Check for updates

CORONAVIRUS

Applying machine learning to identify unrecognized COVID-19 deaths recorded as other causes of death in the United States

Mathew V. Kiang[1], Zehang Richard Li[2], Elizabeth Wrigley-Field[3], Rafeya V. Raquib[4], Dielle J. Lundberg[4], Eugenio Paglino[5,6], Benjamin Huynh[7], Kirsten Bibbins-Domingo[8], M. Maria Glymour[9], Andrew C. Stokes[4]*

Copyright © 2026 The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original U.S. Government Works. Distributed under a Creative Commons Attribution NonCommercial License 4.0 (CC BY-NC).

The actual number of US deaths caused by severe acute respiratory syndrome coronavirus 2 infection has been investigated and debated since the start of the COVID-19 pandemic. Here, we use machine learning trained on US death certificates from March 2020 to December 2021 to predict 155,536 (95% uncertainty interval: 150,062 to 161,112) unrecognized COVID-19 deaths. This indicates that 19% more COVID-19 deaths occurred in the US than officially reported. Predicted unrecognized COVID-19 deaths occurred disproportionately among decedents with less than a high school education; decedents identified as Hispanic, American Indian, Alaska Native, Asian, and/or Black; counties with lower household incomes and worse preexisting health; and counties in the South. These findings suggest that the US death investigation system undercounted COVID-19 deaths unevenly, hiding the true extent of inequities.

INTRODUCTION

Accurate and timely mortality statistics are critical for health system responses during public health emergencies (*1*). Throughout the COVID-19 pandemic in the United States (US), official COVID-19 mortality reporting was often delayed or incomplete (*2*). Policy-makers in federal, state, and county governments—along with decision-makers in hospitals, schools, prisons, jails, long-term care facilities, and other organizations—frequently used these data in the pandemic response to inform protective behaviors and health policies (*3*–*7*). For this reason, inaccuracies and delays in mortality data could influence both policy and population health (*8*, *9*).

Although almost all deaths in the US are recorded, cause-of-death assignment is a complex and resource-intensive process with varying accuracy (*10*, *11*). A COVID-19 death may go unrecognized because of factors related to the decedent and their family (e.g., attitudes and stigma about COVID-19 and testing behaviors) (*12*, *13*), death investigators (e.g., political roles, conflicts of interests, and expertise) (*14*–*16*), the community (e.g., COVID-19 testing rates and distribution of preexisting health conditions that complicate diagnosis) (*17*, *18*), health systems (e.g., funding for death investigation offices, types of offices in a county and state, and related policies) (*19*–*21*), and larger systems (e.g., impacts of racism, ableism, and classism) (*22*–*26*). Given these determinants, it is possible that COVID-19 deaths were undercounted in the US and that undercounting occurred unevenly across sociodemographic groups.

Prior efforts to quantify unrecognized COVID-19 deaths

Research on unrecognized COVID-19 deaths in the US has largely relied on excess mortality models, which estimate mortality attributable to the pandemic by comparing observed all-cause deaths to those expected on the basis of prepandemic trends (*27*). Prior studies (with different modeling specifications) have estimated that excess mortality exceeded reported COVID-19 mortality by 28% in early 2020 (*28*), 38% in 2020 (*29*), and 14% in 2020 (*30*).

Excess mortality models have largely relied on all-cause mortality data, which include deaths from external causes (e.g., intentional or unintentional injuries) unlikely to be caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection in the short term (*31*, *32*). A prior study excluded external-cause deaths to estimate natural-cause excess deaths, which exceeded reported COVID-19 deaths by 16% from March 2020 to August 2022 (*33*). This study and a prior study from California showed strong temporal concordance between reported COVID-19 deaths and excess deaths reported to other natural causes, suggesting that many such deaths were unrecognized COVID-19 deaths (*33*, *34*).

As more detailed mortality data have become available, researchers have sought to produce more refined estimates of excess mortality (*35*, *36*). One prior study used age-, state-, and cause-specific excess mortality models to conclude that 84% of pandemic-era all-cause excess mortality could be directly attributed to SARS-CoV-2 infection (*37*). Excess mortality models have value in the context of emerging health crises as a means of estimating total mortality attributable to the crisis and, in the case of the pandemic, approximating unrecognized COVID-19 deaths (*8*, *27*). This prior evidence, however, is suggestive but not conclusive because excess mortality models do not directly distinguish between deaths from SARS-CoV-2 infection and those caused by pandemic-era health care interruptions and/or social and economic impacts among individuals with no recent SARS-CoV-2 infection (*38*–*40*). Differentiating deaths due to infection from deaths

[1]Department of Epidemiology and Population Health, Stanford University, Stanford, CA, USA. [2]Department of Statistics, University of California Santa Cruz, Santa Cruz, CA, USA. [3]Department of Sociology and Minnesota Population Center, University of Minnesota, Minneapolis, MN, USA. [4]Department of Global Health, Boston University School of Public Health, Boston, MA, USA. [5]Max Planck – University of Helsinki Center for Social Inequalities in Population Health, Finland. [6]Helsinki Institute for Demography and Population Health, Faculty of Social Sciences, University of Helsinki, Finland. [7]Department of Environmental Health and Engineering, Johns Hopkins University, Baltimore, MD, USA. [8]Department of Epidemiology and Biostatistics, University of California San Francisco, San Francisco, CA, USA. [9]Department of Epidemiology, Boston University School of Public Health, Boston, MA, USA.
*Corresponding author. Email: acstokes@bu.edu

due to pandemic response policies is at the heart of the policy debate about appropriate emergency response, so this uncertainty is a critical limitation (*41*).

Potential inequities in unrecognized COVID-19 deaths

Although a large body of research has examined geographic and sociodemographic inequities in reported COVID-19 and excess mortality during the pandemic (*42*–*50*), few studies have explored inequities in COVID-19 deaths that have gone unrecognized. One prior study found that excess deaths attributed to non–COVID-19 causes were more common in counties with lower socioeconomic status, greater prevalence of preexisting health conditions, and a greater fraction of non-Hispanic Black residents (*51*). Another study found that excess deaths attributed to non–COVID-19 natural causes were more common in rural areas, the South, and the West (*33*). While these studies did not examine unrecognized COVID-19 deaths directly, they suggest that the death investigation system may have performed unevenly during the pandemic (*8*).

Applying machine learning to estimate unrecognized COVID-19 deaths

Machine learning algorithms used to classify the cause of death may outperform excess death models if there is a reliable set of correctly classified deaths available to train the algorithm. Wrigley-Field *et al.* (*52*) proposed using cause and place of death information to differentiate deaths directly and indirectly related to COVID-19. The logic behind this approach is that although individuals who died outside a hospital were often not tested for SARS-CoV-2 infection, such testing was nearly universal among hospital patients, starting as early as April 2020 (*53*–*56*). Furthermore, while there were many excess natural-cause deaths attributable to the pandemic that were reported to causes other than COVID-19 in out-of-hospital settings, the number of reported COVID-19 deaths in hospitals was similar to the number of excess natural-cause deaths, suggesting that COVID-19 deaths were accurately identified here (*34*). Thus, in-hospital deaths, during this period of near-universal testing, provide a pool of high-quality training data for classification of whether a death was due to SARS-CoV-2 infection for use in predicting whether deaths in out-of-hospital settings were likely to be COVID-19 related. Here, we applied predictive machine learning to this strategy, building on prior research on machine learning for predictive inference to improve the accuracy of disease diagnosis and postmortem death investigation (*57*–*61*).

This study had two research questions: (i) How many unrecognized COVID-19 deaths occurred from March 2020 to December 2021 in the US, and (ii) did unrecognized COVID-19 deaths occur in some sociodemographic populations disproportionately? We quantified unrecognized COVID-19 deaths using two distinct but related metrics: (i) the predicted number of unrecognized COVID-19 deaths in out-of-hospital settings (i.e., the difference between the number of out-of-hospital deaths predicted by our machine learning algorithm as likely to be COVID-19 related and the officially reported number of out-of-hospital COVID-19 deaths) and (ii) an adjusted reporting ratio (ARR) that estimated underreporting of COVID-19 deaths across all settings of death by weighting the estimated degree of misclassification of COVID-19 deaths in out-of-hospital settings by the proportion of deaths that occurred in out-of-hospital settings. The ARR metric aligns with our research question of examining whether unrecognized COVID-19 deaths occurred in some sociodemographic populations disproportionately (across all settings of death).

