removing the sequential growth of the errors resulting from the increasing nature of the counts themselves. Since we use only five time points, to construct the interval, we opt for the more conservative choice of simply taking the maximum—or the 100th percentile—of the five errors (for instance, the 95th percentile is not well defined for five data points).

Before turning to certain theoretical guarantees, we first discuss some evidence for the exchangeability of the errors in our setting. In Figures 5 and B4, we show that the assumption of exchangeability of the past five *normalized* errors for the CLEP predictor is indeed reasonable for both 7-day-ahead and 14-day-ahead predictions. For a 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 the exchangeability of these six 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 note that the error can be computed only by the end of day $t-j$ (or the morning of day $t-j+1$). If the errors are exchangeable, then their rank has a uniform distribution on $\{1, 2, 3, 4, 5, 6\}$, with a mean of 3.5. To approximate this numerically for 7-day-ahead predictions, we calculate 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. Figure 5 plots the average rank of the six errors for each of the 6 worst affected counties as well as 6 randomly selected counties (see Section 5.2 for further discussion on these counties).

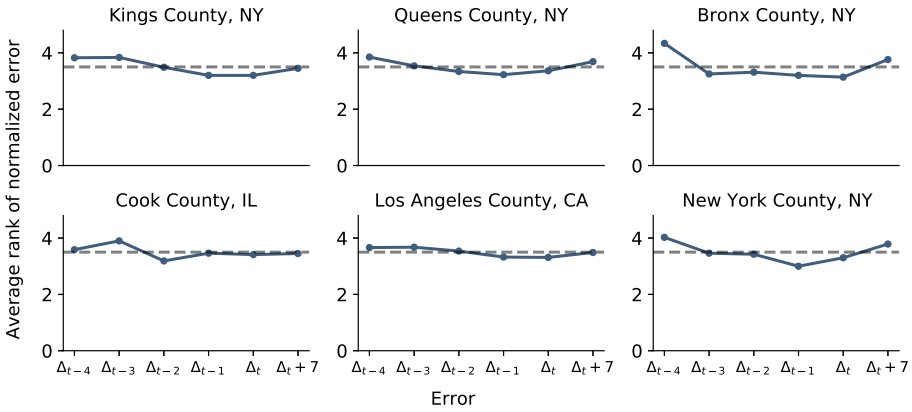
The errors in Figure 5 exhibit fairly consistent average ranks (the lines are fairly flat around 3.5), providing evidence for the exchangeability of the errors. We note that the ranks are discrete random variables, making it difficult and misleading to include standard errors in Figure 5. Consequently, we provide a visualization of the complete empirical distribution of the rank of these six errors over this period as heatmaps in Figure B2 in Appendix B. We observe no clear patterns from the heatmaps, providing further evidence for the exchangeability of the errors. We also provide corresponding results for 14-day-ahead predictions, where we rank the errors Δ_{t+14} and $\Delta_{t-j}, j = 0, \ldots, 4$ in Figure B4 in Appendix B.

The observations from Figures 5, B2 and B4 provide a heuristic justification for the construction of MEPI even though it is certainly not a formal proof (since the average ranks being approximately equal to 3.5 is not sufficient to prove exchangeability of the six errors). Moreover, we refer the interested reader to Appendix B.2 for further discussion on MEPIs, where we provide further discussion on our choices of considering only the past 5 errors and normalization to define MEPI (see Figure B3).

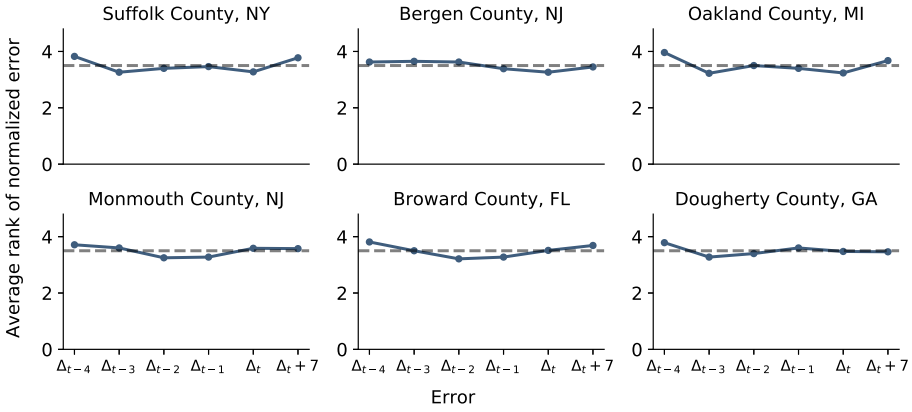
4.4. **Theoretical guarantees for MEPI coverage.** In order to obtain a rough baseline coverage for the MEPIs, we now reproduce some of the theoretical computations from the conformal literature. For a given county and a fixed time t, and a parameter k, if the six errors in the set $\{\Delta_{t+k}, \Delta_t,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

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

Recall the definition (4.4a) for Coverage($\mathcal{T}$) for a given period of days $\mathcal{T}$. Given (4.5), we may believe that Coverage($\mathcal{T}$) $\approx 83\%$ holds for large $|\mathcal{T}|$, where the coverage was defined in (4.4a).



(a) Six worst-affected counties



(b) Six randomly selected counties

Figure 5. Plots for investigating exchangeability of normalized errors of *7-day-ahead* CLEP predictions based on the previous 5 errors made at time t, over the period t = March 26, ..., Jun 13 (80 days). 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). We provide a heatmap visualization of the complete distribution of the ranks in Figure B2.

However, a few challenges remain to take claim (4.5) as a proof for the stronger claim that MEPI achieves 83% coverage as defined by (4.4a).

On the one hand, the probability in (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 $\widehat{\text{PI}}_{t+k}$ 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 (recall the definition (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 (Shafer & Vovk, 2008), has been crucial for establishing theoretical guarantees in conformal inference. In our case, this law—stated as Proposition 1 in Section 3.4 in (Shafer & Vovk, 2008)—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 (exchangeability over the entire period) 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, we later show in Section 5.3 that MEPIs with CLEP achieved good coverage with narrow widths for COVID-19 cumulative death count predictions.

5. Prediction results for March 22 to June 20

In this article, 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 (which are also the top two performing individual predictors). Finally, we report the performance of the coverage and length of the MEPIs for this CLEP. Note that the CLEP that combined all five predictors performed worse than the CLEP that combined the best two predictors since it tended to contain poor predictors with nonzero weights. A Python script containing the code that produced the results in this section is 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.

To compute these errors, on the morning of day $t+1$ we first create $\mathcal{C}_t$—the collection of counties in the United States 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 nontrivial 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

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

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

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

The percentage error (MAPE) captures the relative errors of the predictors independently of the scale of the counts (whereas 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 for which we report the 10th percentile (p10), 50th percentile (median), and 90th percentile (p90) of the errors for the k-day-ahead predictions over the period $t \in \{\text{March } 22, \ldots, \text{June } 20\}$ for $k \in \{3, 5, 7, 14\}$.[11] 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, respectively.[12]

The large (p90) MAEs for the separate (exponential) model for the larger time horizons can be attributed to the fact that exponential fit in (3.1) tends to overpredict for large horizons. On the other hand, the large errors for the demographics shared predictor are potentially due to overfitting as well as the recursive susbtitution procedure to obtain longer horizon estimates in the exponential fit in (3.5). In the early stages of this project (and the COVID-19 outbreak in the United States), these predictors provided a reasonable fit for short-term (3- and 5-day-ahead) predictions in late March to mid-April, but as the pandemic has evolved over time, the quality of their performance has decreased.

In Figure 6, we plot all three errors from 5.1(a)-(c) as a function of time over March 22 to June 20 for the expanded shared exponential predictor, the separate county linear predictor, and the CLEP that combines the two. The adaptive weighting scheme from (3.7) results in the CLEP having a MAE that is usually similar to, or smaller than, the MAE of the two single predictors.

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

[12] In the first version of this article submitted on May 16, 2020, on arXiv, we reported results for the period March 22 to May 10 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.

	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	14.80	24.08	106.97	23.08	37.60	>1000	31.52	51.52	>1000	89.71	190.97	>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	21.81	39.61	77.78	44.57	79.36	596.04	93.83	147.66	>1000	656.55	>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

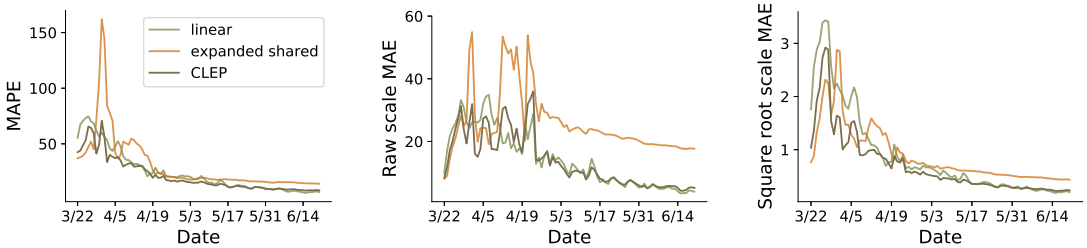
(B) Summary statistics of raw-scale MAE

	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	0.73	1.03	2.78	1.24	1.79	6.63	1.84	2.65	35.56	5.49	8.33	>1000
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
CLEP	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

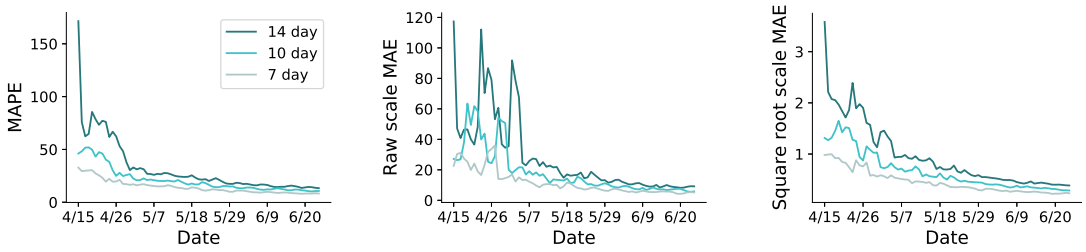
Table 6. Summary statistics of mean absolute errors (MAE) as defined in (5.1), based on **(A)** the mean absolute percentage error (MAPE), **(B)** the raw-scale MAE, and **(C)** the square-root-scale MAE. The results are presented for the 3, 5, 7, and 14-days-ahead forecasts for each of the predictors considered in this article, 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.

