A Novel Machine Learning Approach to Identify COVID-19 Deaths Among Excess Deaths Reported to Non-COVID-19 Causes

Mathew V. Kiang; Andrew C. Stokes

ABSTRACT

The total number of deaths caused by the SARS-CoV-2 virus in the United States has been heavily debated since the start of the COVID-19 pandemic. Researchers have previously developed excess mortality models to estimate the number of deaths that would have occurred in the absence of the pandemic and the number of COVID-19 deaths that were not reported to COVID-19. However, estimates of excess deaths represent an upper-bound of total COVID-19 deaths as some of these deaths were likely related to health care interruptions and others to the pandemic's social and economic effects. We use a machine learning approach to leverage information from death certificate data, county characteristics related to population, health systems, and the death investigation system, and county-month trends in excess deaths and reported COVID-19 deaths to produce refined estimates of the total number of COVID-19 deaths throughout the United States from 2020 through 2022.

PROPOSAL

Background:

Since the beginning of the COVID-19 pandemic, it has been widely recognized that official COVID-19 death counts which include death certificates on which COVID-19 is listed as a cause or contributor to deaths are likely to be an incomplete account of the total mortality impacts of the pandemic in the United States.[1–5] There remains however a significant scholarly and public discussion about several related points. First, were "uncounted deaths" (deaths related to the pandemic that did not list COVID-19 on the certificate) mostly a feature of the early pandemic when health systems were unprepared to recognize COVID-19 deaths?[6,7] Second, have over-counts of COVID-19 deaths occurred in any substantial manner during the pandemic?[7,8] Third, do "uncounted deaths" primarily reflect deaths caused by the SARS-CoV-2 virus or deaths caused by pandemic-related health care interruptions, the pandemic's social and economic effects, and/or public health policies related to the pandemic such as lockdowns and physical distancing?[9–14] Lastly and relatedly, what was the total number of deaths caused by the SARS-CoV-2 virus, adding both deaths recognized in official counts and "uncounted deaths"?[15]

Excess mortality models are one approach that has been utilized throughout the public health literature to better understand the accuracy of official COVID-19 mortality statistics.[16] Excess mortality refers to the difference between the observed number of deaths that occurred during a given period and the number of deaths that would be expected based on earlier mortality trends.[17] Recent estimates have found that an estimated 1,179,024 excess deaths occurred during the first 2 years of the pandemic, with 634,830 excess deaths occurring in the first year and 544,194 in the second year.[7] Most prior studies have found that the number of excess deaths in the United States has substantially exceeded the number of COVID-19 deaths.[1,2,4,5,9] Prior comparisons of excess deaths to COVID-19 deaths have suggested that there have been approximately 10% to 35% more excess deaths than COVID-19 deaths, but estimates have varied across modeling specifications and time periods studied.[1,2,4,5,7,15,18]

Another important complexity that has contributed to differences between models (especially across analyses that use data from different health systems, states, or regions) is how excess deaths, COVID-19 deaths, and "uncounted deaths" differ in important ways demographically, geographically, and temporally.[5,7,10,12,19–29] Structural racism has caused substantial racial and ethnic inequities in COVID-19 and excess mortality during the pandemic, which have varied regionally and across the rural-urban continuum and temporally as the pandemic has progressed.[19,20,30–33] Relatedly, prior research has shown that the proportion of excess deaths not reported to COVID-19 deaths has varied demographically. For example, predominately Black communities have experienced more "uncounted deaths", which suggests that the true racial and ethnic inequities in COVID-19 mortality may be even larger than revealed in official death counts.[4,34] Geographic analyses have also revealed that excess deaths were less likely to be reported as COVID-19 deaths in the Mountain division, in the South, and in nonmetropolitan counties.[15] Temporally, there was a very significant spike in "uncounted deaths" early in the pandemic when death investigators were unfamiliar with the manifestations of COVID-19.[15,35,36] After this initial spike, it is possible that reporting may have gradually improved as guidelines were developed and refinements were made to the national vital statistics system.[15,37–39]

A factor that could serve as one mechanism explaining geographic and temporal differences in "uncounted deaths" is how the death investigation system operates and is funded.[40] In particular, cause-of-death assignment often varies significantly between in-hospital and out-of-hospital settings.[10,41–43] While inaccuracies do exist with in-hospital assignment practices, physicians have significantly more information available to them about decedents and often have far more training than other death investigators.[44–46] The United States death investigation system is a patchwork system with distinct offices across counties and states.[42] While many counties (typically more urban counties) have forensic pathologists working as medical examiners, other counties (typically more rural counties) have coroners, sheriff-coroners, and justices of the peace who are generally elected officials with limited medical training.[40,47] Another factor is that there is significant variability in the amount of funding that death investigation offices have available in each county, meaning counties had differential resources available to investigate COVID-19 deaths through interviews and post-mortem testing.[48] In line with this discussion, prior research during the pandemic has shown that most "uncounted deaths" occurred in out-of-hospital settings and were more common in areas without medical examiners.[10,41]

While estimating the precise number of COVID-19 deaths that have occurred during the pandemic given these challenges may appear as a largely technical project, accurate estimates of the burden of infectious diseases such as COVID-19 have critical importance for informing ongoing pandemic-related health policy and for future public health emergency preparedness. We offer three examples here to illustrate this point.[40,49] First, differential under-counting of COVID-19 deaths could disguise demographic groups and geographic areas that had significant COVID-19 mortality burdens that either went unnoticed or were underrecognized in official death counts.[34] This suggests that improving the accuracy of death counts may be important for ongoing investigation of the long-term impacts of the pandemic on marginalized groups. Second, many social and health policies were designed using official COVID-19 death data. One striking example is the FEMA funeral assistance program, which is an important social program to reduce the financial impact of pandemic-related losses on families.[6] This program, however, requires individuals to submit a death certificate that lists

COVID-19, meaning that “uncounted deaths” are not eligible and that inequities in reporting would extend to this social benefit. Third, COVID-19 deaths are just one issue related to cause-of-death assignment in the death investigation system. Prior research has shown that other causes of death such as drug overdose, Alzhemier’s disease and related dementias, and deaths in police custody are also under-counted in the United States.[46,50–52] We believe new approaches for identifying COVID-19 uncounted deaths could potentially be adapted to these other causes of death, and methods developed using retrospective data could ultimately be leveraged for improved surveillance.