RESULTS

Finding 1: National estimate of unrecognized COVID-19 deaths

From March 2020 to December 2021, our estimate of total COVID-19 deaths in the US {995,787 deaths [95% uncertainty interval (UI): 990,313 to 1,001,363]} exceeded the 840,251 COVID-19 deaths officially reported (i.e., those with COVID-19 listed anywhere on their death certificate) by 19% [ARR: 1.19 (95% UI: 1.18 to 1.19)]. This equals 155,536 predicted unrecognized COVID-19 deaths (95% UI: 150,062 to 161,112) in out-of-hospital settings during this period.

The number of COVID-19 deaths in decedents' homes was 160% higher than officially reported [ARR: 2.60 (95% UI: 2.56 to 2.65)], indicating that there were 111,245 predicted unrecognized COVID-19 deaths that occurred in homes (95% UI: 108,372 to 114,210) (Table 1). Predicted unrecognized COVID-19 deaths also occurred in hospice settings [17,346 deaths (95% UI: 16,832 to 17,873)], outpatient and emergency room settings [14,832 deaths (95% UI: 14,364 to 15,315)], and other settings not elsewhere classified [9452 deaths (95% UI: 9011 to 9926)].

We found similar model performance across racial and ethnic groups, sex, and educational attainment (table S1 and fig. S1), the rural-urban continuum (table S2), and states (table S3). Sensitivity analyses demonstrated that our national estimate remained consistent even when we made a different model selection {ARR: 1.17 [95% confidence interval (CI): 1.16 to 1.18]} (table S4). We also found that our results were conservative compared to other (empirically untestable) assumptions of misclassification (ARR increased up to a maximum of 1.44).

Finding 2: Geographic and temporal differences in unrecognized deaths

COVID-19 deaths were most likely to go unrecognized in the South (Table 1). The estimated number of COVID-19 deaths was 31% higher than officially reported (95% UI: 29 to 33%) in the West South Central division, 26% higher than officially reported (95% UI: 24 to 27%) in the Middle Atlantic division, and 25% higher than officially reported (95% UI: 23 to 28%) in the East South Central division.

Figure 1 represents the degree of predicted underreporting of COVID-19 deaths in the US by county and state. The states that we identified where COVID-19 deaths were most likely to go unrecognized were Alabama [ARR: 1.67 (95% UI: 1.61 to 1.74)], Oklahoma [ARR: 1.51 (95% UI: 1.46 to 1.57)], and South Carolina [ARR: 1.47 (95% UI: 1.42 to 1.52)]. The states with the largest estimates of predicted unrecognized COVID-19 deaths were Texas [24,024 deaths (95% UI: 22,381 to 25,740)], New York [23,005 deaths (95% UI: 21,690 to 24,330)], California [11,613 deaths (95% UI: 10,408 to 12,846)], Alabama [11,501 deaths (95% UI: 10,435 to 12,583)], Florida [7718 deaths (95% UI: 6848 to 8592)], and South Carolina [7224 deaths (95% UI: 6468 to 8027)] (table S5).

Figure 2 shows the temporal evolution of COVID-19 mortality reporting by pandemic wave and month. COVID-19 deaths were most likely to go unrecognized in the initial wave of the pandemic [March to May 2020; highest ARR during this wave: 1.49 (95% UI: 1.48 to 1.51)]. Predicted unrecognized COVID-19 deaths declined in likelihood but continued to occur in subsequent waves [highest monthly ARR during the second wave: 1.19 (95% UI: 1.17 to 1.21); during the Alpha wave: 1.34 (95% UI: 1.32 to 1.36); during the Delta wave: 1.23 (95% UI: 1.22 to 1.25); during the start of the Omicron wave: 1.09 (95% UI: 1.07 to 1.11)]. By month, the largest absolute number of predicted unrecognized COVID-19 deaths occurred in the months of January 2021

Table 1. Estimates of total predicted, officially reported, and predicted unrecognized COVID-19 deaths in the United States overall, by census division, and by place of death, March 2020 to December 2021. The ARR shows the extent by which total predicted COVID-19 deaths exceeded officially reported COVID-19 deaths during the period. For example, an ARR of 1.19 indicates that our model predicted that total COVID-19 deaths exceeded official reports by 19%. Our model assumes that inpatient deaths have an ARR of 1. We estimated 95% UIs by using 5000 bootstrapped samples of the out-of-hospital predictions and reporting the 2.5th and 97.5th quantiles. Predictions are based on the median of 5000 bootstrapped samples; subgroup estimates may not sum exactly to the overall total.

	Absolute number of COVID-19 deaths			
Place	**Total predicted (95% UI)**	**Officially reported**	**Predicted unrecognized (95% UI)**	**Adjusted reporting ratio (95% UI)**
United States	995,787 (990,313 to 1,001,363)	840,251	155,536 (150,062 to 161,112)	1.19 (1.18 to 1.19)
Census divisions				
East North Central	137,635 (135,714 to 139,658)	123,922	13,713 (11,792 to 15,736)	1.11 (1.10 to 1.13)
East South Central	76,775 (75,463 to 78,156)	61,238	15,537 (14,225 to 16,918)	1.25 (1.23 to 1.28)
Middle Atlantic	161,483 (159,574 to 163,391)	128,502	32,981 (31,072 to 34,889)	1.26 (1.24 to 1.27)
Mountain	72,044 (70,946 to 73,233)	62,138	9,906 (8,808 to 11,095)	1.16 (1.14 to 1.18)
New England	38,739 (37,775 to 39,683)	32,306	6,433 (5,469 to 7,377)	1.20 (1.17 to 1.23)
Pacific	109,658 (108,417 to 110,941)	99,067	10,591 (9,350 to 11,874)	1.11 (1.09 to 1.12)
South Atlantic	189,612 (187,581 to 191,593)	163,877	25,735 (23,704 to 27,716)	1.16 (1.14 to 1.17)
West North Central	56,445 (55,072 to 57,877)	52,058	4,387 (3,014 to 5,819)	1.08 (1.06 to 1.11)
West South Central	153,402 (151,351 to 155,461)	117,143	36,259 (34,208 to 38,318)	1.31 (1.29 to 1.33)
Places of death (out-of-hospital only)				
Home	180,560 (177,687 to 183,525)	69,315	111,245 (108,372 to 114,210)	2.60 (2.56 to 2.65)
Hospice	41,544 (41,030 to 42,071)	24,198	17,346 (16,832 to 17,873)	1.72 (1.70 to 1.74)
Hospital: Dead on arrival	1,256 (1,211 to 1,303)	805	451 (406 to 498)	1.56 (1.50 to 1.62)
Hospital: Outpatient or emergency room	43,282 (42,814 to 43,765)	28,450	14,832 (14,364 to 15,315)	1.52 (1.50 to 1.54)
Nursing home or long-term care facility	129,229 (127,222 to 131,241)	127,038	2,191 (184 to 4203)	1.02 (1.00 to 1.03)
Other	25,728 (25,287 to 26,202)	16,276	9,452 (9,011 to 9,926)	1.58 (1.55 to 1.61)
Unknown	133 (121 to 145)	113	20 (8 to 32)	1.18 (1.07 to 1.28)

[35,665 deaths (95% UI: 33,248 to 38,208)] and April 2020 [32,110 deaths (95% UI: 30,978 to 33,256)] (table S6). As shown in the bottom panel of Fig. 2, the peaks in predicted unrecognized COVID-19 deaths correspond to the monthly peaks in official COVID-19 deaths: The largest number of official COVID-19 deaths occurred in January 2021 along with another peak in April 2020. This correspondence supports the finding that the unrecognized COVID-19 deaths we predicted were COVID-19 deaths.