Figure 6. Plots of the mean absolute error (MAE) of three different predictors (the expanded shared exponential predictor, the separate county linear predictor, and the combined linear and exponential predictors [CLEP] that combines the two 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. The weights in (3.7) used to create the CLEP are adaptive over time.



(a) MAPE (b) Raw-scale MAE (c) Square-root-scale MAE

CLEP performance for longer horizons: Next, in Figure 7, we plot the performance of the best performing CLEP for longer horizons. In particular, we plot the three MAEs (raw-scale, percentage-scale, and square-root-scale), for $7-, 10-$, and 14-day-ahead predictions. The 7-day-ahead CLEP predictor has the lowest MAE and the MAE increases as the prediction horizon increases (aggregate statistics over time for 7-day-ahead and 14-day-ahead MAEs are reported in Table 6). The increases in MAE in mid-late April was 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 resulted in the CLEP greatly overpredicting deaths in New York in mid-late April 7 days later (on April 21) for the 7-day-ahead CLEP, 10 days later (on April 24) for the 10-day-ahead CLEP, and 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 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).



(a) MAPE (b) Raw-scale MAE (c) Square-root-scale MAE

Figure 7. Plots of (a) the mean absolute percentage error (MAPE), (b) the raw-scale mean absolute error (MAE), and (c) the square-root-scale MAE over time for the best performing CLEP based on 7-, 10-, and 14-day-ahead prediction horizons from April 11, 2020, to June 20, 2020.

We further evaluate the performance of the best performing CLEP for longer prediction horizons in Figure 8, where all predictions were made over the period April 11–June 20. We plot the distribution of raw-scale MAE of k-day ahead predictions for $k = \{3, 5, 7, 10, 14, 21\}$ as boxplots, and the median value of the raw scale MAE for $k = \{1, \ldots, 21\}$ as a line. From Figure 8, we observe that the MAEs increase roughly linearly with the horizon for up to 21 days.

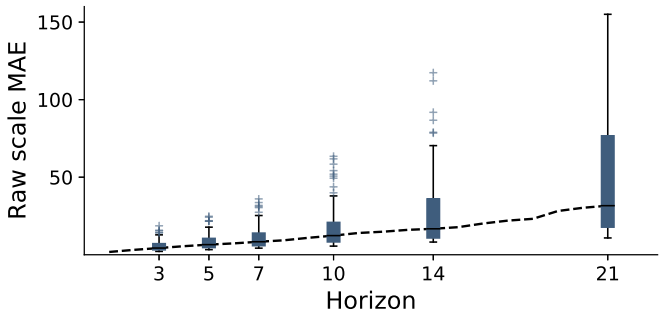


Figure 8. Box plots of the raw-scale mean absolute error (MAE) for k-day ahead predictions for $k = \{3, 5, 7, 10, 14, 21\}$, and the median value of the raw scale MAE for $k = \{1, \ldots, 21\}$ (as a dashed black line). For the 21-day-ahead prediction, the eight raw-scale MAEs, which are larger than 160, are not shown on the box plot since their magnitudes obfuscate the rest of the plot.

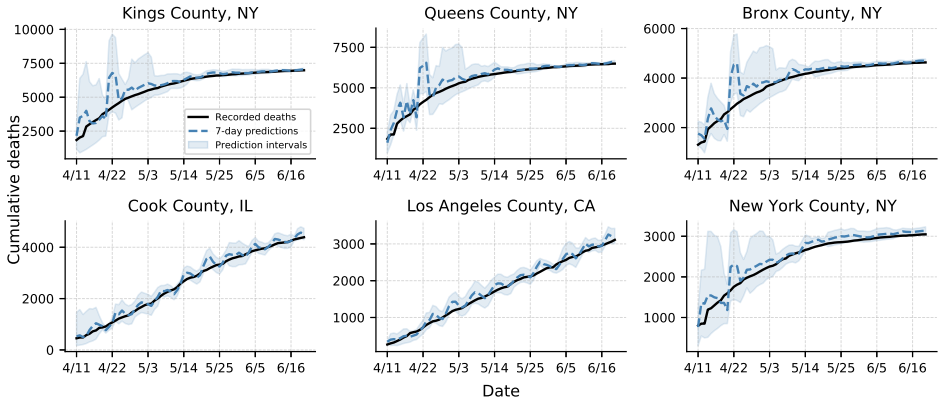
Putting together the results from Table 6, Figures 6, 7 and 8, we find that the adaptive combination used for building our ensemble predictor CLEP is able to leverage the advantages of linear and exponential predictors, and, by improving upon the MAE of single predictors, is able to provide very good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

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 April 11 – June 20 for the 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:

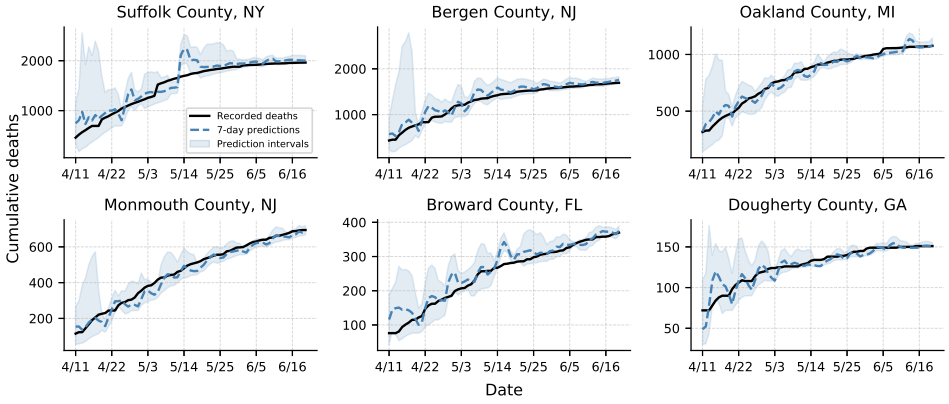
- The six worst-affected counties on June 20: Kings County, NY; Queens County, NY; Bronx County, NY; Cook County, IL; Los Angeles County, CA; and New York County, NY.
- Six randomly selected counties: Suffolk County, NY; Bergen County, NY; Oakland County, MI; Monmouth County, NJ; Broward County, FL;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

7- and 14-day-ahead predictions: In Figure 9, we present the 7-day-ahead prediction results for the six worst-affected counties in the top panel (a), and for the six randomly selected counties in the bottom panel (b). The solid black line denotes the recorded death counts, the dashed blue line denotes the CLEP 7-day-ahead predictions, and the 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 (blue dashed lines in Figures 9 and 10) need not be, since they are updated daily. To understand why, recall Section 3.7 where we discuss the monotonicity constraints for predictions made on a given day t ($\widehat{\mathrm{E}}[\mathrm{deaths}^c_{t+k+1}|t] \geq \widehat{\mathrm{E}}[\mathrm{deaths}^c_{t+k}|t]$), but, it is possible to have nonmonotonic trends for predictions made on different days, for example, we can have $\widehat{\mathrm{E}}[\mathrm{deaths}^c_{t+k}|t] > \widehat{\mathrm{E}}[\mathrm{deaths}^c_{t+k+1}|t+1]$.

From Figure 9(a), we observe that, among the worst-affected counties, the CLEP appears to fit the data very well for Cook County, Illinois, and Los Angeles County, California. After initially



(a) Worst-affected counties on June 20



(b) randomly selected counties

Figure 9. A grid of line charts displaying the *7-day-ahead* performance of the best performing combined linear and exponential predictors (CLEP) and maximum (absolute) error prediction intervals (MEPI) for the cumulative death counts due to COVID-19 between April 11, 2020, and June 20, 2020. The observed data is shown as black lines, the CLEP predictions are shown as dashed blue lines, and the corresponding 7-day-ahead MEPIs are shown as shaded blue regions. In panel **(a)**, the MEPI coverage for Kings County is 92%, Queens County is 80%, Bronx County is 90%, Cook County is 90%, Los Angeles County is 89%, and New York County is 86%. In panel **(b)**, the MEPI coverage for Suffolk County is 85%, Bergen county is 96%, Oakland county is 87%, Monmouth county is 86%, Broward County is 89%, and Dougherty County is 86%. It is worth noting that the theoretical results guarantee a coverage of 83% under certain assumptions – see (4.5).