Research Objectives:

In this analysis, we use a novel machine learning approach to leverage individual-level demographic information from multiple cause of death data, county-level characteristics related to demographics, population, health systems, and the death investigation system, and county-month trends in excess deaths and reported COVID-19 deaths by place of death to produce refined estimates of the total number of COVID-19 deaths throughout the United States from 2020 through 2022. In addition to highlighting the performance and utility of our approach, we will also produce updated estimates of COVID-19 mortality for counties and states and identify geographic areas and demographic groups (e.g. by age, sex, race, ethnicity, gender, and education) who had large numbers of uncounted COVID-19 deaths. By sharing a new approach for improving estimates of COVID-19 deaths to account for inaccuracies and inequities, we will generate discussion about new approaches for monitoring multiple types of deaths in the United States to account for the possibility of inconsistent cause-of-death assignment, death investigator bias, and differential under-funding of death investigation systems to mislead public health researchers and policy-makers about the burden of disease.

Data:

We used restricted-access multiple cause of death files from the National Center for Health Statistics (NCHS) for March 2020 through December 2021. These data contain all 6.35 million deaths that occurred in the US during this time period and contain information such as underlying cause of death and up to 20 contributory causes of death (coded in ICD-10) as well as decedent information such as age (in years), month and year of death, county of residence and death, educational attainment, marital status, sex, smoking status, and race and ethnicity. Importantly, the death certificates also contain information about the place of death (e.g., in hospital, at home, long-term care facility, etc.). Based on previous work, we created derived variables using contributing cause of death codes including the presence of diabetes, pneumonia, kidney failure, essential hypertension, and hyperlipidemia, ICD-10 codes.

We combined these individual-level data with both time-invariant and time-varying contextual data. Our time-invariant contextual data includes the rural-urban continuum code of the county of residence, the proportion of the population that is non-Hispanic White, the proportion of the population that is Hispanic, the proportion of the population that is non-Hispanic Black, the proportion of the population that is over 65 years of age, median household income, the proportion of the population that owns a home, county-level income inequality, the proportion of the county with diabetes, county-level obesity and smoking rates, and the proportion of the county that reports poor or fair health. For all time-invariant covariates, we used data from before 2020 to prevent data leakage. In addition, we included two time-varying covariates at the county-month level: the CDC community transmission level categorization and the percent

of the population that received a vaccine. In sensitivity analyses, we also included county of residence and each of the NCHS 113 cause of death codes as fixed effects.

Research Methods:

Analytic Approach

Because COVID-19 deaths among those under 25 years of age is uncommon, we removed them from the data for a final data set of 6.2 million deaths. We further split these data into a gold standard analytic data set and a prediction data set. The gold standard analytic data set consisted of all deaths that occurred in the hospital among those receiving in-patient care (N=1.98 million), reflecting our belief that in-hospital, in-patient deaths during this time period were more likely to be accurately classified as COVID-19 deaths. These gold standard data were used for model tuning, model selection, and building a final classification model. The remaining, out-of-hospital (or in-hospital emergency room or dead on arrival) deaths (N=4.24 million) are used during the prediction phase.

Creating a Classification Model

We divided the gold standard analytic data set into 60-20-20 train-validate-test splits. During hyperparameter tuning, models were trained on 60% of the data and validated against 20% of the data until an optimal set of hyperparameters was found. The model was then refit with the finalized hyperparameters on 80% of the data and evaluated against the 20% test set, which was only used for model evaluation (i.e., never used for model training). Using true holdout data for our test split ensures that our model performance metrics accurately reflect the expected out-of-sample performance.

We fit five types of models: logistic regression with ElasticNet regularization, logistic regression with LASSO regularization, random forests, LightGBM, and XGBoost. We used Bayesian optimization, which is a sequential tuning optimization that explores the hyperparameter space using an acquisition function to continually find a set of hyperparameters that improves upon a pre-specified metric. We used the area under the receiver operating characteristics curve (AUC ROC) as our primary model performance metric. We allowed the Bayesian optimization to continue for up to 125 iterations with two tweaks. First, tuning stopped early if there was no model performance improvement for 25 iterations. Second, if there was no model performance improvement for 7 iterations, the model would select a high variance area of the hyperparameter space to explore.

For each of the five model types, we fit our primary covariate set and three additional sensitivity sets. Our primary covariate set consisted of all information available on the death certificate and both time-invariant and time-varying contextual variables based on the decedent's county of residence. The three sensitivity covariate sets included adding county fixed effects to county for unobserved factors that may vary across areas (but not time), contributing cause of death fixed effects, and both county and contributing cause of death fixed effects.

For the regularized logistic regressions, we preprocessed the data by creating binary indicator variables for all categorical predictors, creating a third-degree polynomial of age, and performing a mean-centered, unit-standard deviation normalization on all numeric predictors.

For the remaining tree-based models, we preprocessed the data by using one-hot coding for all categorical covariates. For all models, we used mean imputation for time-invariant county-level contextual variables with missing data and downsampled observations to address the slight imbalance in COVID-19 classifications.

Model Selection

After tuning and refitting all models, we selected the best model with the highest AUC ROC value. This model was then refit to all (i.e., 100%) of the gold-standard data and was used to create our final classification model to classify COVID-19 deaths in the prediction (i.e., out of hospital) data set.