Finding 3: Differences in unrecognized deaths by decedent characteristics

We next examined the predicted accuracy of COVID-19 reporting by decedent characteristics as recorded on death certificates (Fig. 3). For age, COVID-19 deaths between the ages of 65 to 84 years were most likely to go unrecognized [ARR for 65 to 74 years: 1.21 (95% UI: 1.21 to 1.22); ARR for 75 to 84 years: 1.22 (95% UI: 1.21 to 1.23)]. COVID-19 deaths recorded as male were more likely to go unrecognized than deaths recorded as female [ARR for male: 1.22 (95% UI: 1.21 to 1.23); ARR for female: 1.14 (95% UI: 1.13 to 1.15)]. COVID-19 deaths were also more likely to go unrecognized for those identified as having less than any high school education compared to those recorded as having more education [ARR for less than high school: 1.29 (95% UI: 1.28 to 1.31); ARR for high school diploma or equivalent: 1.18 (95% UI: 1.17 to 1.18); ARR for some college: 1.15 (95% UI: 1.14 to 1.16)]. There was also variability by marital status.

COVID-19 deaths among individuals identified as Hispanic were most likely to go unrecognized [ARR: 1.31 (95% UI: 1.30 to 1.32)]. Compared to decedents recorded as non-Hispanic white [ARR: 1.15 (95% UI: 1.14 to 1.16)], deaths were also more likely to go unrecognized if the decedent was listed on certificates as non-Hispanic American Indian and Alaska Native [ARR: 1.24 (95% UI: 1.21 to 1.28)], non-Hispanic Asian [ARR excluding Indian: 1.24 (95% UI: 1.22 to 1.27); ARR for Indian: 1.23 (95% UI: 1.19 to 1.26)], or non-Hispanic Black [ARR: 1.19 (95% UI: 1.18 to 1.20)]. In absolute terms, 78,561 predicted unrecognized COVID-19 deaths (95% UI: 73,783 to 83,447) occurred among non-Hispanic white adults, 43,132 deaths (95% UI: 41,632 to 44,622) among Hispanic adults, 23,617 deaths (95% UI: 22,269 to 24,985) among non-Hispanic Black adults, 5772 deaths (95% UI: 5285 to 6328) among non-Hispanic Asian adults, and 2318 deaths (95% UI: 1982 to 2650) among non-Hispanic American Indian and Alaska Native adults (table S7). The unrecognized COVID-19 deaths that our machine learning approach predicted were attributed to a variety of underlying causes of death including Alzheimer's disease and related dementia, cardiovascular disease, and diabetes (table S8).

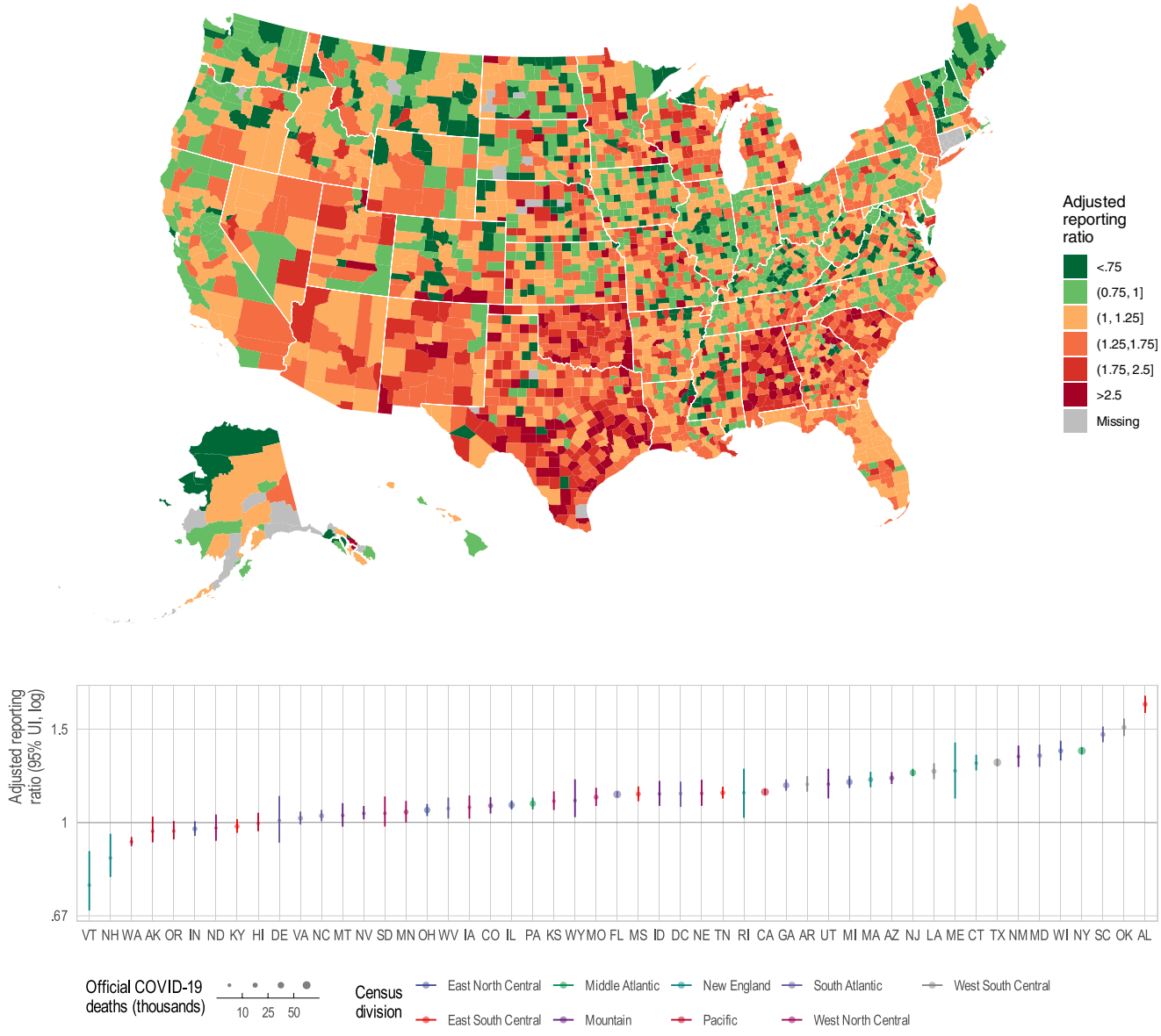

Fig. 1. ARRs comparing total predicted COVID-19 deaths to those officially reported in each county and state, March 2020 to December 2021. The ARR shows the extent by which total predicted COVID-19 deaths exceeded officially reported COVID-19 deaths during the period. A limited number of counties had ARRs < 1, which suggests that there were more officially reported COVID-19 deaths than total predicted COVID-19 deaths. One reason that a county could have an ARR < 1 is if death certifiers recorded people as dying from COVID-19 when they had COVID-19 but actually died from another unrelated cause. The map (top) shows ARRs for each county. Counties with missing data had missing county-level covariates that prevented prediction. The forest plot (bottom) shows ARRs for each state, sorted from states with the lowest ARR (e.g., Vermont where there were 24% fewer total predicted COVID-19 deaths than officially reported) to those with the largest ARR (e.g., Alabama where there were 67% more total predicted COVID-19 deaths than officially reported).

Finding 4: Differences in unrecognized deaths by county characteristics

Our predictions showed that reporting of COVID-19 deaths also varied by county-level characteristics (Fig. 4). COVID-19 deaths were more likely to go unrecognized in counties in the lowest quintile of median household income [ARR: 1.34 (95% UI: 1.31 to 1.36)] and the lowest quintile of the proportion of residents with some or more college [ARR: 1.37 (95% UI: 1.33 to 1.41)]. COVID-19 deaths were more likely to go unrecognized in counties with the most residents reporting poor or fair health [ARR: 1.30 (95% UI: 1.28 to 1.33)] and with a higher percentage of residents living with diabetes [ARR: 1.33 (95% UI: 1.30 to 1.35)]. COVID-19 death reporting accuracy also varied across the rural-urban continuum.