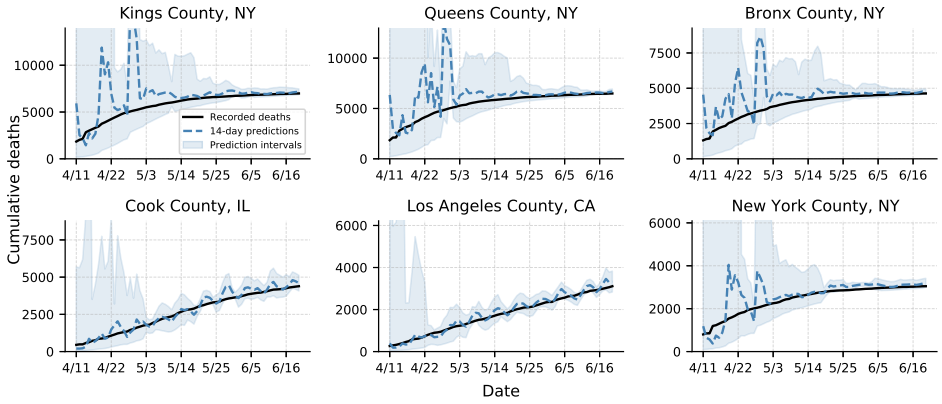
overpredicting the number of deaths in the four New York counties in mid-late April (recall our discussion on the conversion of 'probable' deaths to confirmed deaths in New York on April 14), our predictor also performs very well on these New York counties. Moreover, the MEPIs have reasonable width and appear to cover the recorded values for the majority of the days (detailed results on MEPIs are presented in Section 5.3).

From Figure 9(b), we find that our predictors and MEPI also 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, for Suffolk County, New York, there is a sudden uptick in cumulative deaths on May 5, leading to a fluctuation in the predictions shortly thereafter (similar to the New York 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 wider MEPIs for the earlier dates.

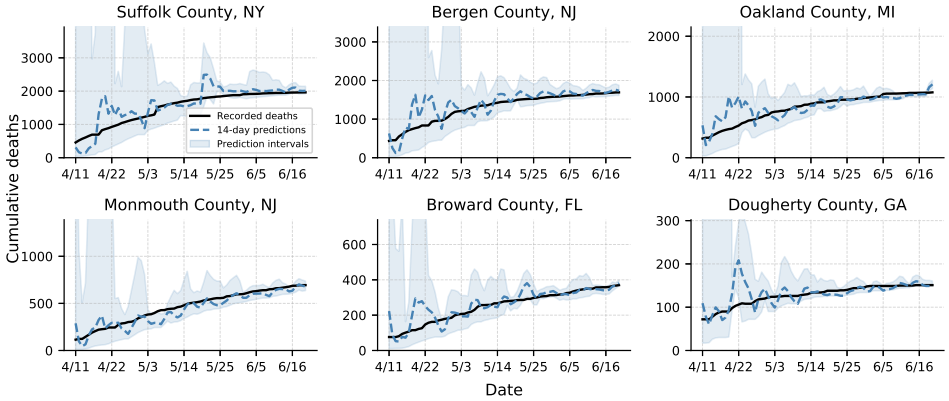
Figure 10 similarly plots the 14-day-ahead CLEP and MEPI predictions for the same two sets of counties discussed above. While the CLEP predictions appear to be fairly similar to the observed cumulative death counts, and the MEPIs display good coverage overall, the MEPIs are generally wider, especially at the beginning of the prediction period. These wider MEPIs are not surprising given the fact that before mid-April, 14-day-ahead predictions were based on data with very few death counts, which led the expanded shared predictor to significantly underestimate the death counts in the beginning of the observation period period (and thus leading to larger normalized errors and wider MEPIs). Furthermore, we again observe that for the counties in New York State, CLEP greatly overestimates the cumulative recorded death counts toward the end of April due to the upward revision of the death counts in New York State on April 14. Note that these sharp peaks in Figure 10 occur on 7 days of the corresponding sharp peaks for the 7-day-ahead CLEP predictions in Figure 9. These observations also suggest that one may possibly use large errors from our predictors as a warning flag for anomaly or reporting error/sudden revision in the data. We leave a detailed investigation on this aspect as an interesting future direction.

Visualization of CLEP weights: Figure 11 plots the weight of the linear predictor in the combined linear and exponential predictors (CLEP) over time for the same counties we have considered thus far. These weights were defined in Section 3.6, and the same set of weights were used for all of the k-day-ahead predictions; recall (3.6), (3.7), and (3.8). We found that for counties with a large number of cumulative deaths, the prediction of the CLEP is more similar to the linear predictor in late May and June than in March and April. For example, for the six worst affected counties on June 20 (Figure 11(a)), the average weight of the linear predictor in the CLEP is larger than 0.91 from May 17 to June 20. In contrast, the average weight of linear predictor of these six counties is less than 0.5 from March 23 to March 31. The weights appear unstable in Late March-early April reflecting the fact that it was harder to reliably predict deaths earlier in the outbreak due to low death counts, and that unexpected modifications to the counts (e.g. upticks) can lead to temporarily unstable predictions. However, in Figure 11, we observe that the Suffolk county weights exhibit surprising variability around May 10. Once again, these sudden changes are caused by revisions in the COVID-19 death counts in this county.

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 (4.4a) and average normalized length (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 article presented results until May 10, and thus the period May 11 to June 20 serves as an additional validation. Recall that



(a) Worst-affected counties on June 20

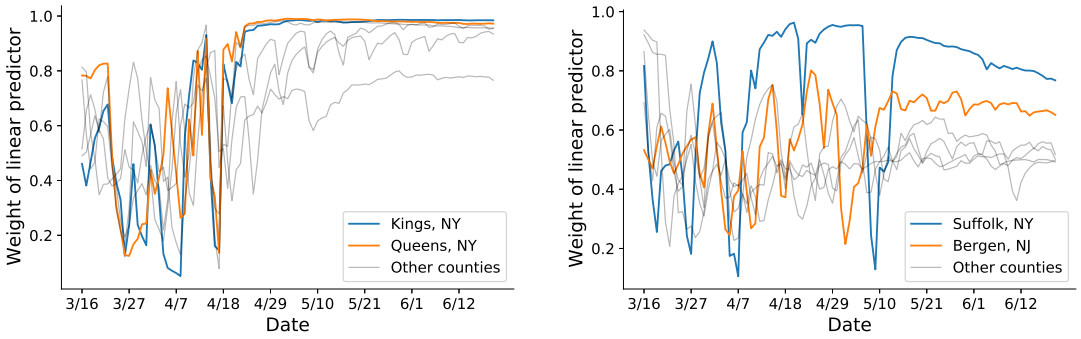


(b) Randomly selected counties

Figure 10. A grid of line charts displaying the *14-day-ahead* performance of the best performing combined linear and exponential predictors (CLEP) and maximum (absolute)error prediction intervals (MEPI) for the cumulative death counts due to COVID-19 between April 11, 2020 and June 20, 2020. The observed data is shown as black lines, the CLEP predictions are shown as dashed blue lines, and the corresponding 14-day-ahead MEPIs are shown as shaded blue regions. In panel **(a)**, the MEPI coverage for Kings County is 99%, Queens County is 99%, Bronx County is 99%, Cook County is 93%, Los Angeles County is 96%, and New York County is 89%. In panel **(b)**, the MEPI coverage for Suffolk County is 92%, Bergen County is 99%, Oakland County is 89%, Monmouth County is 94%, Broward County is 97%, and Dougherty County is 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 at the beginning of the observation period.

our methods were proposed and tuned prior to May 10, except for the transformation change from logarithm to square-root in (3.7).

We evaluate the 7-day-ahead and 14-day-ahead MEPIs, that is, $k = 7$ and 14 in (4.2a), for the CLEP that combines the expanded shared and separate linear predictors, and summarize the results in Figure 12 and Figure 13, respectively.



(a) Worst affected counties (b) Random counties

Figure 11. Line plots displaying the weights of the linear predictor in the combined linear and exponential predictors (CLEP) over time for the six worst affected counties in panel **(a)**, and six random counties in panel **(b)**. To make the plots more easily digestible, we highlight only two counties in each plot, displaying the remaining four counties in the background. In panel (a), the other four counties are Bronx County, NY; Cook County, IL; Los Angeles County, CA; and New York County, NY. In panel (b), the other four counties are Oakland County, MI; Monmouth County, NJ; Broward County, FL; and Dougherty County, GA.

Coverage: We compute the observed coverage from (4.4a) of the 7-day-ahead MEPIs across all counties in the United States for April 11–May 10, and May 11–June 20, and plot the histogram of these values in Figure 12(a) and 12(c), respectively. Panel (e) of Figure 12 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 only the days between April 11 and June 20 for which the county had at least 10 cumulative deaths. On June 20, the median number of days since 10 deaths is 58. From these plots, we observe that we achieve excellent coverage for the majority of the counties. Finally, Figure 13 shows the corresponding results for 14-day-ahead MEPIs, that is, coverage for April 11–May 10 in panel (a), May 11–June 20 in panel (c), and over the county-specific period for 700 *hard-hit* counties in panel (e), where we call a county hard-hit if it had at least 10 cumulative deaths by June 11.

