Curating a COVID-19 data repository and forecasting county-level death counts in the United States

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

[1]Department of Statistics, University of California, Berkeley
[2]Department of EECS, University of California, Berkeley
[3]Department of Pharmaceutical Chemistry, University of California, San Francisco
[4]Chan Zuckerberg Biohub, San Francisco
[5]Center for Computational Biology, University of California, Berkeley
[6]Division of Biostatistics, University of California, Berkeley

April 15, 2020

†Authors ordered alphabetically. All authors contributed significantly to this work.
*Corresponding authors
This project was initiated on March 21, 2020, with the goal of helping aid the allocation of supplies to different hospitals in the U.S., in partnership with the non-profit Response4Life.

Abstract

In recent weeks, the novel Coronavirus causing COVID-19 has dramatically changed the shape of our global society and economy to an extent modern civilization has never experienced. In this paper, we collate a large data repository containing COVID-19 information from a range of different sources.[1]. We use this data to develop several predictors for forecasting the short-term (e.g., over the next week) trajectory of COVID-19-related recorded deaths at the county-level in the United States using data from January 22, 2020, to April 8, 2020. Specifically, we produce several different predictors and combine their 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, an exponential predictor that pools data together across counties, and a demographics-based exponential predictor. We also incorporate a linear predictor and demographic features into our ensemble. The hope is that an understanding of the expected number of deaths over the next week or so will help guide necessary county-specific decision-making and provide a realistic picture of the direction in which we are heading.

[1] All collected data and code for modeling, along with visualizations, are updated daily and available at `https://github.com/Yu-Group/covid19-severity-prediction`

1 Introduction

Our goal is to both provide access to a large data repository (that combines data collected by a range of different sources) and to provide a predictor to forecast short-term COVID-19 mortality at the county-level in the United States. Predicting the short-term impact of the virus in terms of the number of deaths (e.g., over the next week) is critical for many reasons. Not only can it help elucidate the overall impacts of the virus, but it can also help guide difficult policy decisions, such as when to impose/ease lock-downs. While many other studies focus on predicting the long-term trajectory of COVID-19, these approaches are currently difficult to verify due to a lack of data. On the other hand, predictions for immediate short-term trajectories are much easier to verify and are likely much more accurate than long-term forecasts (at least in terms of the short-term predictions for which they are designed).

In this paper, we focus on predicting confirmed deaths, rather than confirmed cases, since confirmed cases fail to accurately capture the true prevalence of the virus due to limited testing availability. Moreover, comparing different counties based on confirmed cases is difficult since some counties have performed many more tests than others: the number of positive tests does not equal the number of actual cases. We note that the confirmed death count is also likely to be an undercount of the number of true COVID-19 deaths (since it seems as though in many cases only deaths occurring in hospitals are being counted). Nonetheless, the confirmed death count is believed to be more reliable than the confirmed case count.

Unsurprisingly death rates are still climbing (as of today, April 10, 2020) across almost all counties, but they are climbing faster in some counties relative to others. On the one hand, our predictors accurately predict the number of deaths a week or so into the future for counties experiencing exponential growth in death counts. On the other hand, we found that it is harder to predict the death counts for counties that have started exhibiting either sub-exponential (slower than exponential) or super-exponential (faster than exponential) growth.

There is a large number of papers covering many dimensions of COVID-19 (see Section 5 for related work) as researchers across academia and industry refocus their efforts towards combating this universal viral threat we face. However, to the best of the authors' knowledge, there is no related work addressing *county level* predictions of COVID-19.

Making both data and the methods used in this paper accessible to others is key to ensuring the usefulness of these resources. Thus the data, code, and predictors we discuss in this paper are all updated daily and are available on GitHub[2]. Of particular note is the data that we have curated from a wide range of sources. This data includes a wide variety of COVID-19 related information in addition to the county-level death counts and demographics data that we use to develop the predictors in this paper. The results in this paper contain case and death information in the U.S.

[2] As of April 14, there have been 307 clones, 510 unique visitors, and 4,497 views of the Github repository: `https://github.com/Yu-Group/covid19-severity-prediction`

from January 22, 2020 to April 8, 2020, but the data and forecasts in the GitHub repository update daily.

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, which is updated daily with new information. Specifically, we have compiled and cleaned a large corpus of hospital-level and county-level data from a variety of public sources to aid data science efforts to combat COVID-19. We are continually updating and adding to this repository. Currently, it includes data on COVID-19-related cases, deaths, demographics, health resource availability, health risk factors, social vulnerability, and other COVID-19-related information. We provide a small snapshot of data sources from the repository below:

1. USA Facts [3]: contains cumulative COVID-19-related confirmed cases and death counts by U.S. county dating back to January 22, 2020, currently updated daily.
2. The New York Times [7]: similar to the USA Facts dataset, but it includes aggregated death counts only in New York city without county breakdowns.
3. Area Health Resources Files [2]: contains county-level data on health facilities, health professions, income estimates, mortality rates, demographics, and socioeconomic and environmental characteristics.
4. County Health Rankings & Roadmaps [1]: contains estimates of various health outcomes and behaviors for each county, including the percentage of adults who are current smokers (2017).
5. The CDC's Diagnosed Diabetes Atlas [4]: contains the estimated (age-adjusted) percentage of people who have been diagnosed with diabetes per county (from 2016).
6. The CDC's Interactive Atlas of Heart Disease and Stroke [6]: contains the estimated heart disease and stroke death rate per 100,000 (all ages, all races/ethnicities, both genders, 2014-2016).
7. Kaiser Health News [5]: contains information on the number of hospitals and the number of ICU beds in each county.