DISCUSSION

Implications for the COVID-19 pandemic

Our machine learning approach predicted 995,787 COVID-19 deaths in the US from March 2020 to December 2021 and 155,536 unrecognized COVID-19 deaths. This finding adds to the evidence base documenting uncounted COVID-19 deaths nationally (*33*, *34*, *37*). Prior attempts to study unrecognized COVID-19 deaths have largely relied

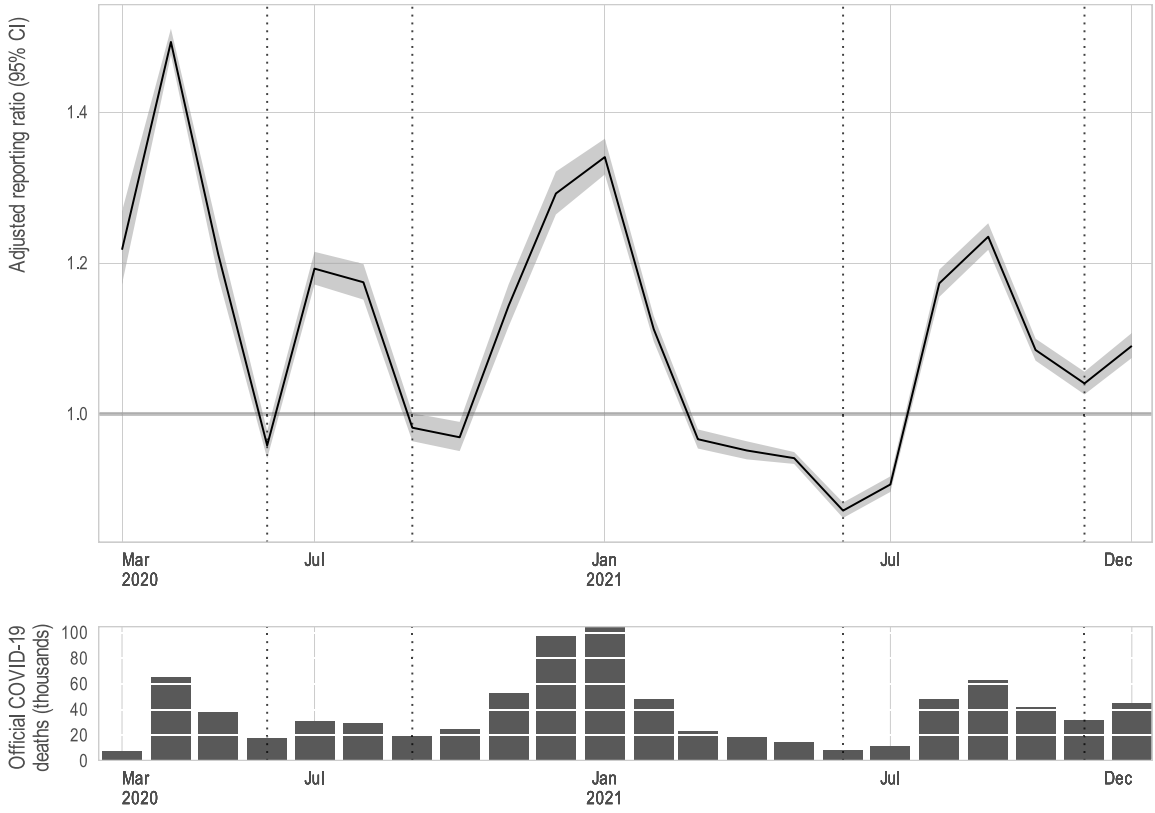

Fig. 2. ARRs comparing total predicted COVID-19 deaths to those officially reported over each month and pandemic wave, March 2020 to December 2021. The ARR shows the extent by which total predicted COVID-19 deaths exceeded officially reported COVID-19 deaths during the period. The line graph (top) shows ARRs (black line) with 95% UIs (gray shading) in the US for each month. The bar chart (bottom) shows the absolute number of officially reported COVID-19 deaths each month. We used dotted lines to delineate the initial wave (March to May 2020), second wave (June to August 2020), Alpha wave (September 2020 to May 2021), Delta wave (June to October 2021), and start of the Omicron wave (November to December 2021).

on excess mortality models, which cannot directly distinguish whether an excess death was due to COVID-19 or the pandemic's indirect effects (*41*). Our national estimate of total predicted COVID-19 deaths is similar to the estimate of the US Centers for Disease Control and Prevention (CDC) of just under 1 million excess deaths for this period (*62*). Excess mortality models assume that projected historical mortality trends accurately estimate expected mortality in the absence of the pandemic and are sensitive to the choice of historical period (*27*). Our machine learning approach relies on different assumptions—(i) that there was no measurement error for COVID-19 deaths in hospitals and (ii) that a model trained on in-hospital deaths is transportable to out-of-hospital settings. It also enables subgroup estimates that are typically infeasible with excess mortality models. Estimates in this study should be interpreted alongside other estimates derived using different methodologies to support evidence triangulation (*63*).

The primary contribution of this study is our predictions showing that unrecognized COVID-19 deaths occurred unevenly, disproportionately affecting decedents identified as Hispanic, American Indian, Alaska Native, Black, and Asian; decedents with less than high school education; and decedents who resided in counties with lower socioeconomic status and worse prepandemic health. Together, this study indicates that the US death investigation system reported COVID-19 deaths in a systematically inequitable way that hid the true extent of pandemic mortality and inequities. While this study does not isolate mechanisms in the death investigation system, the communities affected by the undercounting of COVID-19 deaths could be interpreted as a pattern of structural racism, classism, and ableism in the death investigation system that warrants further research and policy attention (*22*–*26*).

Some public health groups and community organizations have expressed concerns that incomplete reporting of COVID-19 deaths contributed to an insufficient pandemic response (*5*, *25*, *40*, *64*–*66*). Many of the communities in which we predicted unrecognized COVID-19 deaths were also the populations where the largest number of total COVID-19 deaths occurred (e.g., Latino populations, Indigenous populations, and people with preexisting health conditions or disabilities), a consequence of health and social policy failures (*67*). In this way, whether intentional or not, undercounting of COVID-19 deaths has functioned to absolve politicians and health policy-makers of responsibility for some of these failures.

This study did not directly examine actors, policies, and structures in death investigation and other health systems that may have contributed to inaccuracies in COVID-19 death reporting. Death investigators faced difficult circumstances during the pandemic, and many put themselves at occupational health risk to investigate COVID-19 deaths (*14*). This study's predictions are suggestive of biased outcomes in the investigations carried out in this challenging context, a possibility that the field's professional ethics recognize as a risk to guard against