The coverage observed for the two periods in panels (a) and (c) of Figure 12 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 decreases slightly 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, which had very good coverage with the prediction interval. On the flip side, the *normalized* interval lengths for the MEPIs for such counties would typically be larger.

Figure 13(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% (the median coverage is 100% for

both periods). In panel (e), for the hard-hit counties, which had at least 10 cumulative deaths by 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 during 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 article.

Normalized length: Next, we discuss the other evaluation metric of the MEPIs, their normalized length from (4.4b). In panels (b) and (d) of Figure 12, 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 700 hard-hit counties as did panel (e): those with at least 10 deaths for at least 10 days in the period April 11 to June 20 (or, equivalently, counties with a cumulative death count of at least 10 on June 11).

Recall that the normalized length is defined as the length of the MEPI over the recorded number of deaths (4.4b), and note that more than 70% of counties in the United States recorded two or less COVID-19 deaths by May 1. For these counties, a normalized length of 2 corresponds to an unnormalized prediction interval length of 4. It is thus not surprising to see that the average normalized length of MEPI for a nontrivial 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 for these (county-specific) periods is much smaller (the median is 0.470).

Turning to 14-day-ahead MEPIs in Figure 13, panels (b) and (d) show that that the normalized length for the 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 the hard-hit counties with more than 10 deaths, with a median average normalized length of 1.027—but this is roughly twice the median size of 0.470 for the 7-day-ahead MEPIs in Figure 12(f).

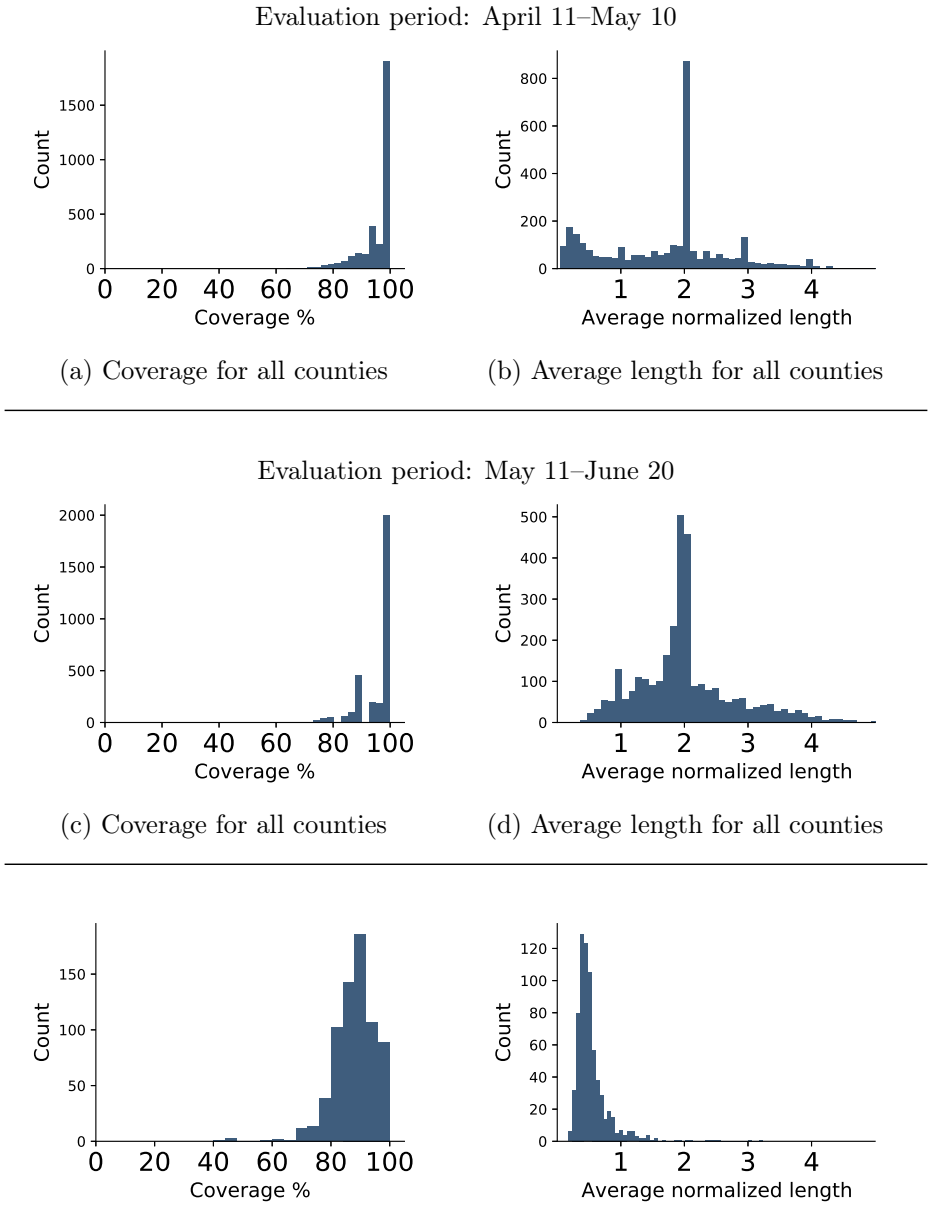
Overall, Figures 12 and 13 show that our MEPIs provide a reasonable balance between coverage and length for up to 14 days in future, especially when the cumulative counts are not too small.

6. Related work

Several recent works have tried to predict the number of cases and deaths related to COVID-19. Even more recently, the CDC has started aggregating results from several models.[13]. But, to the best of our knowledge, ours was the first work focusing on predictions at the county level. During the time period of our work, a direct comparison with other models to our own was difficult for several reasons: (1) other existing models typically relied on strong assumptions and the authors did not usually provide prediction errors (on a validation set) on their models, (2) we did not have access to a direct implementation of their models (or results), and (3) their models focused on substantially longer time horizons than ours does, making the results difficult to compare.

Two recent works (Ferguson et al., 2020; Murray & COVID-19 Health Service Utilization Forecasting Team, 2020) focused on modeling death counts at the state level in the United States. The earlier versions of the model by Murray et al. (also referred to as the IHME model) were based on Farr's Law with feedback from Wuhan data. In the 1990s, Bregman and Langmuir (1990) used Farr's law to predict that the total number of cases from the AIDS pandemic would diminish by the