Predicting COVID-19 Deaths

To quantify potential over- or under-reporting of official COVID-19 deaths, we used the final classification model to identify COVID-19 deaths on the prediction (i.e., out of hospital) data set across a variety of individual- and county-level strata. We estimated the adjusted reporting ratio (ARR) for each stratum as

$$ARR_i = \frac{{D_i}^H + \widetilde{D_\iota}}{{D_i}^T}$$

where i is our strata of interest, ${D_i}^H$ is the number of official in-hospital COVID-19 deaths, $\widetilde{D_\iota}$ is the number of predicted out-of-hospital COVID-19 deaths, and ${D_i}^T$ is the total number of official COVID-19 deaths. We estimated 95% uncertainty intervals (95% UI) by using 5,000 bootstrapped samples of the out-of-hospital predictions and reporting the 2.5th and 97.5th quantiles.

Preliminary Results:

Across all models, the XGBoost models had the highest ROC AUC of 0.902, while maintaining high accuracy, sensitivity, and specificity (Figure 1). In fact, our preferred model, the XGBoost with no county and no contributing cause data had the highest performance in 8 of the 13 measures used and was within less than 1% of the best performing model for the remaining 5 measures (Table S1) and exhibited excellent performance across all standard evaluations (Figures S1-S4).

The adjusted reporting ratio (ratio of predicted COVID-19 deaths over observed COVID-19 deaths) was 1.28 for the total population. The ratio differed across individual and community-level factors, including such factors as age, educational attainment, race and ethnicity, and urban-rural status (Figures 1-2). There was also substantial spatial temporal variation in the adjusted reporting ratio by state and over pandemic months (Figure 3).

Next Steps:

Having identified significant differences in the adjusted reporting ratio across social and demographic characteristics, states, and pandemic months, we will expand our research by further refining the spatial resolution of the analysis. We will also perform regression

modeling to examine the association of the reporting ratio with public health and health care factors as well as features of the death investigation system.

REFERENCES

1. Weinberger DM, Chen J, Cohen T, et al. Estimation of Excess Deaths Associated With the COVID-19 Pandemic in the United States, March to May 2020. *JAMA Intern Med*. 2020;180(10):1336-1344.

2. Woolf SH, Chapman DA, Sabo RT, Zimmerman EB. Excess Deaths From COVID-19 and Other Causes in the US, March 1, 2020, to January 2, 2021. *JAMA*. April 2021. https://jamanetwork.com/journals/jama/fullarticle/2778361?guestAccessKey=8445def5-44ef-401d-8e01-002d0fbbadf5&utm_source=silverchair&utm_medium=email&utm_campaign=article_alert-jama&utm_content=olf&utm_term=040221. Accessed April 2, 2021.

3. Woolf SH, Chapman DA, Sabo RT, Weinberger DM, Hill L, Taylor DDH. Excess Deaths From COVID-19 and Other Causes, March-July 2020. *JAMA*. 2020;324(15):1562-1564.

4. Stokes AC, Lundberg DJ, Elo IT, Hempstead K, Bor J, Preston SH. COVID-19 and excess mortality in the United States: A county-level analysis. *PLoS Med*. 2021;18(5):e1003571.

5. Ackley CA, Lundberg DJ, Ma L, Elo IT, Preston SH, Stokes AC. County-level estimates of excess mortality associated with COVID-19 in the United States. *SSM Popul Health*. 2022;17:101021.

6. FEMA. COVID-19 Funeral Assistance. U.S. Department of Homeland Security. https://www.fema.gov/disaster/coronavirus/economic/funeral-assistance. Published July 11, 2023. Accessed July 17, 2023.

7. Paglino E, Lundberg DJ, Zhou Z, et al. Monthly excess mortality across counties in the United States during the COVID-19 pandemic, March 2020 to February 2022. *Sci Adv*. 2023;9(25). doi:10.1126/sciadv.adf9742

8. Wen LS. We are overcounting covid deaths and hospitalizations. That's a problem. *The Washington Post*. https://www.washingtonpost.com/opinions/2023/01/13/covid-pandemic-deaths-hospitalizations-overcounting/. Published January 13, 2023. Accessed July 13, 2023.

9. Lee WE, Woo Park S, Weinberger DM, et al. Direct and indirect mortality impacts of the COVID-19 pandemic in the United States, March 1, 2020 to January 1, 2022. *Elife*. 2023;12. doi:10.7554/eLife.77562

10. Chen YH, Stokes AC, Aschmann HE, et al. Excess natural-cause deaths in California by cause and setting: March 2020 through February 2021. *PNAS Nexus*. June 2022. doi:10.1093/pnasnexus/pgac079

11. Glei DA. The US Midlife Mortality Crisis Continues: Excess Cause-Specific Mortality During 2020. *Am J Epidemiol*. March 2022. doi:10.1093/aje/kwac055

12. Chen R, Aschmann HE, Chen YH, et al. Racial and Ethnic Disparities in Estimated Excess Mortality From External Causes in the US, March to December 2020. *JAMA Intern Med*. May 2022. doi:10.1001/jamainternmed.2022.1461

13. Vamos EP, Khunti K. Indirect effects of the COVID-19 pandemic on people with type 2 diabetes: time to urgently move into a recovery phase. *BMJ Qual Saf*. October 2021.

doi:10.1136/bmjqs-2021-014079

14. Caleb AM, Gallin S. Policies of Exclusion: The Impact of COVID-19 on People with Disabilities. *Saint Louis University Journal of Health Law & Policy*. 2021;14(2):7.

15. Paglino E, Lundberg DJ, Zhou Z, et al. Differences between reported COVID-19 deaths and estimated excess deaths in counties across the United States, March 2020 to February 2022. *medRxiv*. January 2023. doi:10.1101/2023.01.16.23284633

16. Wang H, Paulson KR, Pease SA, et al. Estimating excess mortality due to the COVID-19 pandemic: a systematic analysis of COVID-19-related mortality, 2020–21. *Lancet*. 2022;399(10334):1513-1536.

17. Beaney T, Clarke JM, Jain V, et al. Excess mortality: the gold standard in measuring the impact of COVID-19 worldwide? *J R Soc Med*. 2020;113(9):329-334.