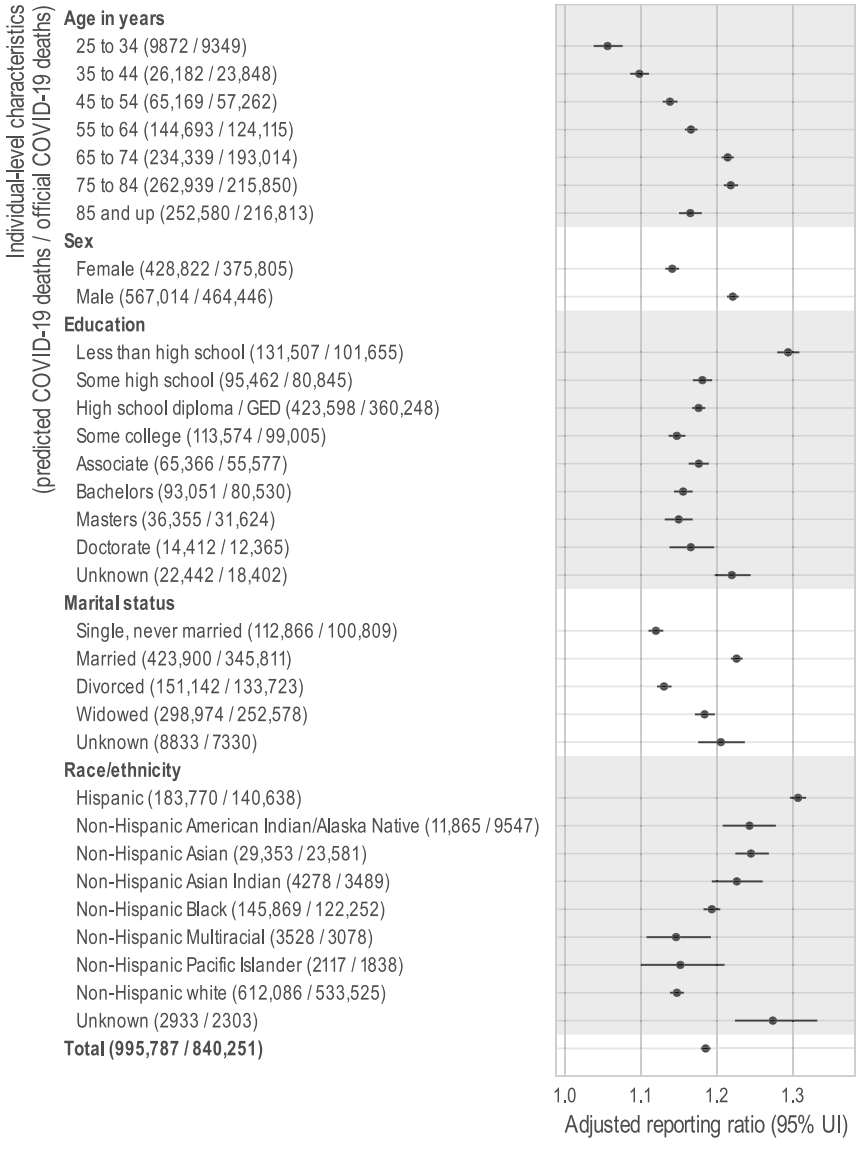

Fig. 3. ARRs comparing total predicted COVID-19 deaths to those officially reported by decedent characteristics, March 2020 to December 2021. The ARR shows the extent by which total predicted COVID-19 deaths exceeded officially reported COVID-19 deaths during the period. Decedent characteristics are reported by families and death investigators on death certificates and are limited to the above options for each characteristic. Potential misclassification and biases in sociodemographic data listed on death certificates are discussed in the Supplementary Text. Predictions are based on the median of 5000 bootstrapped samples; subgroup estimates may not sum exactly to the overall total. GED, General Educational Development test.

(*68*). Some policies and structural factors that may be relevant include inadequate and inequitable funding for death investigation offices to hire staff and conduct investigations (e.g., postmortem COVID-19 testing) (*69*, *70*), variability in training required for death investigators (*14*, *16*, *71*), and elected roles presenting conflicts of interests (i.e., sheriff-coroners) (*11*, *19*). While community variables (i.e., access to COVID-19 testing, health system distrust, and preexisting conditions that may complicate death investigation) could all also influence reporting accuracy, a well-functioning death investigation system would ideally account for these determinants (*12*, *13*, *17*, *18*). Another mechanism that may have contributed to inaccuracies in reporting is COVID-19 stigma and related political attitudes (*72*, *73*). These factors may have affected whether a decedent sought testing in the weeks before their death, whether their family members reported COVID-19 symptoms to an investigator, and whether investigators followed established procedures (*12*, *14*, *74*–*76*).

Broader implications for death investigation systems

The machine learning approach used in this study could be further developed and adapted for use in public health responses in a number of settings where cause-of-death investigation is incomplete, delayed, and/or suspected to be biased (*77*, *78*). In the case of COVID-19, numerous policies and programs (e.g., Federal Emergency Management Agency funeral assistance program) were designed around officially reported COVID-19 deaths (*4*, *8*, *79*). Other settings that may be particularly relevant for the application of these methods are deaths related to drug overdose (*80*), deaths in police custody (*15*), deaths related

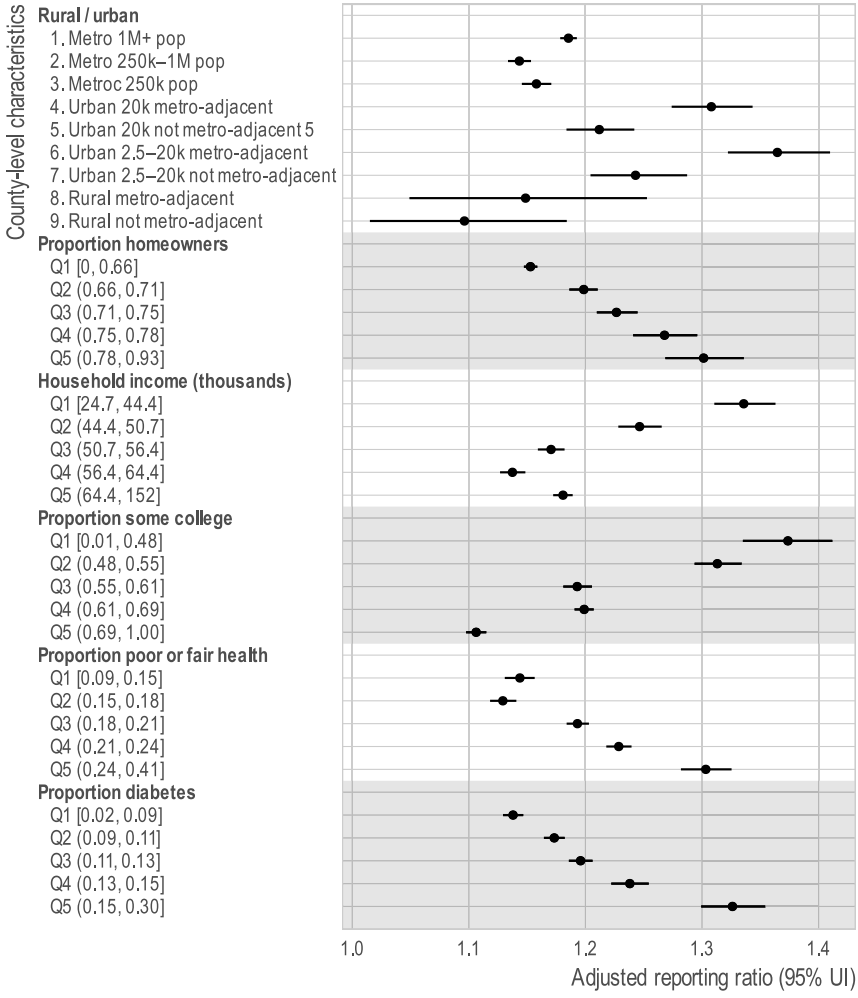

Fig. 4. ARRs comparing total predicted COVID-19 deaths to those officially reported by county of residence characteristics, March 2020 to December 2021. The ARR shows the extent by which total predicted COVID-19 deaths exceeded officially reported COVID-19 deaths during the period. County characteristics are divided into quintiles, with Q1 representing the lowest values and Q5 the highest values. The specific values for each quintile are included in parentheses. For example, the ARR for Q1 for percent homeowners is the ARR for counties where the percentage of residents who are homeowners is in the lowest 20% of all counties (i.e., counties where between 0 and 66% are homeowners). Figure S2 presents the number of total and officially reported COVID-19 deaths for each category.

to Alzheimer's disease and related dementias (*81*), and other public health emergencies like extreme heat–related deaths (*20*, *82*). While machine learning approaches may have utility for improving estimates using existing data, these tools should not come at the expense of larger investments in death investigation system reform, efforts to study and address bias in death investigation systems, and programs to build death investigation infrastructure globally (*1*, *10*, *11*, *77*).

MATERIALS AND METHODS

Study design

Our machine learning approach consisted of four general methodological steps:

1) We tuned and selected a model trained on inpatient COVID-19 deaths using nested cross-validation.

2) We used this model to predict COVID-19 deaths in out-of-hospital settings (all other settings of death).

3) We evaluated differences in unrecognized COVID-19 deaths across sociodemographic populations. We estimated (i) the predicted number of unrecognized COVID-19 in out-of-hospital settings and (ii) the ARR.

4) We conducted additional analyses for model fairness and model misclassification.

We describe these steps below and provide additional detail in the Supplementary Text. This study did not involve human participants and was deemed exempt from review by the Stanford University Institutional Review Board.