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

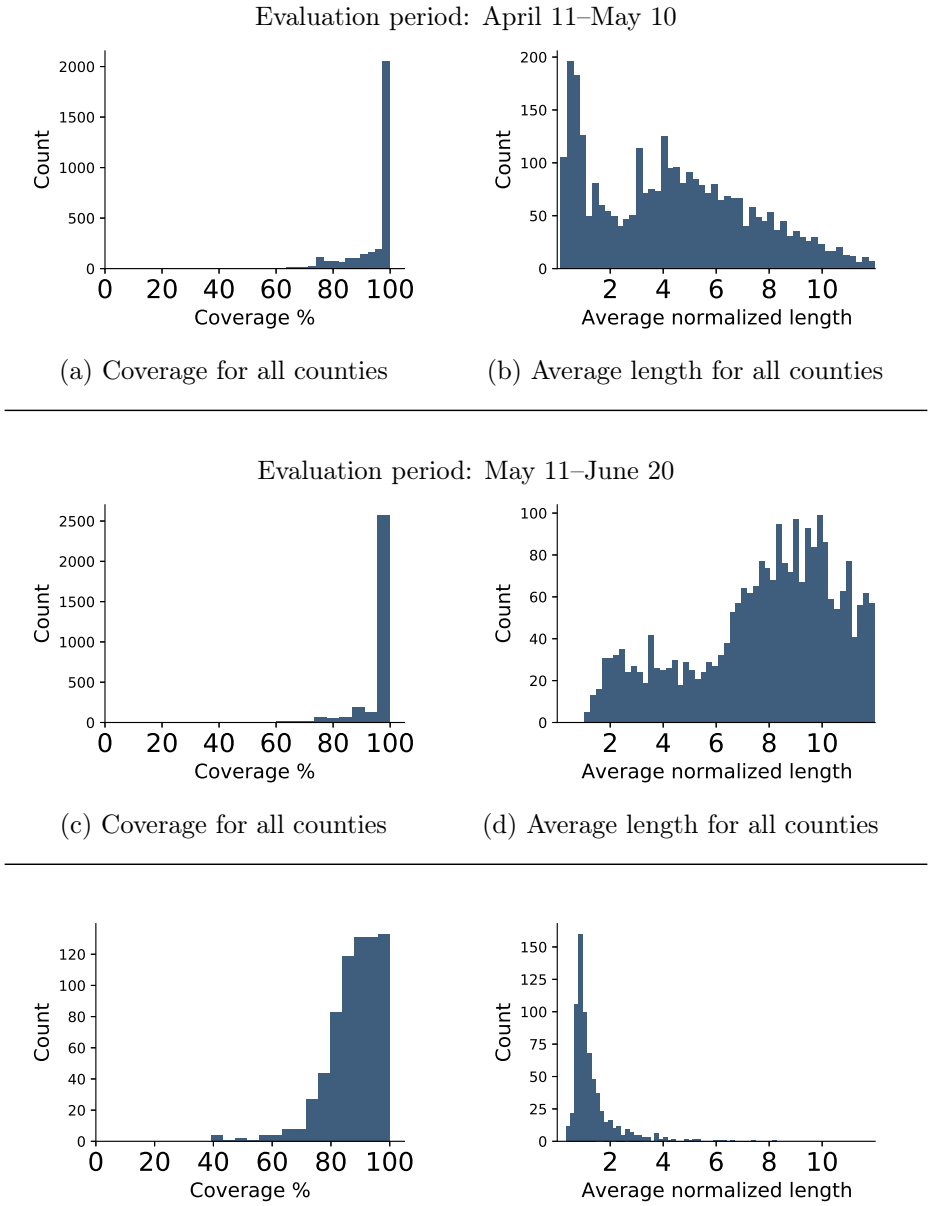


(a) Coverage for all counties (b) Average length for all counties

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

(e) Coverage for hard-hit counties (f) Average length for hard-hit counties

Figure 12. Histograms showing the performance of *7-day-ahead maximum (absolute) error prediction intervals* for county-level cumulative death counts. For each county, we compute the observed coverage and average normalized length for the period April 11–May 10, 2020 and plot the histogram of these values across counties respectively in the top panels (a) and (b). Respective results for the period May 11–June 20, 2020 are plotted in the middle panels (c) and (d). For the bottom two panels (e) and (f), we plot the histogram across 700 *hard-hit* counties where we call a county hard-hit if it had at least 10 cumulative deaths by June 11. In these two panels, for each county the observed coverage and average normalized length are computed only over those days in April 11 to June 20 for which that county's cumulative death count is at least 10. See Figure 13 for similar plots for 14-day-ahead MEPIs.



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

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

(e) Coverage for hard-hit counties (f) Average length for hard-hit counties

Figure 13. Histograms showing the performance of the *14-day-ahead maximum (absolute) error prediction intervals intervals* for county-level cumulative death counts. For each county, we compute the observed coverage and average normalized length for the period April 11–May 10, 2020 and plot the histogram of these values across counties respectively in the top panels **(a)** and **(b)**. Respective results for the period May 11–June 20, 2020 are plotted in the middle panels **(c)** and **(d)**. For the bottom two panels **(e)** and **(f)**, we plot the histogram across 700 *hard-hit* counties where we call a county hard-hit if it had at least 10 cumulative deaths by June 11. In these two panels, for each county the observed coverage and average normalized length are computed only over those days in April 11 to June 20 for which that county's cumulative death count is at least 10. See Figure 12 for similar plots for 7-day-ahead MEPIs.

mid-1990s and the total number of cases in the United States would be around 200,000. However, by 2018, it is estimated that at least 1,200,000 (1.2 million) people have had HIV in the United States, and at least 700,000 perished from it (as per the data from Centers for Disease Control and Prevention, 2018b). While the AIDS pandemic is very different from the COVID-19 pandemic, it is still useful to keep this historical performance of Farr's law in mind. The second approach, by Ferguson et al. (2020), uses an individual-based simulation model with parameters chosen based on prior knowledge.

Another approach uses exponential smoothing from time-series predictors to estimate day-level COVID-19 cases at the national level (Elmousalami & Hassanien, 2020). In addition, several works use compartmental epidemiological predictors such as SIR, SEIR, and SIRD (Becker & Chivers, 2020; Fanelli & Piazza, 2020; Pei & Shaman, 2020) to provide simulations at the national level (where S, E, I, R and D respectively stand for susceptible-exposed-infectious-recovered-deceased individuals). Other works (Hsiang et al., 2020; Peak et al., 2020) simulate the effect of social distancing policies either in the future for the United States or in a retrospective manner for China. Finally, several papers estimate epidemiological parameters retrospectively based on data from China (Kucharski et al., 2020; Wang et al., 2020).

During the revision of this article, another work was published by Chiang et al. (2020) that appeared in medRxiv on June 8, 2020[14] after the submission of our article to arXiv in mid-May. Chiang et al. use models based on Hawkes's process to provide county-level predictions for new daily cases as well as new death counts. Of note is that the authors used an approach similar to ours, fitting a CLEP but 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 hyperparameters in the CLEP in our own future work. Unfortunately, we were unable to 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,[15] 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 predictions are not publicly available (such as on their GitHub repository) for either the period up to May 20 analyzed in their paper or in the period following. The authors do not report the performance of their confidence intervals in their 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, for example, once a week for the 7-day-ahead predictions (rather than daily as in this article). 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.

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

[15]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

Our models are being used to support the nonprofit Response4Life[16] in determining 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 use our forecasts to calculate the COVID pandemic severity index, a 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 U.S. hospitals with respect to (i) total deaths so far and (ii) predicted deaths for the next 7 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 is 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 article, 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 in planning and coping with the COVID-19 pandemic. 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.

Furthermore, our data repository as well as forecasting and interval methodology will be useful for academic purposes. Our data repository has already been used for data science education, and by 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, code and results can be found at https://covidseverity.com.

Lastly, inspired by the recent work of Chiang et al. (2020), we are beginning our investigation into adaptive tuning (over time) of μ, c and other hyperparameters for CLEP, in the hope of improving its performance.

Acknowledgments

Bin Yu acknowledges the support of CITRIS Grant 48801, a research award by Amazon, and Amazon web services (AWS). The authors would like to thank many people for help with this effort. Our acknowledgement section is unusually long because it reflects the nature of an ER-like collaborative project. We were greatly energized by the tremendous support and the outpouring of

[16] https://response4life.org/.

help that we received not only from other research groups, but also high school students, medical staff, ER doctors, and several volunteers who signed up at Response4Life.

We would like to first thank the Response4Life team (Don Landwirth and Rick Brennan in particular), and volunteers for building the base for this project. We would also like to thank Max Shen's IEOR group at Berkeley: Junyu Cao, Shunan Jiang, Pelagie Elimbi Moudio for helpful inputs in the early stages of the project. We thank Aaron Kornblith and David Jaffe for advice from a medical perspective, especially Aaron for a great deal of useful feedback. We want to mention special thanks to Sam Scarpino for sharing data with us.

We would like to thank Danqing Wang, Abhineet Agarwal, and Maya Shen for their help in improving our visualization website https://covidseverity.com over the summer, and support from Google, in particular, Cat Allman and Peter Norvig. We would also like to thank the high school students Matthew Shen, Anthony Rio, Miles Bishop, Josh Davis, and Dylan Goetting for helping us to collect valuable hospital related information.

Finally, we acknowledge helpful input on various data and modeling aspects from many others including Ying Lu, Tina Eliassi-Rad, Jas Sekhon, Philip Stark, Jacob Steinhardt, Nick Jewell, Valerie Isham, Sri Satish Ambati, Rob Crockett, Marty Elisco, Valerie Karplus, Marynia Kolak, Andreas Lange, Qinyun Lin, Suzanne Tamang, Brian Yandell and Tarek Zohdi. We also acknowledge many critical and constructive comments from anonymous referees, an AE, the editor-in-chief, and the dataviz editor. Their comments have led to the discussion on data biases, additional substantive results, improved visualization, and overall enhanced readability.

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

References

Angelopoulos, A. N., Pathak, R., Varma, R., & Jordan, M. I. (2020). On identifying and mitigating bias in the estimation of the covid-19 case fatality rate [https://hdsr.mitpress.mit.edu/pub/y9vc2u36]. *Harvard Data Science Review*, *0*. https://doi.org/10.1162/99608f92.f01ee285

Apple Inc. (2020). Apple Mobility Trends Reports. https://www.apple.com/covid19/mobility

Becker, M., & Chivers, C. (2020). Announcing CHIME, a tool for covid-19 capacity planning. http://predictivehealthcare.pennmedicine.org/2020/03/14/accouncing-chime.html

Bregman, D. J., & Langmuir, A. D. (1990). Farr's law applied to AIDS projections. *JAMA*, *263*(11), 1522–1525. https://doi.org/10.1001/jama.1990.03440110088033

Bureau of Transportation Statistics. (2020). Airline origin and destination survey (db1b). https://transtats.bts.gov/Databases.asp?Mode_ID=1&Mode_Desc=Aviation&Subject_ID2=0

Centers for Disease Control and Prevention. (2018a). HIV in the United States and Dependent Areas. http://nccd.cdc.gov/DHDSPAtlas

Centers for Disease Control and Prevention. (2018b). Interactive Atlas of Heart Disease and Stroke. https://www.cdc.gov/hiv/statistics/overview/ataglance.html

Centers for Disease Control and Prevention, Agency for Toxic Substances and Disease Registry, & Geospatial Research, Analysis, and Services Program. (2018). Social Vulnerability Index Database. https://svi.cdc.gov/data-and-tools-download.html

Centers for Disease Control and Prevention, Division of Diabetes Translation, & US Diabetes Surveillance System. (2016). Diagnosed diabetes atlas. https://www.cdc.gov/diabetes/data

Centers for Medicare & Medicaid Services. (2017). Chronic Conditions Prevalence State/County Level: All Beneficiaries by Age, 2007-2017. https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Chronic-Conditions/CC_Main

Centers for Medicares & Medicaid Services. (2018). Case mix index file. https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/FY2020-IPPS-Final-Rule-Home-Page-Items/FY2020-IPPS-Final-Rule-Data-Files

Centers for Medicares & Medicaid Services. (2020). 2020 reporting cycle: Teaching hospital list. https://www.cms.gov/OpenPayments/Downloads/2020-Reporting-Cycle-Teaching-Hospital-List-PDF-.pdf

Chiang, W.-H., Liu, X., & Mohler, G. (2020). Hawkes process modeling of COVID-19 with mobility leading indicators and spatial covariates. *medRxiv 2020.06.06.20124149*.