Further details, as well as the full corpus of data, are available on GitHub. Note that similar but complementary county-level data was recently aggregated and released in another study [18].

For the predictive approaches we discuss in this paper, we primarily use the county-level case and death reports provided by USA Facts. In Sec 3 and Sec 4, we use the number of deaths and cases from January 22, 2020 to April 8, 2020, along with some county-level demographics and health data.

Table 1: Overview of the 5 predictors used here. The best model is a combination of the expanded shared predictor and the linear predictor (see Sec 3.6).

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

3 Predictors for forecasting short-term death counts

Our approach involves fitting several different statistical methods. Since each method captures slightly different data trends, we also evaluate weighted combinations of these models. The five predictors we consider in this paper are:

1. **A separate-county exponential predictor (the "separate" predictors)**: a series of separate predictors built for each county using only data from that county, used to predict deaths in that county.

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

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

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

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

After fitting 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 [26]. Higher weight is given to those predictors which have more accurate predictions, especially on recent time points. In practice, we find that the CLEP that combines only the *expanded shared predictor* and the *linear predictor* has the best predictive performance.

For the rest of this section, we expand upon the individual predictors and the weighting procedure for the CLEP ensembles.

3.1 The separate-county exponential predictors (the "separate" predictors)

Our initial approach aims to capture the reported exponential growth of COVID-19 cases and deaths [21]. Hence, we approximate the best fit 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:

$$\mathrm{E}(\text{deaths}_t \mid t) = e^{\beta_0+\beta_1 t}, \quad t = 1, \ldots, 5, \tag{1}$$

where t denotes the day, and we fit a separate predictor for each county. The coefficients β_0 and β_1 are fit for each county using maximum likelihood estimation under a Poisson generalized linear model (GLM) with t as the independent variable and deaths_t as the observed variable. If the first death in a county occurred less than 5 days prior to fitting the predictor, only the days from the first death were used for the fit. If there is only one day's worth of data, we simply predict the most recent value for future values. We also fitted exponential predictors to the full time-series of available data for each county, but due to the rapidly shifting trends, these performed worse than our 5-day predictors. Besides, we found that predictors fit using 6 days of data yielded similar results to predictors fit using 5 days of data, and using 4 days of data performed slightly worse, where we cross-validated on previous days' data.

To handle possible over dispersion of data (when the variance is larger than the mean), we also estimated β_0, β_1 by fitting a negative binomial regression model (in place of Poisson GLM) with inverse-scale parameter taking values in $\{0.05, 0.15, 1\}$. However, when evaluating our predictions on earlier data (before April 1), this yields a larger mean absolute error than the Poisson GLM for counties with more than 10 deaths.

3.2 The shared-county exponential predictor (the "shared" predictor)

To incorporate additional data into our predictions (i.e., extending beyond using 5 data points from each county to fit separate predictors), we fit an predictor that combines data across the counties. In particular, we produce a single "shared" predictor that pools information from counties across the nation to predict future deaths in the individual counties.

The format of the shared predictor is slightly different from the separate county predictors. First, instead of only including the most recent 5 days of data from each county, we include all days after the third death in each county. Thus the data from many of the counties extend substantially further back than 5 days. Second, we pool all of the county-specific data points together and fit a single predictor. By using data that extends much further back, the early-stage data from counties that are now much further along could inform the predictions for current early-stage counties. Third, instead of basing the exponential predictor prediction on time t, we base the prediction on the (logarithm of the) previous day's death count. This log-transformation makes the counties comparable since the outbreaks began at different time points in each county. The shared predictor is given as follows:

$$\mathrm{E}(\text{deaths}_t \mid t) = e^{\beta_0 + \beta_1 \log(\text{deaths}_{t-1}+1)}, \tag{2}$$

where the coefficients β_0 and β_1 are fitted by maximizing the log-likelihood corresponding to Poisson GLM (like that in the separate county model (1)).

3.3 The expanded shared 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 confirmed *cases* in the county as this may give an additional indication to the severity of an outbreak, as well as the number of confirmed deaths and cases in *neighboring* counties. Let cases_t, neigh_deaths_t, neigh_cases_t respectively denote the number of cases in the county at time t, the total number of deaths across all neighboring counties at time t, and the total number of cases across all neighboring counties at time t. Then our (expanded) predictor to predict the number of confirmed deaths k days into the future is given by

$$\mathrm{E}[\text{deaths}_t | t] = e^{\beta_0 + \beta_1 \log(\text{deaths}_{t-1}+1) + \beta_2 \log(\text{cases}_{t-k}+1) + \beta_3 \log(\text{neigh\_deaths}_{t-k}+1) + \beta_4 \log(\text{neigh\_cases}_{t-k}+1)}, \tag{3}$$

where the coefficients $\{\beta_i\}_{i=0}^{4}$ are shared across all counties and are fitted using the Poisson GLM. Note that while fitting the model, e.g., at time t, while we use the death count of the county till time $t-1$, we use the new features (cases in the current county, cases in neighboring counties, and deaths in neighboring counties) only up to time $t-k$. For predicting the death count for a given county k days into the future (say $t+k$), we iteratively use the day-by-day sequential predictions for the death counts for that county, and use the information for the other features only till time t[3]. It may be possible to jointly predict the new features along with the number of deaths, but we leave this to future work.