18. Ruhm CJ. The Evolution of Excess Deaths in the United States During the First Two Years of the COVID-19 Pandemic. *Am J Epidemiol*. May 2023. doi:10.1093/aje/kwad127

19. Lundberg DJ, Wrigley-Field E, Cho A, et al. COVID-19 Mortality by Race and Ethnicity in US Metropolitan and Nonmetropolitan Areas, March 2020 to February 2022. *JAMA Netw Open*. 2023;6(5):e2311098.

20. Polyakova M, Udalova V, Kocks G, Genadek K, Finlay K, Finkelstein AN. Racial Disparities In Excess All-Cause Mortality During The Early COVID-19 Pandemic Varied Substantially Across States. *Health Aff* . 2021;40(2):307-316.

21. Luck AN, Stokes AC, Hempstead K, Paglino E, Preston SH. Associations between mortality from COVID-19 and other causes: A state-level analysis. *PLoS One*. 2023;18(3):e0281683.

22. Chen JT, Krieger N. Revealing the Unequal Burden of COVID-19 by Income, Race/Ethnicity, and Household Crowding: US County Versus Zip Code Analyses. *J Public Health Manag Pract*. 2021;27 Suppl 1, COVID-19 and Public Health: Looking Back, Moing Forward:S43-S56.

23. Rossen LM, Branum AM, Ahmad FB, Sutton P, Anderson RN. Excess Deaths Associated with COVID-19, by Age and Race and Ethnicity - United States, January 26-October 3, 2020. *MMWR Morb Mortal Wkly Rep*. 2020;69(42):1522-1527.

24. Truman BI, Chang MH, Moonesinghe R. Provisional COVID-19 Age-Adjusted Death Rates, by Race and Ethnicity - United States, 2020-2021. *MMWR Morb Mortal Wkly Rep*. 2022;71(17):601-605.

25. Wrigley-Field E, Garcia S, Leider JP, Van Riper D. COVID-19 Mortality At The Neighborhood Level: Racial And Ethnic Inequalities Deepened In Minnesota In 2020. *Health Aff* . 2021;40(10):1644-1653.

26. Chen YH, Matthay EC, Chen R, et al. Excess Mortality in California by Education During the COVID-19 Pandemic. *Am J Prev Med*. 2022;63(5):827.

27. Aschmann HE, Riley AR, Chen R, et al. Dynamics of racial disparities in all-cause mortality during the COVID-19 pandemic. *Proc Natl Acad Sci U S A*. 2022;119(40):e2210941119.

28. Riley AR, Chen YH, Matthay EC, et al. Excess mortality among Latino people in California during the COVID-19 pandemic. *SSM Popul Health*. 2021;15:100860.

29. Pathak EB, Menard J, Garcia RB, Salemi JL. Social class, race/ethnicity, and COVID-19 mortality among working age adults in the United States. *bioRxiv*. November 2021. doi:10.1101/2021.11.23.21266759

30. Zalla LC, Martin CL, Edwards JK, Gartner DR, Noppert GA. A Geography of Risk: Structural Racism and Coronavirus Disease 2019 Mortality in the United States. *Am J Epidemiol*. 2021;190(8):1439-1446.

31. Tan SB, deSouza P, Raifman M. Structural Racism and COVID-19 in the USA: a County-Level Empirical Analysis. *J Racial Ethn Health Disparities*. January 2021. doi:10.1007/s40615-020-00948-8

32. Egede LE, Walker RJ. Structural Racism, Social Risk Factors, and Covid-19 - A Dangerous Convergence for Black Americans. *N Engl J Med*. 2020;383(12):e77.

33. Millett GA, Jones AT, Benkeser D, et al. Assessing differential impacts of COVID-19 on black communities. *Ann Epidemiol*. 2020;47:37-44.

34. Cronin CJ, Evans WN. Excess mortality from COVID and non-COVID causes in minority populations. *Proc Natl Acad Sci U S A*. 2021;118(39). doi:10.1073/pnas.2101386118

35. Woolf SH, Chapman DA, Sabo RT, Zimmerman EB. Excess Deaths From COVID-19 and Other Causes in the US, March 1, 2020, to January 2, 2021. *JAMA*. 2021;325(17):1786-1789.

36. Mulligan CB, Arnott RD. The Young were not Spared: What Death Certificates Reveal about Non-Covid Excess Deaths. *Inquiry*. 2022;59:469580221139016.

37. Abobaker A, Raba AA, Alzwi A. Extrapulmonary and atypical clinical presentations of COVID-19. *J Med Virol*. 2020;92(11):2458-2464.

38. Ahmad FB, Anderson RN, Knight K, Rossen LM, Sutton PD. Advancements in the National Vital Statistics System to Meet the Real-Time Data Needs of a Pandemic. *Am J Public Health*. 2021;111(12):2133-2140.

39. Council of State and Territorial Epidemiologists. Interim Guidance for Public Health Surveillance Programs for Classification of COVID-19- associated Deaths among COVID-19 Cases. CTSE. https://cdn.ymaws.com/www.cste.org/resource/resmgr/pdfs/pdfs2/20211222_interim-guidance.pdf. Published December 22, 2021. Accessed May 20, 2022.

40. Stokes AC, Lundberg DJ, Bor J, Bibbins-Domingo K. Excess Deaths During the COVID-19 Pandemic: Implications for US Death Investigation Systems. *Am J Public Health*. 2021;111(S2):S53-S54.

41. Stokes AC, Lundberg DJ, Bor J, Elo IT, Hempstead K, Preston SH. Association of Health Care Factors With Excess Deaths Not Assigned to COVID-19 in the US. *JAMA Netw Open*. 2021;4(9):e2125287.

42. Institute of Medicine. *Medicolegal Death Investigation System: Workshop Summary*. Washington, DC: The National Academies Press; 2003.

43. Hanzlick RL, Fudenberg J. Coroner versus Medical Examiner Systems: Can We End the Debate? *Academic Forensic Pathology*. 2014;4(1):10-17.