Assumptions underlying our approach

Our approach to identifying unrecognized COVID-19 deaths involved training a machine learning algorithm on inpatient COVID-19 deaths to predict COVID-19 deaths in other settings (*60*, *61*). Two core assumptions underlie this approach. First, the approach assumes that there was no measurement error in COVID-19 cause-of-death assignment in inpatient hospitals. This assumption is supported by the

widespread (and often mandatory and universal) COVID-19 testing of patients in these settings in 2020 and 2021 (*53*–*56*), evidence that excess natural-cause deaths have equaled reported COVID-19 deaths in hospital settings (*34*), and research showing that clinicians in hospital settings adhered to national reporting guidance more closely than other death certifiers (*2*, *14*, *19*, *83*). While it is possible that this assumption was partially violated in the early months of the pandemic, less than one-third of COVID-19 deaths occurred out of hospital during this period (*84*), and potential bias would result in an underestimation of unrecognized COVID-19 deaths, not an overestimation. Second, our approach assumes that the model is transportable from inpatient hospital to out-of-hospital settings. This includes assuming that predictors of COVID-19 deaths (i) were recorded with the same level of accuracy (i.e., that sociodemographic variables were either reported accurately or that any misreporting occurred similarly across settings of death), (ii) follow the same distribution, and (iii) have the same strengths of association in inpatient hospital versus out-of-hospital settings. This assumption is empirically untestable and represents the largest source of potential bias in our approach. However, it gains credence from the specific predictors that carry the most weight in our models, whose accuracy and predictive power are unlikely to vary substantially by place of death. We also tested the effect of differences in the distribution of predictors between inpatient hospital and out-of-hospital settings on the performance of our machine learning model. We provide a more detailed discussion of these assumptions in the "Rationale for our approach" section of the Supplementary Text.

Data

We used restricted-use individual-level multiple cause-of-death files from the National Center for Health Statistics to predict total COVID-19 mortality and estimate unrecognized COVID-19 deaths. There were 6.86 million recorded US deaths in 2020 and 2021. Our sample included adults 25 years and older dying from natural causes from March 2020 through December 2021. Deaths from external causes were excluded because we hypothesized that deaths from SARS-CoV-2 infection would be unlikely to be attributed to external causes of death (e.g., drug poisoning and homicide). There were 1,878,456 death certificates from adults 25 years or older dying from natural causes in inpatient hospital settings. This was the "gold standard training data" where we assumed no measurement error in COVID-19 deaths and which we used to build our model. There were 3,850,161 death certificates from adults 25 years or older dying from natural causes across all other settings of death. This was the "out-of-hospital prediction data" where we sought to predict whether deaths were likely to be COVID-19 related. We defined officially reported COVID-19 deaths as death certificates with ICD-10 code U07.1 listed as an underlying or contributing cause of death, consistent with the CDC definition (*85*).

For our machine learning model, we used information on causes of death (i.e., underlying and up to 20 contributing), decedent characteristics (i.e., age, sex, race, ethnicity, educational attainment, marital status, and smoking status as recorded on the death certificate, which has important limitations given that it is typically sourced from other records or surviving relatives as described in the Supplementary Text), temporal information (i.e., pandemic month of death and seasonality), geography (i.e., county and state of death and residence), and place of death. The eight places of death were hospital inpatient, hospital outpatient or emergency department, decedent's home, nursing home or long-term care facility, hospice, dead on arrival to a medical facility, other, or unknown (*86*). We also linked decedents to several county of residence characteristics: (i) rurality, (ii) percentage who were homeowners, (iii) median household income, (iv) percentage who completed some college, (v) percentage with poor or fair health, and (vi) percentage with diabetes. We hypothesized that these factors may have influenced reporting accuracy by affecting access to COVID-19 testing, resources available to death investigators (i.e., ability to undertake postmortem testing), and case complexity (i.e., likelihood of misassignment to comorbidities) (*2*, *17*, *87*–*89*). We also linked decedents to two time-varying county factors: (i) monthly vaccination rates and (ii) community COVID-19 transmission levels.

Model selection and tuning

We compared four modeling approaches and four covariate sets for a total of 16 candidate models. We considered logistic regressions with elastic net regularization as our reference model, along with three tree-based alternatives that could offer potential improvements. These included random forests (*90*), extreme gradient boosting trees (XGBoost) (*91*), and light gradient boosting machine models (LightGBM) (*92*). The four covariate sets we considered were as follows: a "naïve" covariate set with all predictors along with county and contributing cause of death fixed effects, a "county fixed effects only" covariate set with all predictors along with county fixed effects, a "contributing causes fixed effects only" covariate set with all predictors along with contributing cause of death fixed effects, and a "no fixed effects" covariate set with all predictors and no fixed effects.

We assessed the performance of each of the 16 candidate models in classifying COVID-19 deaths within the pool of inpatient deaths (for which we assumed that there was no measurement error for officially reported COVID-19 deaths) using a nested cross-validation approach with a 60-20-20 split. Specifically, we used 60% of the inpatient deaths for model fitting, 20% for validation and hyperparameter tuning (i.e., Bayesian optimization as detailed in the Supplementary Text), and 20% for testing out-of-sample model performance. We executed this procedure a total of five times across mutually exclusive 20% test sets and calculated the out-of-sample performance as the average across the five test sets.

We selected the XGBoost model with the "no fixed effects" covariate set as our primary prediction model on the basis of it having the highest area under the receiver operator curve (ROC AUC: 0.90; sensitivity: 0.79; specificity: 0.85). We also selected LightGBM with the "no fixed effects" covariate set as a sensitivity analysis on the basis of its high performance.

Prediction of unrecognized COVID-19 deaths

We refit the selected primary model using all inpatient data and then used the model to predict COVID-19 deaths in out-of-hospital settings (all other settings of death). We estimated the absolute number of predicted unrecognized COVID-19 deaths in a population of interest by comparing the number of predicted COVID-19 deaths in out-of-hospital settings for a population to the officially reported number of COVID-19 deaths in out-of-hospital settings for that population. We estimated model-based uncertainty by using 5000 bootstrapped samples of the out-of-hospital predictions and reporting the 2.5th and 97.5th quantiles. Our predictions are based on the median of 5000 bootstrapped samples.

We also calculated the ARR, which was quantified as total predicted COVID-19 deaths (the sum of official in-hospital COVID-19 deaths and predicted out-of-hospital COVID-19 deaths) divided by

official COVID-19 deaths across all settings of death. The ARR allows for a comparison of potential underreporting of COVID-19 deaths across all settings of death and accounts for differences in the percentage of deaths in out-of-hospital settings across populations.

An ARR > 1 indicates that our machine learning approach predicted more COVID-19 deaths in a population than the number of officially reported COVID-19 deaths. This suggests that COVID-19 deaths were underreported in official COVID-19 death counts for that population. An ARR < 1 indicates that COVID-19 deaths were overreported relative to the machine learning predictions. Populations and geographic areas with an ARR < 1 were infrequent. One reason that a population or geographic area could have an ARR < 1 is if death certifiers recorded people as dying from COVID-19 when they had COVID-19 but actually died from another unrelated cause. Especially early in the pandemic, states and death certifiers had different protocols that influenced how conservative they were with assigning COVID-19 deaths. For example, while most states defined COVID-19 deaths as deaths occurring within 30 days of a COVID-19 diagnosis, Massachusetts included deaths occurring within 60 days until March 2022 (*26*).

Model fairness and misclassification

We assessed model fairness across race and ethnicity, sex, educational attainment, the rural-urban continuum, and states by calculating the ROC AUC parity, false negative rate parity, and false positive rate parity in the inpatient data. Specifically, we used fivefold cross-validation. This involved splitting inpatient data into five subsets and using four of the subsets to predict the fifth subset. This procedure was conducted five times, and results were averaged. To confirm that the fairness metrics we selected for this study were not masking absolute differences in predictive performance, we also constructed subgroup-specific precision-recall curves for racial groups.

Given that we do not know the true COVID-19 status for any subset of out-of-hospital deaths, it was not possible to estimate model misclassification metrics for the model's performance among out-of-hospital deaths. Instead, we estimated the sensitivity of our results to model misclassification on the basis of model performance for inpatient deaths. We adapted a prior approach (*93*) that involved estimating the false discovery rate and false omission rate among the inpatient deaths and applying these metrics to the out-of-hospital predictions by drawing from a binomial distribution and switching the outcome. Further details about this sensitivity analysis for model misclassification are provided in the Supplementary Text.