For this predictor, we found it beneficial to implement feature scaling and regularization. We

[3] More precisely, first we estimate $\widehat{\text{deaths}}_{t+1}$ using $(\text{deaths}_t, \text{cases}_{t-k+1}, \text{neigh\_deaths}_{t-k+1}, \text{neigh\_cases}_{t-k})$. Then, for $j = 1, 2, \ldots, k-1$ we recursively plug-in $(\widehat{\text{deaths}}_{t+j}, \text{cases}_{t-k+j+1}, \text{neigh\_deaths}_{t-k+j+1}, \text{neigh\_cases}_{t-k+j+1})$ in equation (3) to estimate $\widehat{\text{deaths}}_{t+j+1}$, and finally obtain an estimate $\widehat{\text{deaths}}_{t+k}$ for k-days ahead.

scaled all features to have mean 0 and variance 1 and applied elastic net with an equal penalty on the ℓ_1 and ℓ_2 regularization terms. The regularization penalty of 0.01 was chosen through cross-validation on previous days' data.

3.4 The demographics shared predictor (the "demographics shared" predictor)

The demographics shared county exponential predictor (the "demographics shared" predictor) is again very similar to the shared predictor. However, it includes several static county demographic and healthcare-related features to address the fact that some counties will be affected more severely than others due to several factors. The severity in the counties can depend on (a) their population makeup, e.g., older populations are likely to experience a higher death rate than younger populations, (b) their hospital preparedness, e.g., if a county has very few ICU beds relative to their population, they might experience a higher death rate since the number of ICU beds as this is correlated strongly (0.96) with the number of ventilators [25]), and (c) their population health, e.g., age, smoking history, diabetes, cardiovascular disease, and respiratory diseases are all considered to be likely risk factors for acute COVID-19 infection [15, 24, 16, 14, 28]).

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

$$\mathrm{E}[\text{deaths}_t | t, c] = e^{\beta_1 \log(\text{deaths}_{t-1}+1)+\beta_0+\beta_{d_1} d_1^c+\cdots+\beta_{d_m} d_m^c}, \tag{4}$$

where the coefficients $\{\beta_0, \beta_1, \beta_{d_1}, \dots, \beta_{d_m}\}$ are fitted by maximizing the log-likelihood of the corresponding Poisson generalized linear model. The features we choose fall into three categories:

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

3.5 The separate county linear predictor (the "linear separate" predictor)

We also fit a linear version of the separate county predictors where we use linear regression to generate a linear prediction (rather than an exponential prediction) fit to the most recent 4 days of data in a county. We use this predictor because some counties have started exhibiting sub-exponential growth. For these counties, the exponential predictors introduced in the previous section may not be a good fit to the data. The linear separate predictor is given by

$$\mathrm{E}[\text{deaths}_t | t] = \beta_0 + \beta_1 t, \tag{5}$$

where we fit the coefficients β_0 and β_1 using ordinary least squares. The separate county linear predictor (5) is not a good fit for the counties that still exhibit exponential growth. In the following section, we introduce the Combined Linear and Exponential Predictor (CLEP), which incorporates the abilities of our exponential predictors (to deal with exponential trends) and linear predictor (to deal with sub-exponential trends). In practice, we found that combining the *expanded shared predictor* and the *linear predictor* has the best predictive performance.

3.6 The combined predictors: CLEP

Finally, we consider various combinations of the five predictors we have introduced using an ensemble approach similar to that described in [26]. The Combined Linear and Exponential Predictors (CLEPs) are developed as follows.

Let us first consider the procedure for generating a combined predictor for any *two* of our predictors. Let $\widehat{y}^1_{t+k}$ and $\widehat{y}^2_{t+k}$ be the predictions of (cumulative) deaths by day $t+k$ made on day t by the two predictors that we can index arbitrarily by predictor 1 and 2. Note that we only have access to complete confirmed cases and recorded deaths data up to day $t-1$ on day t, because recorded deaths and confirmed cases are not fully updated until the end of the day. The prediction of the combined estimates of deaths by day $t+k$ can be written as

$$\widehat{y}^{\text{combined}}_{t+k} = w^1_t \widehat{y}^1_{t+k} + w^2_t \widehat{y}^2_{t+k}, \tag{6}$$

where $w^1_t \geq 0$ and $w^2_t \geq 0$ represent the weights of the first and second predictors respectively, and $w^1_t + w^2_t = 1$. We select weights for the two predictors based on their past predictive performance, using an exponential decay term (a function of t). As a result, more recent predictive performance has more influence on the weight term than less recent performance. Let $\widehat{y}^m_i$ (where $m = 1, 2$) denote the predicted number of deaths from predictor m for day i, y_i denote the actual deaths for day i, and $\ell(\widehat{y}^m_i, y_i)$ denote a loss function (used for measuring predictive performance). Then following [26], the exponential weighting term w^{m}_t for predictor m applied on day t is given by