44. Hanzlick R, Combs D. Medical examiner and coroner systems: history and trends. *JAMA*. 1998;279(11):870-874.

45. Dewan S. Failed Autopsies, False Arrests: A Risk of Bias in Death Examinations. *The New York Times*. https://www.nytimes.com/2022/06/20/us/medical-examiners-autopsy-racism.html. Published June 20, 2022. Accessed April 1, 2023.

46. Denham A, Vasu T, Avendano P, Boslett A, Mendoza M, Hill EL. Coroner county systems are associated with a higher likelihood of unclassified drug overdoses compared to medical examiner county systems. *Am J Drug Alcohol Abuse*. 2022;48(5):606-617.

47. Documenting COVID-19 project and USA TODAY Network. Uncounted: Inaccurate death certificates across the country hide the true toll of COVID-19. *USA Today*. https://www.usatoday.com/in-depth/news/nation/2021/12/22/covid-deaths-obscured-inaccurate-death-certificates/8899157002/. Published December 26, 2021.

48. Hanzlick R. Coroner training needs. A numeric and geographic analysis. *JAMA*. 1996;276(21):1775-1778.

49. Xue Y, Lai L, Liu C, Niu Y, Zhao J. Perspectives on the death investigation during the COVID-19 pandemic. *Forensic Sci Int Synerg*. 2020;2:126-128.

50. Stokes AC, Weiss J, Lundberg DJ, et al. Estimates of the Association of Dementia With US Mortality Levels Using Linked Survey and Mortality Records. *JAMA Neurol*. 2020;77(12):1543-1550.

51. Feldman JM, Gruskin S, Coull BA, Krieger N. Quantifying underreporting of law-enforcement-related deaths in United States vital statistics and news-media-based data sources: A capture-recapture analysis. *PLoS Med*. 2017;14(10):e1002399.

52. Feldman JM, Bassett MT. Monitoring Deaths in Police Custody: Public Health Can and Must Do Better. *Am J Public Health*. 2021;111(S2):S69-S72.

Figure 1. Ratio of the predicted total number of COVID-19 deaths to the number of COVID-19 deaths recorded on death certificates across individual-level decedent characteristics

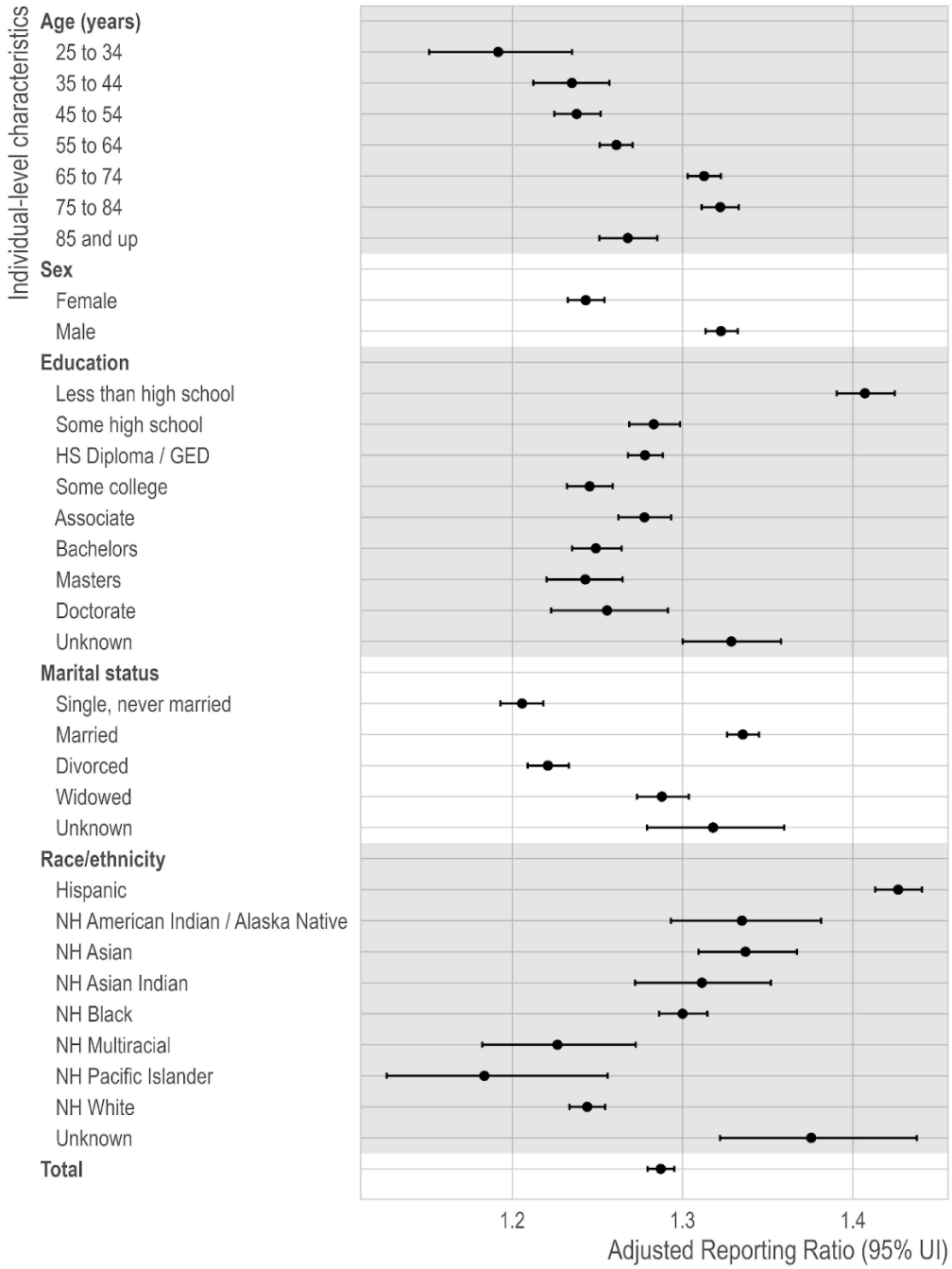


Figure 2. Ratio of the predicted total number of COVID-19 deaths to the number of COVID-19 deaths recorded on death certificates across county-level characteristics