County Health Rankings & Roadmaps. (2020). County Health Rankings & Roadmaps 2020 Measures. https://www.countyhealthrankings.org/explore-health-rankings/measures-data-sources/2020-measures

Definitive Healthcare. (2020). Definitive Healthcare: USA Hospital Beds. https://coronavirus-resources.esri.com/datasets/definitivehc::definitive-healthcare-usa-hospital-beds

Dong, E., Du, H., & Gardner, L. (2020). An interactive web-based dashboard to track covid-19 in real time. *The Lancet Infectious Diseases*, *20*(5), 533–534.

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

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

Fei, Z., Ting, Y., Ronghui, D., Guohui, F., Ying, L., Zhibo, L., Jie, X., Yeming, W., Bin, S., Xiaoying, G., Lulu, G., Yuan, W., Hui, L., Xudong, W., Jiuyang, X., Shengjin, T., Yi, Z., Hua, C., & Bin, C. (2020). Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: A retrospective cohort study. *The Lancet*, 1054–1062. https://www.sciencedirect.com/science/article/pii/S0140673620305663

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

Goh, K. J., Kalimuddin, S., & Chan, K. S. (2020). 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.

Google LLC. (2020). Google COVID-19 Community Mobility Reports. %5Curl%7Bhttps://www.google.com/covid19/mobility/%7D

Granneman, J. (2020). Washington Department of Health clarifies COVID-19 death numbers. https://www.clarkcountytoday.com/news/washington-department-of-health-clarifies-covid-19-death-numbers/

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

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

Health Resources and Services Administration. (2019). Area Health Resources Files. https://data.hrsa.gov/data/download

Health Resources and Services Administration. (2020). Health Professional Shortage Areas - Primary Care. https://data.hrsa.gov/data/download

Homeland Infrastructure Foundation-Level Data. (2020). Hospitals. https://hifld-geoplatform.opendata.arcgis.com/datasets/6ac5e325468c4cb9b905f1728d6fbf0f_0

Hsiang, S., Allen, D., Annan-Phan, S., Bell, K., Bolliger, I., Chong, T., Druckenmiller, H., Hultgren, A., Huang, L. Y., Krasovich, E., Lau, P., Lee, J., Rolf, E., Tseng, J., & Wu, T. (2020). The effect of large-scale anti-contagion policies on the Coronavirus (COVID-19) pandemic. *medRxiv*. https://doi.org/10.1101/2020.03.22.20040642

Institute for Health Metrics and Evaluation. (2017). United States Chronic Respiratory Disease Mortality Rates by County 1980-2014. http://ghdx.healthdata.org/record/ihme-data/united-states-chronic-respiratory-disease-mortality-rates-county-1980-2014

Institute for Health Metrics and Evaluation. (2020). COVID-19: What's New for April 5, 2020. http://www.healthdata.org/sites/default/files/files/Projects/COVID/Estimation_update_040520_3.pdf

Kaiser Health News. (2020). ICU Beds by County. https://khn.org/news/as-coronavirus-spreads-widely-millions-of-older-americans-live-in-counties-with-no-icu-beds/

Katz, J., Lu, D., & Sanger-katz, M. (2020). U.S. Coronavirus death toll is far higher than reported, C.D.C. data suggests. *The New York Times*. https://www.nytimes.com/interactive/2020/04/28/us/coronavirus-death-toll-total.html

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

Kucharski, A. J., Russell, T. W., Diamond, C., Liu, Y., Edmunds, J., Funk, S., & Eggo, R. M. (2020). Early dynamics of transmission and control of COVID-19: A mathematical modelling study. *medRxiv*. https://doi.org/10.1101/2020.01.31.20019901

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

MIT Election Data and Science Lab. (2018). County Presidential Election Returns 2000-2016 [Blog]. https://doi.org/10.7910/DVN/VOQCHQ

Murray, C. J., & COVID-19 Health Service Utilization Forecasting Team, I. H. M. E. (2020). Forecasting COVID-19 impact on hospital bed-days, ICU-days, ventilator-days and deaths by US state in the next 4 months. *medRxiv*. https://doi.org/10.1101/2020.03.27.20043752

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

New York Times. (2020). COVID-19 Data in the United States. https://github.com/nytimes/covid-19-data

Nguyen, Q. P., & Schechtman, K. W. (2020). Confirmed and probable covid-19 deaths, counted two ways. *The COVID Tracking Project*. https://covidtracking.com/blog/confirmed-and-probable-covid-19-deaths-counted-two-ways

Peak, C. M., Kahn, R., Grad, Y. H., Childs, L. M., Li, R., Lipsitch, M., & Buckee, C. O. (2020). Modeling the comparative impact of individual quarantine vs. active monitoring of contacts for the mitigation of COVID-19. *medRxiv*. https://doi.org/10.1101/2020.03.05.20031088

Pei, S., & Shaman, J. (2020). Initial simulation of SARS-CoV2 spread and intervention effects in the continental US. *medRxiv*. https://doi.org/10.1101/2020.03.21.20040303

Qi, D., Yan, X., Tang, X., Peng, J., Yu, Q., Feng, L., Yuan, G., Zhang, A., Chen, Y., Yuan, J., Huang, X., Zhang, X., Hu, P., Song, Y., Qian, C., Sun, Q., Wang, D., Tong, J., & Xiang, J. (2020). Epidemiological and clinical features of 2019-nCoV acute respiratory disease cases in Chongqing municipality, China: A retrospective, descriptive, multiple-center study. *medRxiv*. https://doi.org/10.1101/2020.03.01.20029397

Rubinson, L., Vaughn, F., Nelson, S., Giordano, S., Kallstrom, T., Buckley, T., Burney, T., Hupert, N., Mutter, R., Handrigan, M., et al. (2010). Mechanical ventilators in US acute care hospitals. *Disaster Medicine and Public Health Preparedness*, *4*(3), 199–206. https://www.cambridge.org/core/journals/disaster-medicine-and-public-health-preparedness/article/mechanical-ventilators-in-us-acute-care-hospitals/F1FDBACA53531F2A150D6AD8E96F144D

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

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with python. *9th Python in Science Conference*.

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

Stokes, E. K., Zambrano, L. D., Anderson, K. N., Marder, E. P., Raz, K. M., Felix, S. E. B., Tie, Y., & Fullerton, K. E. (2020). Coronavirus disease 2019 case surveillance—United States, January 22–May 30, 2020. *Morbidity and Mortality Weekly Report*, *69*(24), 759.

United States Census Bureau. (2018). County Adjacency File. https://www.census.gov/geographies/reference-files/2010/geo/county-adjacency.html

United States Department of Agriculture, Economic Research Service. (2018). Poverty estimates for the U.S., states, and counties. https://www.ers.usda.gov/data-products/county-level-data-sets/download-data/

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

USAFacts. (2020). COVID-19 Deaths Data. https://www.reuters.com/article/us-health-coronavirus-who/covid-19-spread-map

Vovk, V., Gammerman, A., & Shafer, G. (2005). *Algorithmic learning in a random world*. Springer Science & Business Media.

Walker, A. S., Jones, L. W., & Gamio, L. (2020). Is the coronavirus death tally inflated? Here's why experts say no. *The New York Times*. https://www.nytimes.com/interactive/2020/06/19/us/us-coronavirus-covid-death-toll.html

Wang, C., Liu, L., Hao, X., Guo, H., Wang, Q., Huang, J., He, N., Yu, H., Lin, X., Pan, A., Wei, S., & Wu, T. (2020). Evolving epidemiology and impact of non-pharmaceutical interventions

on the outbreak of Coronavirus disease 2019 in Wuhan, China. *medRxiv*. https://doi.org/10.1101/2020.03.03.20030593

Wu, J., McCann, A., Katz, J., & Peltier, E. (2020). 364,000 missing deaths: Tracking the true toll of the Coronavirus outbreak. *The New York Times*. https://www.nytimes.com/interactive/2020/04/21/world/coronavirus-missing-deaths.html

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 utilized the new features only in the context of the combined linear and exponential predictors (CLEP) that combines the expanded shared and linear predictors.[17]

A.1. **Social-distancing feature.** Here we consider adding a social distancing feature to the expanded shared model (discussed in Section 3.4). We include an indicator feature in (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 choose 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 find it necessary to regularize this predictor since, without regularization, our 7-day-ahead predictions became infinite in some cases. We regularize this model with the elastic net and an equal penalty of 0.01 for both ℓ_1 and ℓ_2 regularization.

We now compare CLEP with the social-distancing feature included in the expanded shared predictor, with the original CLEP from the main article, for 7-day-ahead prediction of the recorded cumulative death counts. We find 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 (5.1) of 13.95, while the social-distancing variant has an MAE of 14.2. In Figure A1, we plot the behavior of raw-scale mean absolute errors (MAE) 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.

[17]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 redo the investigation with the monotonicity adjustment. We leave any further investigation with these new features (or their variants) for future work.