$$w^m_t \propto \exp\left(-c(1-\mu)\sum_{i=t_0}^{t-1} \mu^{t-i} \ell(\widehat{y}^m_i, y_i)\right), \tag{7}$$

where $\mu \in (0, 1)$ and $c > 0$ are tuning parameters, t_0 represents some past time point, and t represents the day on which the prediction is calculated. Since $\mu < 1$, the μ^{t-i} term represents the greater influence given to more recent predictive performance. Note that the loss terms $\ell(\widehat{y}^m_i, y_i)$ used in the weights are calculated based on the three-day predictions from seven predictors built 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). The influence of each predictor's loss decreases as we go back in time due to the exponential decaying term μ^{t-i}

in the weight expression (7). We chose the past week's 3-day performance in the weights since it yielded good performance for our ensemble predictor for predicting death counts several days in the future.

In [26], the authors choose $\ell(\widehat{y}_i^m, y_i) = |\widehat{y}_i^m - y_i|$ as their loss function, since their errors roughly had a Laplacian distribution. In our case, we found that using this loss function led to vanishing weights due to the heavy-tailed nature of our error distribution. To help address this, we apply a logarithm to the predictions and the true values, and define $\ell(\widehat{y}_i^m, y_i) = |\log(1+\widehat{y}_i^m) - \log(1+y_i)|$, where we add a one inside the logarithm to handle potential zero values. We found that this transformation improved performance in practice.

To generate our predictions, we use the default value of c in [26] which is 1. However, we change the value of μ from the default of 0.9 to 0.5 for two reasons: (i) we found $\mu = 0.5$ yielded better empirical performance, and (ii) it ensured that performance more than a week ago had little influence over the predictor. We chose $t_0 = t - 7$ (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. (Moreoever,the information from more than a week effectively has a vanishing effect due to our choice of μ.) Thus, in practice, we used weights for predictor m of the form:

$$w_t^m \propto \exp\left(-0.5 \sum_{i=t-7}^{t-1} (0.5)^{t-i} \left|\log(1+\widehat{y}_i^m) - \log(1+y_i)\right|\right), \tag{8}$$

where $\widehat{y}_i^m$ is the 3-day ahead prediction from the predictor m trained on data till time $i-3$.

To extend our ensemble approach to developing and Combined Linear and Exponential Predictors (CLEPs) consisting of more than two predictors, the weights are calculated in the same way, and normalized so that $\sum_{m=1}^{M} w_t^m = 1$, where M is the total number of predictors that make up the CLEP.

4 Results

In this paper, we focused on short-term (up to 10 days) predictive accuracy. Table 2 summarizes the Mean Absolute Errors (MAEs) of our predictions for cumulative recorded deaths by April 8th using (a) the un-scaled death counts ($\frac{1}{n}\sum_{i=1}^{n}|\widehat{y}_i - y_i|$), and (b) using a log-scale ($\frac{1}{n}\sum_{i=1}^{n}|\log(\widehat{y}_i + 1) - \log(y_i+1)|$). In each column, the smallest mean absolute error is displayed in bold. We compared different predictors to predict death count for k-days in future for $k \in \{3, 5, 7, 10\}$ and filtered for counties with $\geq j$ deaths for $j \in \{10, 100\}$. In all the choices, either the expanded shared predictor or the CLEP ensemble that combines the expanded shared exponential predictor and the separate county linear predictors perform the best.

In Figure 1, we explore the results of our best performing predictor: the CLEP ensemble predictor with the expanded shared predictor and the linear predictors, focusing on 7-day predictions.

Table 2: Mean Absolute Errors (Top: raw scale; Bottom: log scale) of our predictions for deaths by April 8, 2020. "t-day prediction" ($t = 3, 5, 7, 10$) indicates predictions made t days ahead of April 8. "deaths$\geq j$" ($j = 10, 100$) indicates average error across all counties with nonzero cases and cumulative deaths greater than j on April 8. "Average deaths" are average cumulative deaths across all counties in the corresponding category and are given here for reference. The smallest error in each column is displayed in bold.

(a) Raw scale MAE

	3-day prediction		5-day prediction		7-day prediction		10-day prediction	
	deaths>= 10	>= 100	>= 10	>= 100	>= 10	>= 100	>= 10	>= 100
separate	22.71	95.74	56.64	197.46	87.94	310.57	297.85	1113.44
shared	17.37	84.87	23.01	100.57	40.39	197.24	46.15	201.38
demographics	25.25	133.22	29.05	146.12	56.70	307.74	74.24	379.76
expanded shared	13.02	55.35	15.30	**47.77**	24.02	86.46	31.18	**111.39**
linear	11.10	39.50	21.75	98.74	30.85	142.07	47.64	215.37
CLEP (all 5 predictors)	14.56	66.46	22.52	86.59	33.58	144.77	60.97	249.12
CLEP (expanded +linear)	**8.59**	**29.51**	**14.40**	52.08	**16.40**	**63.14**	**29.77**	123.70
Average deaths (for reference)	75.53	373.29	75.53	373.29	75.53	373.29	75.53	373.29