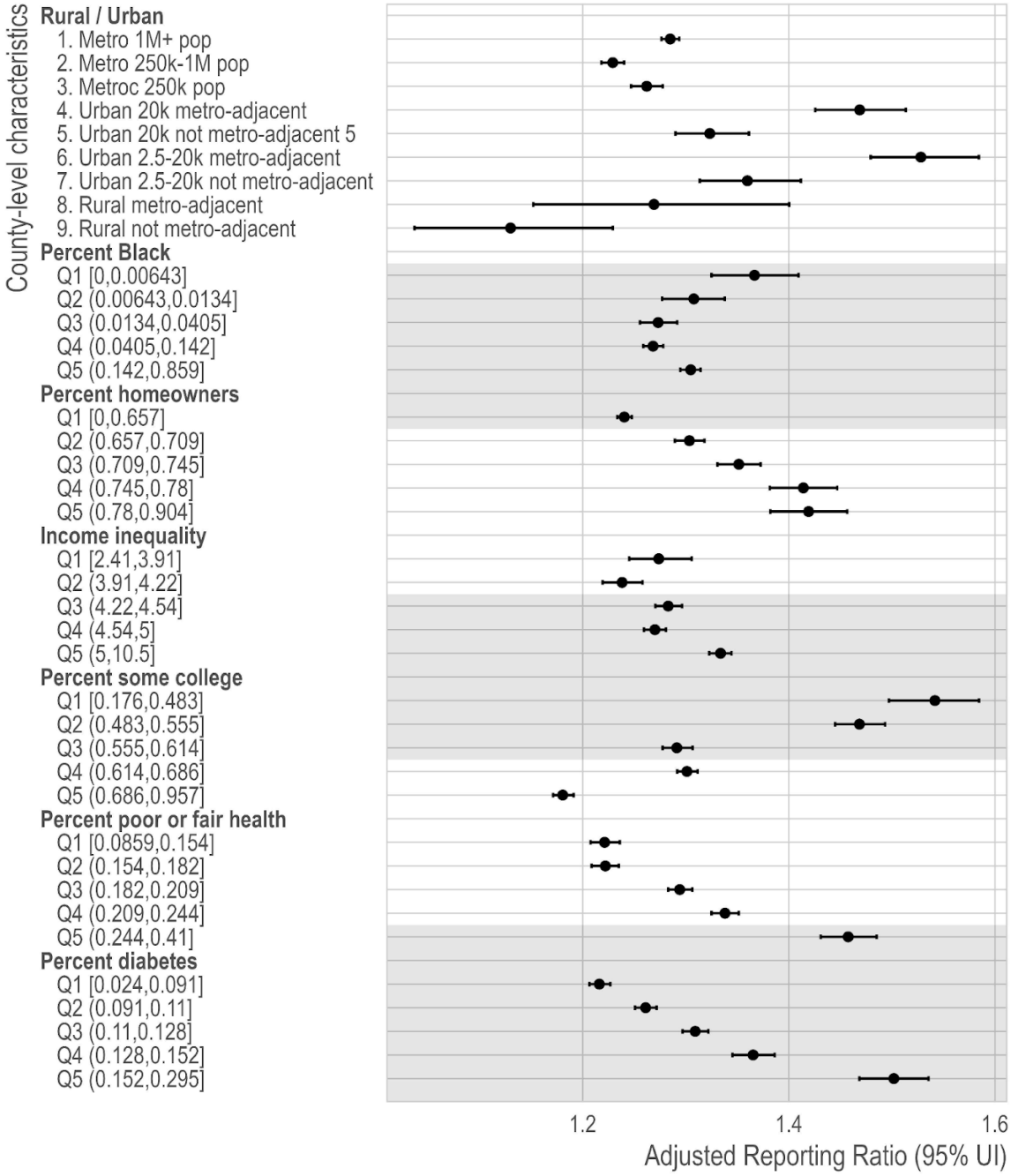
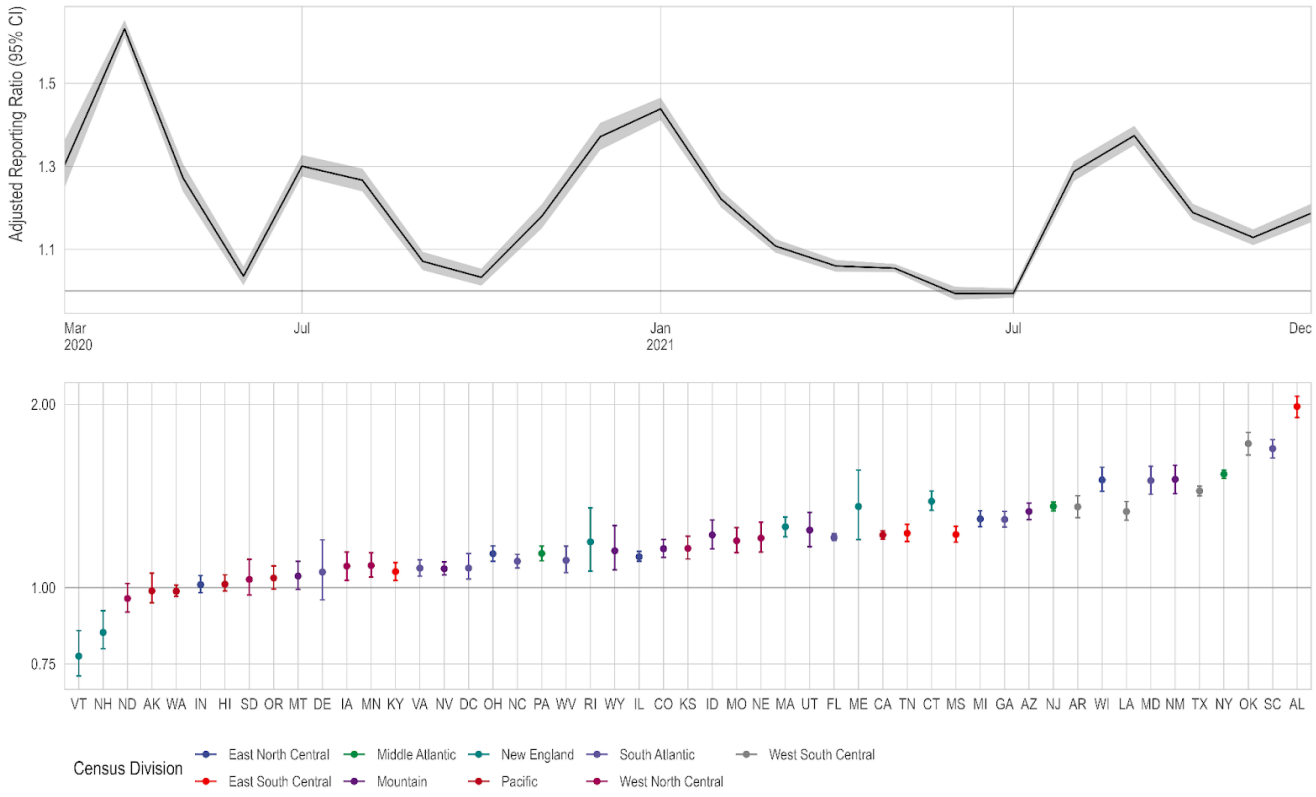