Supplementary Materials

This PDF file includes:

Supplementary Text
Figs. S1 to S10
Tables S1 to S12
References

REFERENCES

1. F. B. Ahmad, R. N. Anderson, K. Knight, L. M. Rossen, P. D. Sutton, Advancements in the National Vital Statistics System to meet the real-time data needs of a pandemic. *Am. J. Public Health* **111**, 2133–2140 (2021).
2. M. Lipsitch, M. T. Bassett, J. S. Brownstein, P. Elliott, D. Eyre, M. Kate Grabowski, J. A. Hay, M. Johansson, S. M. Kissler, D. B. Larremore, J. Layden, J. Lessler, R. Lynfield, D. MacCannell, L. C. Madoff, C. J. E. Metcalf, L. A. Meyers, S. K. Ofori, C. Quinn, A. I. Ramos Bento, N. Reich, S. Riley, R. Rosenfeld, M. H. Samore, R. Sampath, R. B. Slayton, D. L. Swerdlow, S. Truelove, J. K. Varma, Y. H. Grad, Infectious disease surveillance needs for the United States: Lessons from COVID-19. arXiv:2311.13724 (2023).
3. A. Bovell-Ammon, N. C. McCann, M. Mulugeta, S. Ettinger de Cuba, J. Raifman, P. Shafer, Association of the expiration of child tax credit advance payments with food insufficiency in US households. *JAMA Netw. Open* **5**, e2234438 (2022).
4. A. Skinner, K. Flannery, K. Nocka, J. Bor, L. T. Dean, J. Jay, S. K. Lipson, M. B. Cole, E. A. Benfer, R. Scheckman, W. Raderman, D. K. Jones, J. Raifman, A database of US state policies to mitigate COVID-19 and its economic consequences. *BMC Public Health* **22**, 1124 (2022).
5. M. G. T. Denney, R. G. Valdez, Compounding racialized vulnerability: COVID-19 in prisons, jails, and migrant detention centers. *J. Health Polit. Policy Law* **46**, 861–887 (2021).
6. R. Katz, K. Toole, H. Robertson, A. Case, J. Kerr, S. Robinson-Marshall, J. Schermerhorn, S. Orsborn, M. Van Maele, R. Zimmerman, T. Stevens, COVID AMP Coding Team, A. Phelan, C. Carlson, E. Graeden, Open data for COVID-19 policy analysis and mapping. *Sci. Data* **10**, 491 (2023).
7. E. Gutierrez, A. Rubli, T. Tavares, Information and behavioral responses during a pandemic: Evidence from delays in COVID-19 death reports (2021). https://doi.org/10.2139/ssrn.3645317.
8. A. C. Stokes, D. J. Lundberg, J. Bor, K. Bibbins-Domingo, Excess deaths during the COVID-19 pandemic: Implications for US death investigation systems. *Am. J. Public Health* **111**, S53–S54 (2021).
9. L. McGivern, L. Shulman, J. K. Carney, S. Shapiro, E. Bundock, Death certification errors and the effect on mortality statistics. *Public Health Rep.* **132**, 669–675 (2017).
10. National Academies of Sciences, Engineering, and Medicine, *Strengthening Forensic Science in the United States: A Path Forward* (The National Academies Press, 2009).
11. Institute of Medicine, *Medicolegal Death Investigation System: Workshop Summary* (The National Academies Press, 2003).
12. A. Cakanlar, R. Trudel, K. White, Political ideology and the perceived impact of coronavirus prevention behaviors for the self and others. *J. Assoc. Consum. Res.* **7**, 36–44 (2022).
13. S. L. Choi, P. Martin, J. Cho, Y. J. Ryou, M. Heinz, Personality and compliance with COVID-19 protective measures among older Americans: Moderating effects of age, gender, and race/ethnicity. *Pers. Individ. Dif.* **189**, 111499 (2022).
14. D. Bergin, R. Keller, "Uncounted: Inaccurate death certificates across the country hide the true toll of COVID-19," *Missouri Independent*, 22 December 2021. https://missouriindependent.com/2021/12/22/uncounted-inaccurate-death-certificates-across-the-country-hide-the-true-toll-of-covid-19/.
15. J. M. Feldman, S. Gruskin, B. A. Coull, N. Krieger, Quantifying underreporting of law-enforcement-related deaths in United States vital statistics and news-media-based data sources: A capture-recapture analysis. *PLOS Med.* **14**, e1002399 (2017).
16. M. J. Prados, T. Baker, A. N. Beck, D. B. Burghart, R. R. Johnson, D. Klinger, K. Thomas, B. K. Finch, Do sheriff-coroners underreport officer-involved homicides? *Acad. Forensic Pathol.* **12**, 140–148 (2022).
17. M. Kanyangarara, V. Daguise, L. Gual-Gonzalez, A. H. Litwin, J. Korte, C. Ross, M. S. Nolan, SC STRONG Project Team, COVID-19 testing practices, preventive behaviors, and factors associated with test positivity: Population-based statewide survey study. *JMIR Public Health Surveill.* **9**, e34579 (2023).
18. D. Dada, J. N. Djiometio, S. M. McFadden, J. Demeke, D. Vlahov, L. Wilton, M. Wang, L. E. Nelson, Strategies that promote equity in COVID-19 vaccine uptake for Black communities: A review. *J. Urban Health* **99**, 15–27 (2022).
19. A. C. Stokes, D. J. Lundberg, J. Bor, I. T. Elo, K. Hempstead, S. H. Preston, Association of health care factors with excess deaths not assigned to COVID-19 in the US. *JAMA Netw. Open* **4**, e2125287 (2021).
20. A. B. Le, E. G. Brooks, L. A. McNulty, J. R. Gill, J. J. Herstein, J. Rios, S. J. Patlovich, K. C. Jelden, K. K. Schmid, J. J. Lowe, S. G. Gibbs, U.S. medical examiner/coroner capability to handle highly infectious decedents. *Forensic Sci. Med. Pathol.* **15**, 31–40 (2019).
21. L. Ruiz, B. M. Posey, M.-A. Neuilly, M. K. Stohr, C. Hemmens, Certifying death in the United States. *J. Forensic Sci.* **63**, 1138–1145 (2018).
22. D. J. Lundberg, J. A. Chen, Structural ableism in public health and healthcare: A definition and conceptual framework. *Lancet Reg. Health Am.* **30**, 100650 (2024).
23. M. A. Garcia, P. A. Homan, C. García, T. H. Brown, The color of COVID-19: Structural racism and the disproportionate impact of the pandemic on older Black and Latinx adults. *J. Gerontol. B Psychol. Sci. Soc. Sci.* **76**, e75–e80 (2021).
24. L. E. Egede, R. J. Walker, Structural racism, social risk factors, and Covid-19—A dangerous convergence for Black Americans. *N. Engl. J. Med.* **383**, e77 (2020).
25. E. E. Andrews, K. B. Ayers, K. S. Brown, D. S. Dunn, C. R. Pilarski, No body is expendable: Medical rationing and disability justice during the COVID-19 pandemic. *Am. Psychol.* **76**, 451–461 (2021).
26. L. V. Estrada, J. L. Levasseur, A. Maxim, G. A. Benavidez, K. M. Pollack Porter, Structural racism, place, and COVID-19: A narrative review describing how we prepare for an endemic COVID-19 future. *Health Equity* **6**, 356–366 (2022).