(b) Log scale MAE

	3-day prediction		5-day prediction		7-day prediction		10-day prediction	
	deaths>= 10	>= 100	>= 10	>= 100	>= 10	>= 100	>= 10	>= 100
separate	0.30	0.20	0.62	0.54	0.82	0.60	1.43	1.04
shared	**0.17**	0.16	0.28	0.23	0.36	0.36	0.54	0.51
demographics	0.20	0.23	0.26	0.29	0.38	0.49	0.61	0.78
expanded shared	**0.17**	0.13	0.25	**0.12**	0.36	**0.19**	**0.52**	**0.46**
linear	0.24	0.12	0.33	0.26	0.53	0.47	1.07	1.05
CLEP (all 5 predictors)	0.18	0.14	0.31	0.26	0.37	0.34	0.64	0.63
CLEP (expanded +linear)	**0.17**	**0.09**	**0.24**	0.15	**0.27**	0.20	**0.52**	0.65
Average log of deaths (for reference)	4.34	5.93	4.34	5.93	4.34	5.93	4.34	5.93

Since there are over 3,000 counties in the United States, we first focus on results for six of the currently worst affected counties, including Queens County, NY; King County, WA; Kings County, NY; New York County, NY; Orleans County, LA; and Wayne County, MI. Later, we also consider results for six randomly selected counties.

Specifically, Figure 1 displays the number of cumulative deaths (by days since 10 deaths) up to April 8, 2020, for these six worst affected counties. The predicted number of deaths based on the best-performing ensemble predictor (the expanded shared predictor plus the linear predictor) for the 7 days of April 2-8 (inclusive) are displayed as a dashed blue line. These predictions were made on April 1. Overall, the CLEP appears to fit the data closely for three of these counties (Orleans County LA, Queens County NY, and Wayne County MI). However, our predictor (i) greatly under-predicts the number of deaths in Kings County NY (whose growth in the past few days also grew more than would be expected under exponential growth), (ii) slightly under-predicts the number of deaths in New York County NY (whose growth in the past few days also grew more than expected), and (iii) slightly over-predicts the number of deaths in King County WA (which now appears to be exhibiting sub-exponential growth).

Figure 2 shows these line plots for 6 randomly selected counties, and again, we see that our predictors perform very well for three of the counties (Broward County FL, Dougherty County GA, and Monmouth County NJ), but performs slightly less well for Bergen County NJ, Suffolk County NY and Oakland County MI. However, in these counties with poorer performance, the trends in the data between April 4 and April 8 are surprisingly irregular. For instance, Suffolk County NY reported no new deaths on April 5 but reported 60 new deaths on April 6. These strange behaviours in the observed counts are likely due to death reports not being released on the weekend (April 5 was a Sunday).

Figure 3 compares the actual number of deaths by April 8 against the predicted number of deaths by April 8 for all counties based on the ensemble predictor based on the expanded shared and the linear predictors. Note that the predictions were generated on April 1, using data from seven days prior to the prediction date April 8. We observe that overall the predictions are very close to the true values.

5 Related Work

Several recent works have tried to predict the number of cases and deaths related to COVID-19. But to our best knowledge, none of these works have predicted deaths and cases at the county level. In addition, directly comparing other researcher's forecasting results to our own is difficult for several other reasons: (1) the predictors mostly make strong assumptions and involve little data-fitting, (2) we do not have access to a direct implementation of their predictors (or results), and (3) their predictors focus on substantially longer time horizon when compared to ours. Keeping