Figure 3. Monthly and state-level ratios of the predicted total number of COVID-19 deaths to the number of COVID-19 deaths recorded on death certificates



Appendix

Table S1. Performance metrics across all models

Model	Preprocessing	AUC ROC	Sens.	Spec.	Accu.	NPV	PPV	F1	J	MCC	Kappa	Bal. Accu.	Accu. ratio	AUC PR
XGBoost	No county, no contributing cause	**0.902**	0.799	0.849	**0.835**	0.916	0.671	**0.729**	**0.648**	**0.617**	**0.612**	**0.824**	**0.804**	**0.806**
XGBoost	No county	**0.902**	**0.801**	0.846	0.834	**0.917**	0.668	**0.729**	0.647	0.615	0.610	**0.824**	**0.804**	**0.806**
XGBoost	No contributing causes	0.901	0.797	0.849	**0.835**	0.915	0.671	**0.729**	0.646	0.615	0.611	0.823	0.802	0.805
XGBoost	All covariates	0.901	0.798	0.847	0.834	0.916	0.669	0.728	0.645	0.614	0.609	0.823	0.802	0.805
LightGBM	No county, no contributing cause	0.899	0.793	0.849	0.834	0.914	0.670	0.727	0.643	0.613	0.608	0.821	0.798	0.801
LightGBM	No county	0.899	0.795	0.847	0.832	0.914	0.667	0.725	0.641	0.611	0.606	0.821	0.797	0.801
LightGBM	No contributing causes	0.897	0.790	0.848	0.832	0.913	0.668	0.724	0.638	0.608	0.604	0.819	0.794	0.798
LightGBM	All covariates	0.896	0.789	0.848	0.832	0.912	0.667	0.723	0.637	0.608	0.603	0.819	0.792	0.796
Random Forest	No county, no contributing cause	0.895	0.786	0.852	0.833	0.912	0.672	0.725	0.638	0.610	0.606	0.819	0.790	0.794
Random Forest	No county	0.895	0.783	**0.854**	0.834	0.911	0.675	0.725	0.637	0.611	0.607	0.819	0.790	0.794
Random Forest	No contributing causes	0.895	0.785	0.853	0.834	0.911	0.674	0.725	0.638	0.611	0.607	0.819	0.791	0.795
Random Forest	All covariates	0.895	0.784	0.854	0.834	0.911	**0.675**	0.725	0.638	0.611	0.608	0.819	0.790	0.794
BART	No county, no contributing cause	0.894	0.786	0.846	0.829	0.911	0.663	0.720	0.632	0.602	0.598	0.816	0.788	0.793
BART	No county	0.894	0.787	0.846	0.829	0.911	0.663	0.720	0.633	0.603	0.598	0.816	0.789	0.794
Logistic (ElasticNet)	No county, no contributing cause	0.864	0.727	0.850	0.815	0.889	0.651	0.687	0.576	0.558	0.556	0.788	0.729	0.741
Logistic (ElasticNet)	No county	0.864	0.727	0.849	0.815	0.889	0.650	0.687	0.576	0.558	0.556	0.788	0.729	0.741
Logistic (ElasticNet)	No contributing causes	0.869	0.737	0.848	0.817	0.893	0.652	0.692	0.585	0.564	0.562	0.792	0.737	0.744
Logistic (ElasticNet)	All covariates	0.868	0.734	0.850	0.817	0.892	0.654	0.692	0.584	0.565	0.563	0.792	0.737	0.744
Logistic (LASSO)	No county, no contributing cause	0.864	0.726	0.850	0.815	0.889	0.651	0.687	0.576	0.558	0.556	0.788	0.729	0.741
Logistic (LASSO)	No county	0.864	0.727	0.849	0.815	0.889	0.651	0.687	0.576	0.558	0.556	0.788	0.729	0.741

Figure S1. Receiver operating characteristic curve of the primary model (black) and all other models (grey). The white circle represents the model performance at the standard threshold (Pr>.5).