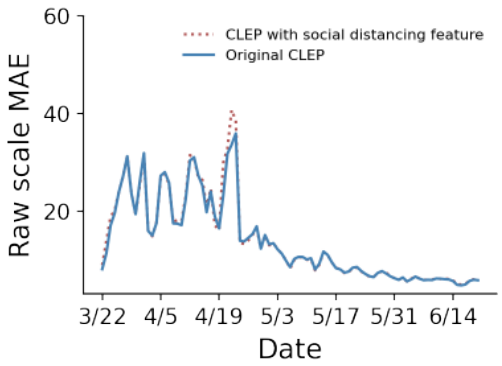


Figure A1. Plots of raw-scale mean absolute errors (MAE) for *7-day-ahead* predictions of two variants of combined linear and exponential predictors (CLEP) combining expanded shared and linear predictors: CLEP with a social distancing indicator feature for whether social distancing was in place in a county for more than 2 weeks or not, and the original CLEP considered in the main article. The social-distancing feature is included in the expanded shared model.

A.2. **Weekday feature.** As illustrated in Section 2.2 and Figure 3(a), the COVID-19 death counts are underreported on Sunday and Monday, which could potentially lead to increased errors for our prediction algorithm. As a result, we consider an additional feature for the two best predictors that we used in our CLEP earlier: the expanded shared and separate linear predictors. We now discuss the details of our investigations with these two predictors one by one.

A.2.1. *Weekday feature for expanded shared predictor.* To address this, we first investigate our expanded shared predictor's (3.4) performance for 3-day-ahead prediction on a per weekday basis and plot the results in Figure A2(a). We observe that 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 introduce an additional indicator feature in (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 A2(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.

We also experiment with a feature that accounted for the predicted day being either a Tuesday or Wednesday, days which are typically overcounted to compensate for undercounting on Sunday and Monday, but initial experiments did not provide promising results.

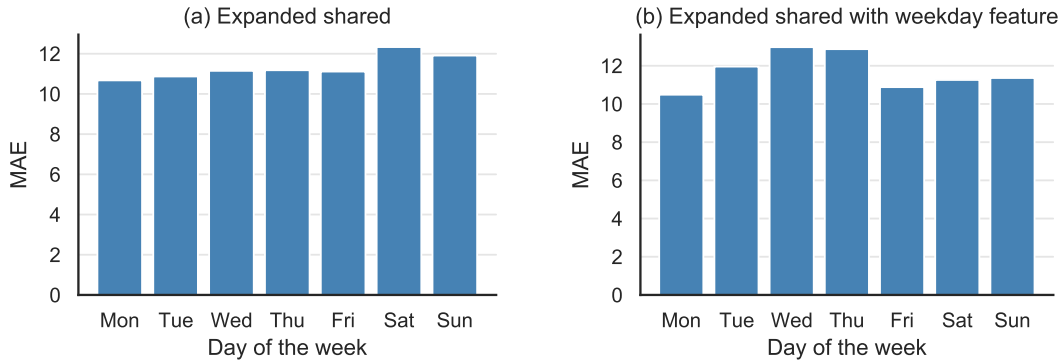


Figure A2. Mean raw-scale mean absolute errors (MAE) by weekdays for the expanded shared predictor for *3-day-ahead* predictions with and without the weekday feature. The MAE for a given day is the MAE for the 3-day-ahead prediction *computed on that day*. So the MAE for Wednesday is the MAE for predicting the cumulative deaths on Friday.

A.2.2. *Weekday feature for separate linear predictors.* 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

$$\widehat{\mathrm{E}}[\mathrm{deaths}^c_{t+1}|t] = \beta^c_0 + \beta^c_1(t+1) + \beta^c_2 v_{t+1}, \tag{A.1}$$

where v_{t+1} indicates whether day $t+1$ is Sunday/Monday or not. For 3-day-ahead predictions ($\widehat{\mathrm{E}}[\mathrm{deaths}^c_{t+3}|t]$), we simply replace $t+1$ by $t+3$, and v_{t+1} by v_{t+3} on the right-hand side of (A.1).

Recall that the original separate linear predictors in Section 3.2 were fit with only 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 $v_t = v_{t-1} = v_{t-2} = v_{t-3} = 0$, that is, the new feature is identically zero in the training data. For such cases, the parameter β^c_2 is not identifiable. To address this issue of nonidentifiability, 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 A3. 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. In addition to this Sunday/Monday feature, we incorporated a feature to account for the overcounting of deaths on Tuesday and Wednesday. However, we did not see any improvement in results from initial experiments, and thereby omit further details here.

Appendix B. Further discussion on MEPI

We now first shed light on why maximum (absolute) error prediction intervals (MEPI) exhibited slightly worse coverage for some of the counties. And then, we provide a further discussion on various choices made for designing MEPI.

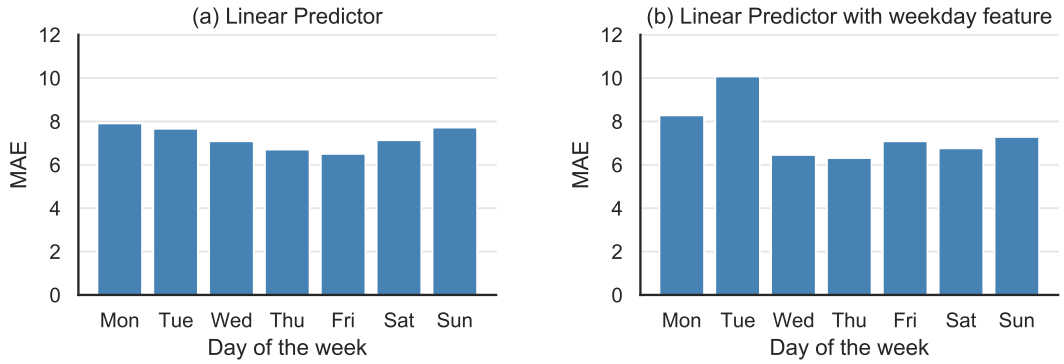


Figure A3. Mean raw-scale mean absolute errors (MAE) by weekdays for the separate linear predictors for *3-day-ahead* predictions with and without the weekday feature. The MAE for a given day is the MAE for the 3-day-ahead prediction *computed on that day*. So the MAE for Wednesday is the MAE for predicting the cumulative deaths on Friday.

B.1. **Counties with poor coverage.** While Figure 12(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. Among 198 such counties, Figure B1 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) recorded only two 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.

B.2. **MEPI vs conformal inference.** Recall that the MEPI (4.2a) can be viewed as a special case of conformal prediction interval (Shafer & Vovk, 2008; Vovk et al., 2005). Here, we provide further discussion on this connection. 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 an 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 (Shafer & Vovk, 2008) and the book (Vovk et al., 2005).

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 these is a more conservative choice (e.g., when compared to taking median or

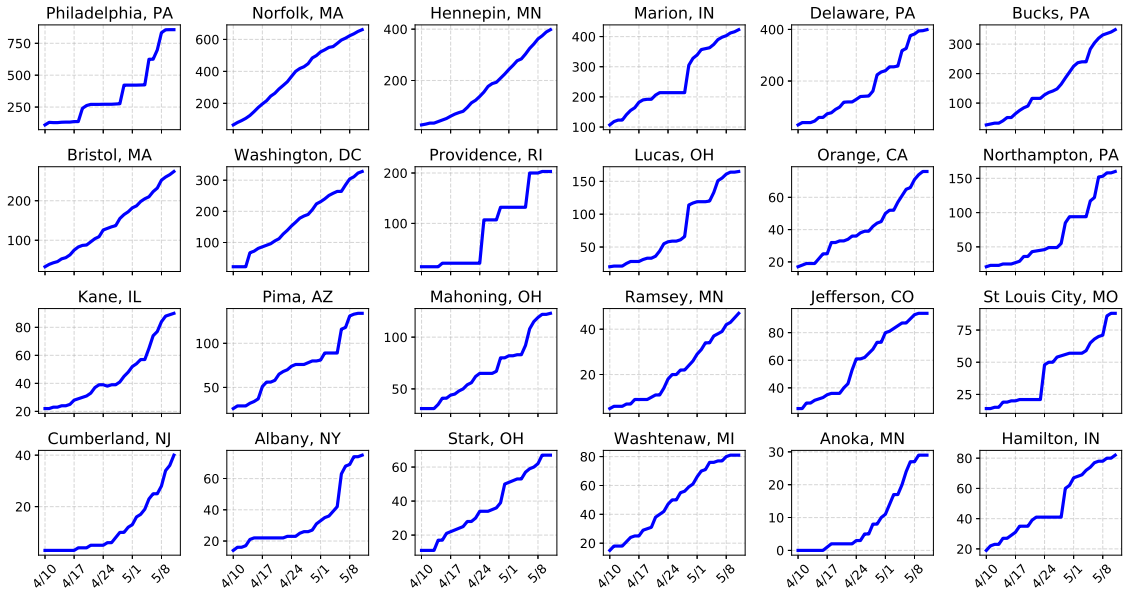


Figure B1. The cumulative death count data from the 24 worst affected counties where the coverage of the *7-day-ahead* maximum (absolute) error prediction intervals (MEPI) is below 0.8 (in Figure 12a).