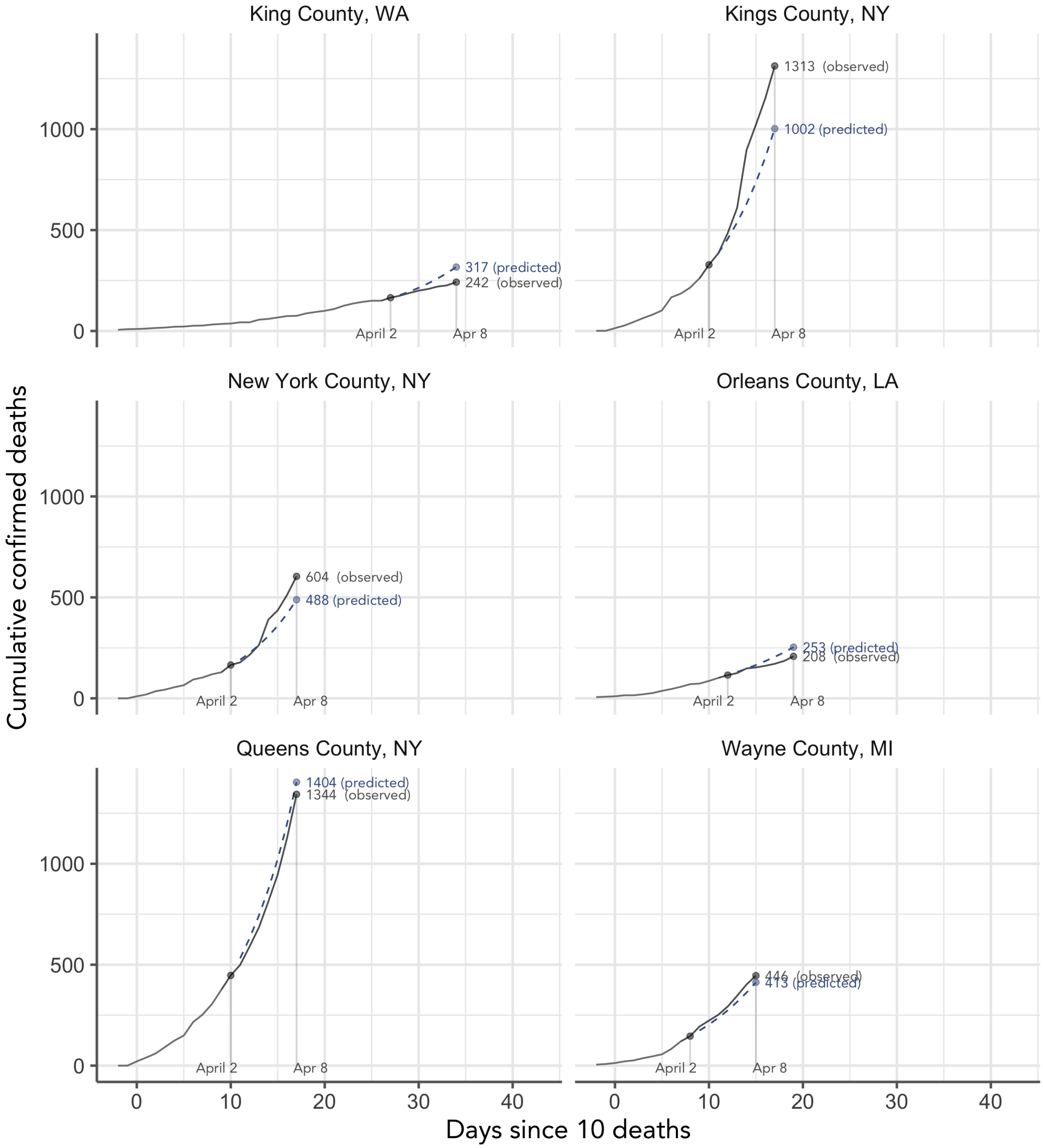

Figure 1: A grid of line charts displaying the cumulative number of confirmed COVID-19 deaths by day measured since 10 deaths for six of the worst-affected counties. The observed data is shown in grey. The seven-day predicted deaths from the ensemble predictor consisting of the expanded shared predictor and the linear predictor based on the data up to April 1 is shown in the dashed blue. Overlapping text annotations are hidden.