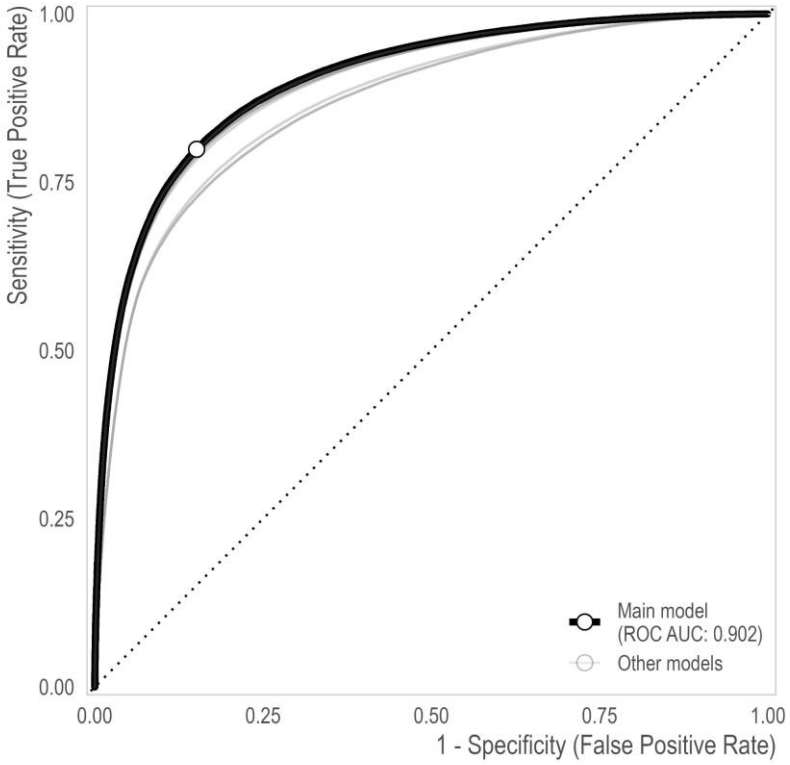


Figure S2. Precision-recall curve of the primary model (black) and all other models (grey). The white circle represents the model performance at the standard threshold (Pr>.5).

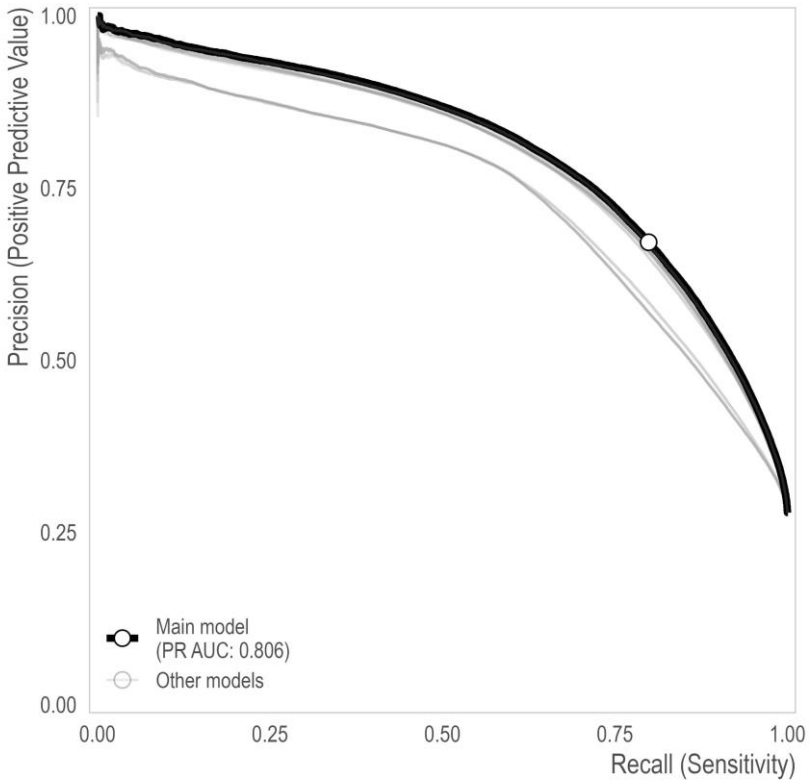


Figure S3. Cumulative gain of the primary model (black) and all other models (grey) compared to a perfect classifier (red) and random chance (blue).

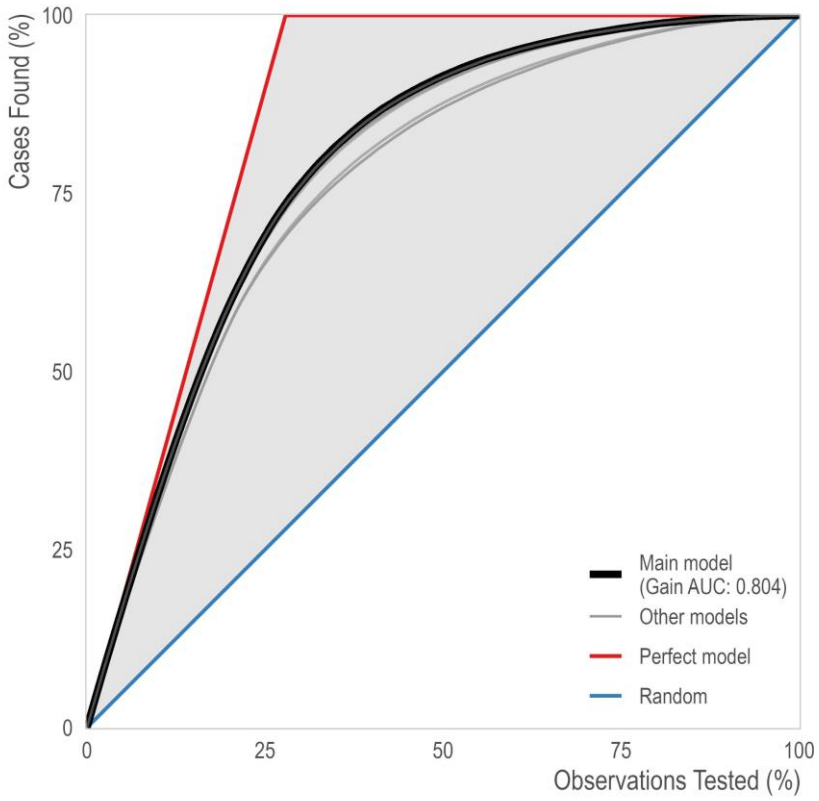


Figure S4. Lift curve of the primary model (black) and all other models (grey).