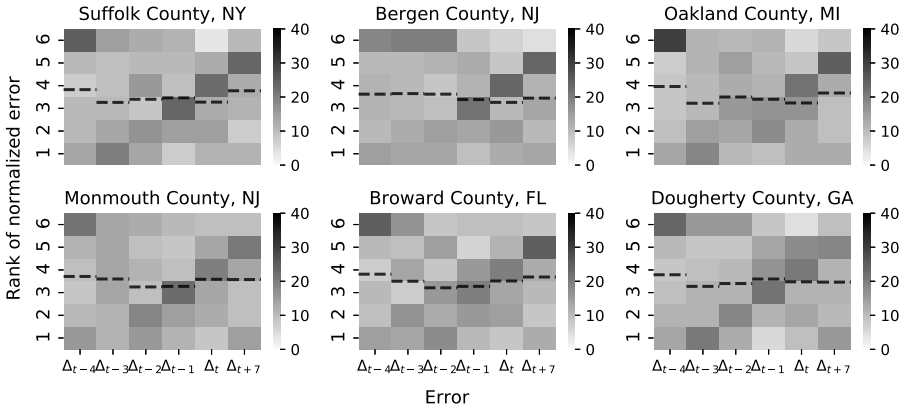
a percentile-based cut-off). Furthermore, in order to compute a 95-percentile value, we need to consider errors for at least the past 20 days. Exchangeability is not likely to hold for such a long horizon.

Normalized vs. unnormalized errors: We now provide some numerical evidence to support our choice of normalized errors to define the MEPI. Figure B3(a) shows the rank distribution of normalized errors of our 7-day-ahead CLEP predictions for the six worst-affected counties over an earlier period (March 26–April 25), and Figure B3(b) shows the (unnormalized) ℓ_1 errors $|\widehat{y}_t - y_t|$ over the same period. We find that, in Figure B3(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 B3(c), we show the rank distribution of normalized errors over a longer window of 10 days. We find 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 find that the average rank substantially deviates from this expected value for many days in this longer window for all displayed counties.

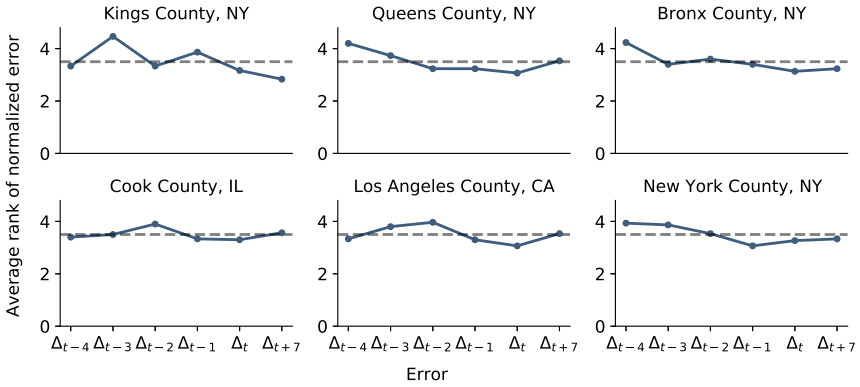
Overall, we believe that putting together the observations from Figures 5 and B3 yield reasonable justification for the two choices we made to define MEPI in (4.2a), namely, the 5-day window (versus the entire past) and the choice of normalized errors (versus the unnormalized absolute errors).

(a) Six worst-affected counties

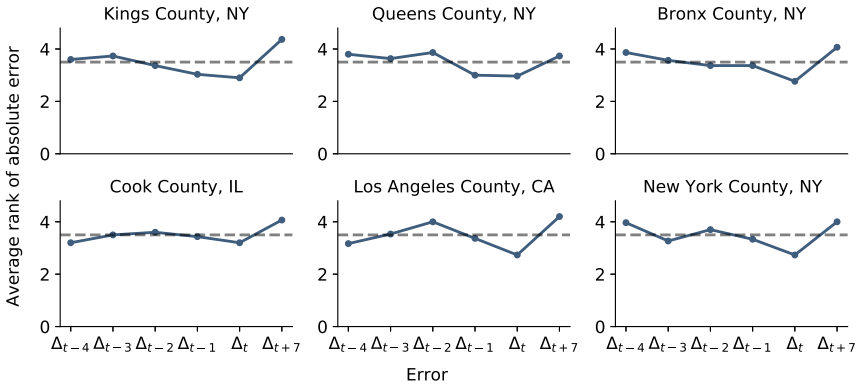


(b) Six randomly selected counties

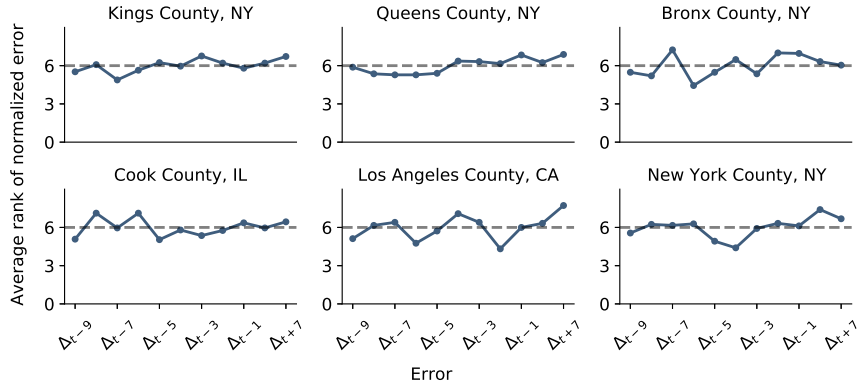
Figure B2. Exploratory data analysis (EDA) plot (heatmap version of Figure 5) for investigating exchangeability of normalized errors of *7-day-ahead* CLEP predictions with its last 5 errors made at time t, over the period t =March 26, ..., Jun 13 (80 days). These heatmaps are obtained as follows. First, for each day t, we rank the errors $\{\Delta_{t+7}, \Delta_t, \Delta_{t-1}, \ldots, \Delta_{t-4}\}$ of our CLEP (with the expanded shared and linear predictors) in increasing order so that the largest error has a rank of 6. Then, for each of the six errors $\{\Delta_{t+7}, \Delta_t, \Delta_{t-1}, \ldots, \Delta_{t-4}\}$, we count the number of days when it is ranked 1, 2, 3, 4, 5 and 6. Finally, we plot these numbers of days in the above heatmaps for (a) the six worst affected counties, and (b) six random counties. In addition, we plot the average rank of each error in dashed black lines. 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. Note that the blue lines in earlier Figure 5 are the same as the solid dotted lines in this figure.



(a) Average rank of normalized errors over 5 days

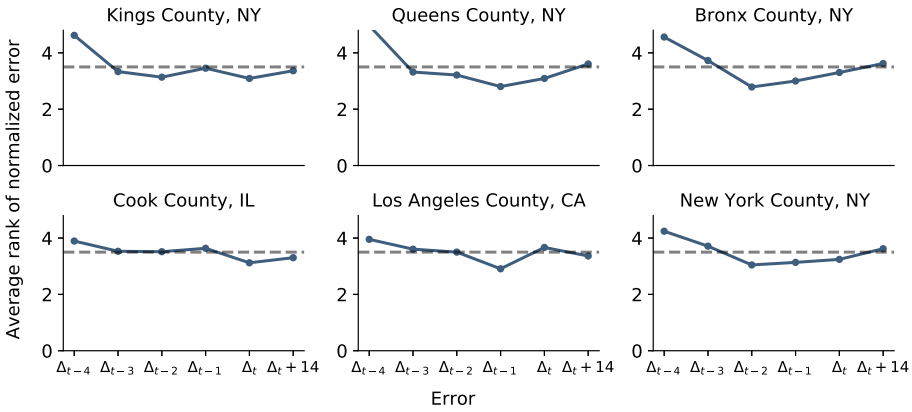


(b) Average rank of absolute unnormalized errors over 5 days

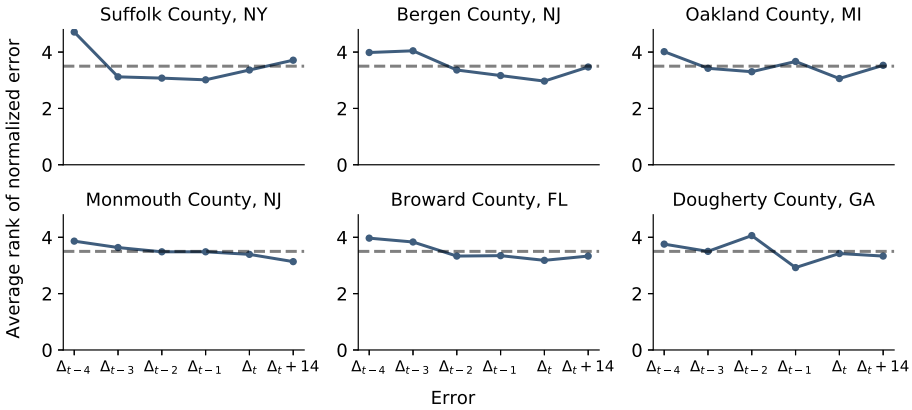


(c) Average rank of normalized errors over a longer window of 10 days

Figure B3. Exploratory data analysis (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 B4. Exploratory data analysis (EDA) plot for investigating exchangeability of normalized errors of *14-day-ahead* combined linear and exponential predictors (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).