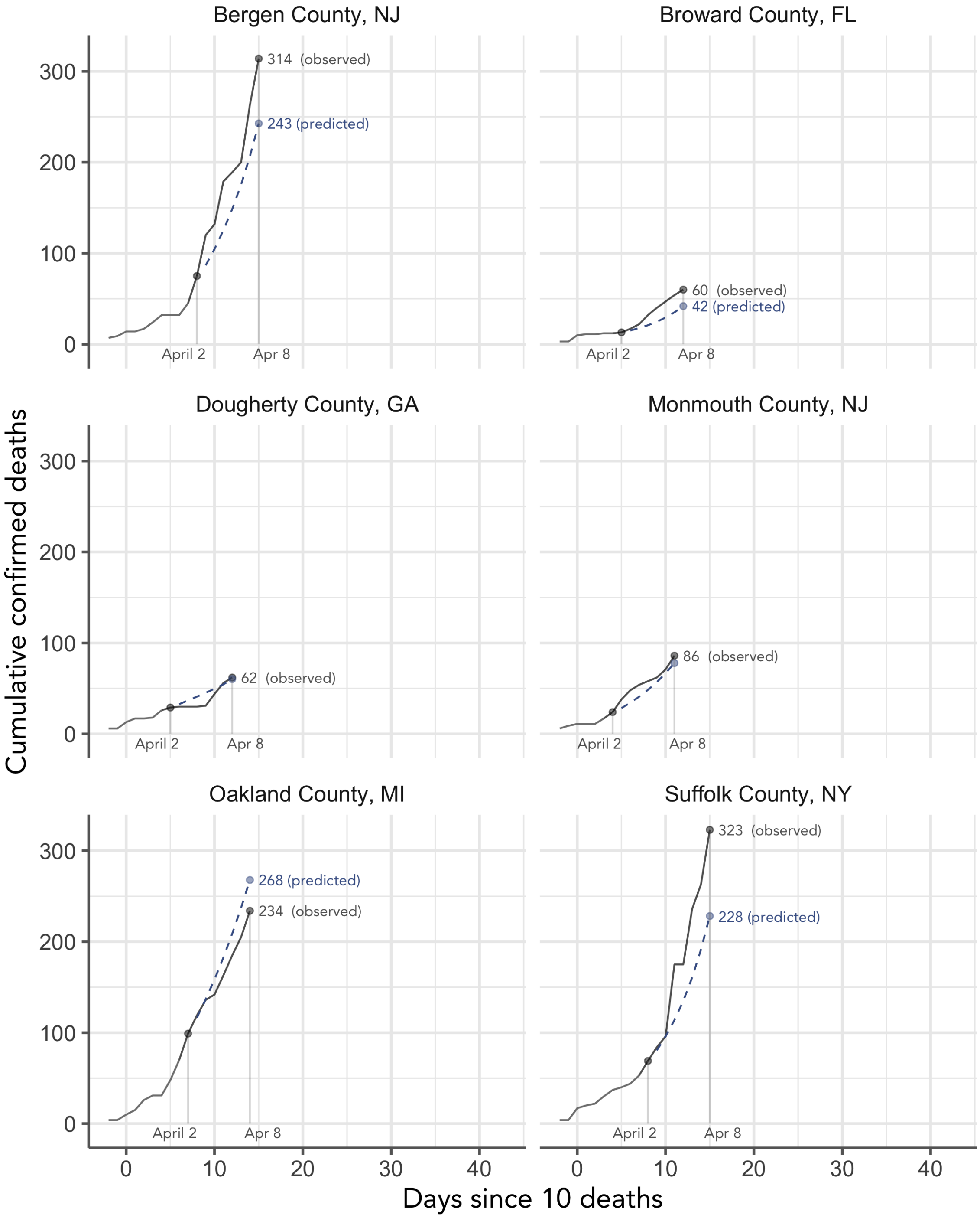

Figure 2: A grid of line charts displaying the cumulative number of confirmed COVID-19 deaths by day measured since 10 deaths for six randomly selected counties. The observed data is shown in grey. The seven-day predicted deaths from the ensemble predictor consisting of the expanded shared predictor and the linear predictor based on the data up to April 1 is shown in the dashed blue. Overlapping text annotations are hidden.

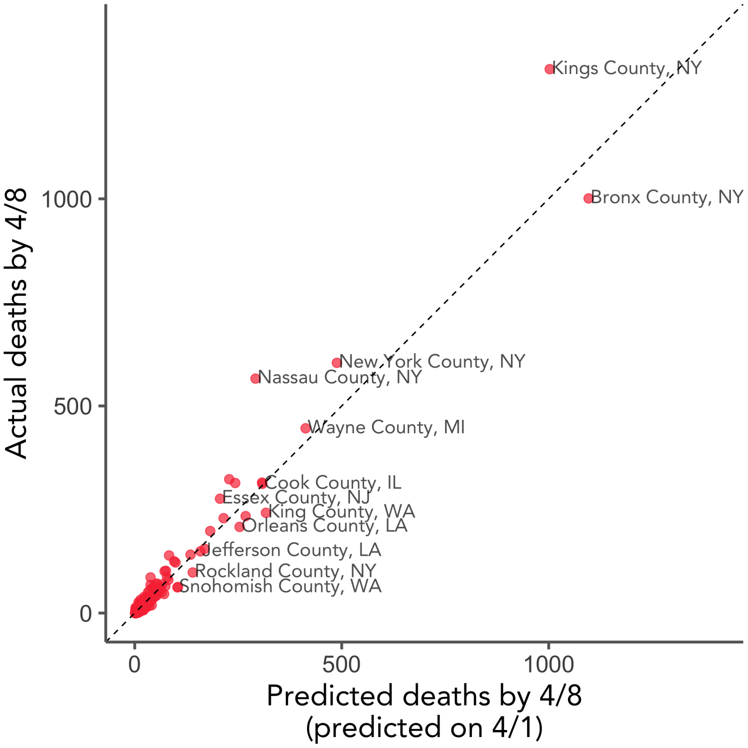

Figure 3: A scatterplot displaying the actual number of deaths on April 8 against the predicted number on deaths by April 8 (where the prediction was generated seven days earlier on April 1).

these points in mind, we now provide a brief summary of recent work on predictive modeling for COVID-19.

Two recent works [20] and [13] have modeled the death counts at the state level in the US. While the Institute for Health Metrics and Evaluation (IHME) [20] model[4] is based on Farr's Law with feedback from Wuhan data, the Imperial College predictor [13] uses an individual-based simulation predictor with parameters chosen based on prior knowledge. On the topic of Farr's Law, we note that a 1990 paper [10] used Farr's Law to predict that the total cases from the AIDS epidemic would diminish by the mid-1990s and the total number of cases would be around 200,000 in the entire lifetime of the AIDS epidemic. It is now estimated that 32 million people have died from the disease so far. While the AIDS pandemic is very different to the COVID-19 pandemic, it is still worth keeping in mind.

Another approach uses exponential smoothing from time series predictors to estimate day-level COVID-19 cases [11]. In addition, several works use compartment epidemiological predictors such as SIR, SEIR and SIRD [12, 23, 9] to provide simulations at the national level. Other works [22, 17] simulate the effect of social distancing policies either in future for USA, or in a retrospective manner for China. Finally, several papers estimate epidemiological parameters retrospectively based on data from China [27, 19].

6 Discussion

We recognize the value that prediction intervals hold and the utility that giving a range of possible predictions provides. However, there are some key challenges to creating good prediction intervals for predicting COVID-19 recorded deaths which implicitly require an objective way of evaluating the quality of such an interval. The curves in Figure 1 and Figure 2 suggest a large source of prediction inaccuracy was due to regime changes, in which the number of deaths would exhibit sharp transitions in behavior (either a sharp uptick in number of record deaths or a change from exponential growth to something closer to linear). Due to the highly dynamic nature of COVID-19, it can be challenging to create good intervals that can account for these sharp regime changes. Furthermore, the progression of theses curves can be subject to external influences such as major policy changes (such as states revoking shelter-in-place orders) or human behavior changes or scientific developments (such as technology allowing multiple people to use a single ventilator). Even without these dynamic considerations, constructing good prediction intervals requires careful thought since confidence interval construction methods are typically built upon assumptions that the underlying probabilistic models hold (approximately or to a desirable extent). We note that validity of bootstrap also hinges upon probabilistic assumptions. These assumptions are very

[4] The IHME model has been updated regularly since its first release on March 26, 2020. Here we briefly described their version released on March 26, 2020. In their model, the authors say that they are no longer using Farr' Law but the precise modeling details were not available until the submission of this manuscript.

difficult to check in a dynamic situation. Furthermore, there has not yet been a consensus way to construct these intervals. As a matter of fact, a widely cited COVID-19 prediction model (at the state level) [20] has been experimenting with different approaches resulting very different prediction interval constructions [8] in different versions of their model. Despite these challenges, we hope to create prediction intervals for our model in future work.

Conclusion In this paper, we introduce a repository of datasets containing COVID-19-related information from a variety of public sources, and we used this data to fit a series of separate (exponential and linear) predictors as well as ensemble predictors each designed to predict the short-term (i.e. on the scale of 1 week) number of county-level deaths. We found that for the majority of counties that appear to exhibit exponential growth, our predictors accurately predict the number of deaths over the next week. Our predictors are less accurate for counties that no longer appear to exhibit exponential growth, either because their reported death counts are growing faster or slower than an exponential rate. We hope that our data repository and our predictors for predicting the short-term trends of the COVID-19 deaths at the county level will help to provide useful information for those who need to make difficult decisions at this critical time in the evolving pandemic.

7 Acknowledgements

The authors would like to thank many people for help with this effort. This includes the response4Life team and volunteers; Max Shen's IEOR group at Berkeley: Shen Group team (IEOR): Junyu Cao, Shunan Jiang, Pelagie Elimbi Moudio; Aaron Kornblith and David Jaffe for advice from a medical perspective, and helpful input from many others including Ying Lu, Tina Eliassi-Rad, Jas Sekhon, Philip Stark, Jacob Steinhardt, Nick Jewell, Valerie Isham, SriSatish Ambati, Rob Crockett, Marty Elisco, Valerie Karplus, Marynia Kolak, Andreas Lange, Qinyun Lin, Suzanne Tamang, Brian Yandell and Tarek Zohdi. Special thanks to Sam Scarpino for sharing his data with us.

References

[1] Adult smoking data. *County Health Rankings & Roadmaps.* Accessed on 04-02-2020 at `https://www.countyhealthrankings.org/explore-health-rankings/measures-data-sources/county-health-rankings-model/health-factors/health-behaviors/tobacco-use/adult-smoking`.

[2] Area health resources files. *National Center for Health Workforce Analysis, Bureau of Health Workforce, Health Resources and Services Administration.* Accessed on 04-02-2020 at `https://data.hrsa.gov/data/download`.

[3] COVID-19 deaths dataset. *USA Facts*. Accessed on 03-31-2020 at https://www.reuters.com/article/us-health-coronavirus-who/covid-19-spread-map.

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

[5] Icu beds by county. *Kaiser Health News*. Accessed on 04-02-2020 at https://khn.org/news/as-coronavirus-spreads-widely-millions-of-older-americans-live-in-counties-with-no-icu-bed

[6] Interactive atlas of heart disease and stroke. *Centers for Disease Control and Prevention, Division for Heart Disease and Stroke Prevention*. Accessed on 04-02-2020 at https://www.cdc.gov/dhdsp/maps/atlas/index.htm.

[7] The New York Times Data. https://www.nytimes.com/interactive/2020/us/coronavirus-us-cases.html. Accessed on 04-01-2020.

[8] COVID-19: What's New for April 5, 2020, 2020. http://www.healthdata.org/sites/default/files/files/Projects/COVID/Estimation_update_040520_3.pdf, Last accessed on 2020-04-13.

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

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

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

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

[13] N. Ferguson, D. Laydon, G. Nedjati-Gilani, N. Imai, K. Ainslie, M. Baguelin, S. Bhatia, A. Boonyasiri, Z. Cucunubá, G. Cuomo-Dannenburg, et al. Impact of non-pharmaceutical interventions (NPIs) to reduce COVID19 mortality and healthcare demand, 2020. Accessed on 04-02-2020 at https://spiral.imperial.ac.uk/bitstream/10044/1/77482/5/Imperial%20College%20COVID19%20NPI%20modelling%2016-03-2020.pdf.

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

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

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

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

[18] B. D. Killeen, J. Y. Wu, K. Shah, A. Zapaishchykova, P. Nikutta, A. Tamhane, S. Chakraborty, J. Wei, T. Gao, M. Thies, et al. A county-level dataset for informing the united states' response to covid-19. *arXiv preprint arXiv:2004.00756*, 2020.

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

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

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

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

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

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

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

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

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

[28] F. Zhou, T. Yu, R. Du, G. Fan, Y. Liu, Z. Liu, J. Xiang, Y. Wang, B. Song, X. Gu, et al. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: A retrospective cohort study. *The Lancet*, 2020.