27. E. Paglino, A. C. Stokes, Excess mortality as a tool to monitor the evolution of health emergencies: Choices, challenges, and future directions. *Am. J. Public Health* **114**, 583–586 (2024).
28. D. M. Weinberger, J. Chen, T. Cohen, F. W. Crawford, F. Mostashari, D. Olson, V. E. Pitzer, N. G. Reich, M. Russi, L. Simonsen, A. Watkins, C. Viboud, Estimation of excess deaths associated with the COVID-19 pandemic in the United States, March to May 2020. *JAMA Intern. Med.* **180**, 1336–1344 (2020).
29. S. H. Woolf, D. A. Chapman, R. T. Sabo, E. B. Zimmerman, Excess deaths from COVID-19 and other causes in the US, March 1, 2020, to January 2, 2021. *JAMA* **325**, 1786–1789 (2021).
30. C. A. Ackley, D. J. Lundberg, L. Ma, I. T. Elo, S. H. Preston, A. C. Stokes, County-level estimates of excess mortality associated with COVID-19 in the United States. *SSM Popul. Health* **17**, 101021 (2022).
31. A. N. Luck, A. C. Stokes, K. Hempstead, E. Paglino, S. H. Preston, Associations between mortality from COVID-19 and other causes: A state-level analysis. *PLOS ONE* **18**, e0281683 (2023).
32. J. S. Faust, C. Du, K. D. Mayes, S.-X. Li, Z. Lin, M. L. Barnett, H. M. Krumholz, Mortality from drug overdoses, homicides, unintentional injuries, motor vehicle crashes, and suicides during the pandemic, March–August 2020. *JAMA* **326**, 84–86 (2021).
33. E. Paglino, D. J. Lundberg, E. Wrigley-Field, Z. Zhou, J. A. Wasserman, R. Raquib, Y.-H. Chen, K. Hempstead, S. H. Preston, I. T. Elo, M. M. Glymour, A. C. Stokes, Excess natural-cause mortality in US counties and its association with reported COVID-19 deaths. *Proc. Natl. Acad. Sci. U.S.A.* **121**, e2313661121 (2024).
34. Y.-H. Chen, A. C. Stokes, H. E. Aschmann, R. Chen, S. DeVost, M. V. Kiang, S. Koliwad, A. R. Riley, M. M. Glymour, K. Bibbins-Domingo, Excess natural-cause deaths in California by cause and setting: March 2020 through February 2021. *PNAS Nexus* **1**, pgac079 (2022).
35. T. B. Foster, L. Fernandez, S. R. Porter, N. Pharris-Ciurej, Racial and ethnic disparities in excess all-cause mortality in the first year of the COVID-19 pandemic. *Demography* **61**, 59–85 (2024).
36. C. Caraballo, D. S. Massey, C. D. Ndumele, T. Haywood, S. Kaleem, T. King, Y. Liu, Y. Lu, M. Nunez-Smith, H. A. Taylor, K. E. Watson, J. Herrin, C. W. Yancy, J. S. Faust, H. M. Krumholz, Excess mortality and years of potential life lost among the Black population in the US, 1999–2020. *JAMA* **329**, 1662–1670 (2023).
37. W.-E. Lee, S. Woo Park, D. M. Weinberger, D. Olson, L. Simonsen, B. T. Grenfell, C. Viboud, Direct and indirect mortality impacts of the COVID-19 pandemic in the United States, March 1, 2020 to January 1, 2022. *eLife* **12**, e77562 (2023).
38. E. P. Vamos, K. Khunti, Indirect effects of the COVID-19 pandemic on people with type 2 diabetes: Time to urgently move into a recovery phase. *BMJ Qual. Saf.* **31**, 483–485 (2022).
39. A. B. Friedman, D. Barfield, G. David, T. Diller, C. Gunnarson, M. Liu, B. V. Vicidomina, R. Zhang, Y. Zhang, S. C. Nigam, Delayed emergencies: The composition and magnitude of non-respiratory emergency department visits during the COVID-19 pandemic. *J Am Coll. Emerg. Physicians Open* **2**, e12349 (2021).
40. I. Akobirshoev, M. Vetter, L. I. Iezzoni, S. R. Rao, M. Mitra, Delayed medical care and unmet care needs due to the COVID-19 pandemic among adults with disabilities in the US. *Health Aff.* **41**, 1505–1512 (2022).
41. C. J. Ruhm, The evolution of excess deaths in the United States during the first 2 years of the COVID-19 pandemic. *Am. J. Epidemiol.* **192**, 1949–1959 (2023).
42. S. Magesh, D. John, W. T. Li, Y. Li, A. Mattingly-App, S. Jain, E. Y. Chang, W. M. Ongkeko, Disparities in COVID-19 outcomes by race, ethnicity, and socioeconomic status: A systematic-review and meta-analysis. *JAMA Netw. Open* **4**, e2134147 (2021).
43. A. Khanijahani, S. Iezadi, K. Gholipour, S. Azami-Aghdash, D. Naghibi, A systematic review of racial/ethnic and socioeconomic disparities in COVID-19. *Int. J. Equity Health* **20**, 248 (2021).
44. D. J. Lundberg, E. Wrigley-Field, A. Cho, R. Raquib, E. O. Nsoesie, E. Paglino, R. Chen, M. V. Kiang, A. R. Riley, Y.-H. Chen, M.-L. Charpignon, K. Hempstead, S. H. Preston, I. T. Elo, M. M. Glymour, A. C. Stokes, COVID-19 mortality by race and ethnicity in US metropolitan and nonmetropolitan areas, March 2020 to February 2022. *JAMA Netw. Open* **6**, e2311098 (2023).
45. E. Paglino, D. J. Lundberg, Z. Zhou, J. A. Wasserman, R. Raquib, A. N. Luck, K. Hempstead, J. Bor, S. H. Preston, I. T. Elo, A. C. Stokes, Monthly excess mortality across counties in the United States during the COVID-19 pandemic, March 2020 to February 2022. *Sci. Adv.* **9**, eadf9742 (2023).
46. Y.-H. Chen, A. R. Riley, K. A. Duchowny, H. E. Aschmann, R. Chen, M. V. Kiang, A. C. Mooney, A. C. Stokes, M. M. Glymour, K. Bibbins-Domingo, COVID-19 mortality and excess mortality among working-age residents in California, USA, by occupational sector: A longitudinal cohort analysis of mortality surveillance data. *Lancet Public Health* **7**, e744–e753 (2022).
47. Y.-H. Chen, E. C. Matthay, R. Chen, M. A. DeVost, K. A. Duchowny, A. R. Riley, K. Bibbins-Domingo, M. Maria Glymour, Excess mortality in California by education during the COVID-19 pandemic. *Am. J. Prev. Med.* **63**, 827–836 (2022).
48. Y.-H. Chen, M. M. Glymour, R. Catalano, A. Fernandez, T. Nguyen, M. Kushel, K. Bibbins-Domingo, Excess mortality in california during the coronavirus disease 2019 pandemic, March to August 2020. *JAMA Intern. Med.* **181**, 705–707 (2020).
49. J. T. Chen, N. Krieger, 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.* **27**, S43–S56 (2021).
50. M. T. Bassett, J. T. Chen, N. Krieger, Variation in racial/ethnic disparities in COVID-19 mortality by age in the United States: A cross-sectional study. *PLOS Med.* **17**, e1003402 (2020).
51. A. C. Stokes, D. J. Lundberg, I. T. Elo, K. Hempstead, J. Bor, S. H. Preston, COVID-19 and excess mortality in the United States: A county-level analysis. *PLOS Med.* **18**, e1003571 (2021).
52. E. Wrigley-Field, J. P. Leider, M. Barbieri, Was COVID-19 mortality differentially reported by race? An analysis leveraging place, timing, and cause of death in Minnesota, United States, OSF [Preprint] (2025). https://doi.org/10.17605/OSF.IO/J6RC9.
53. T. R. Talbot, M. K. Hayden, D. S. Yokoe, A. N. Malani, H. A. Amer, I. C. Kalu, L. K. Logan, R. W. Moehring, S. Munoz-Price, T. N. Palmore, D. J. Weber, S. B. Wright, SHEA Board of Trustees, Asymptomatic screening for severe acute respiratory coronavirus virus 2 (SARS-CoV-2) as an infection prevention measure in healthcare facilities: Challenges and considerations. *Infect. Control Hosp. Epidemiol.* **44**, 2–7 (2023).
54. K. B. Brust, T. Kobayashi, D. J. Diekema, Asymptomatic testing of hospital admissions for SARS-CoV-2: Is it OK to stop? *Clin. Infect. Dis.* **78**, 356–360 (2023).
55. R. B. Sangal, D. R. Peaper, C. Rothenberg, M. L. Landry, L. S. Sussman, R. A. Martinello, A. Ulrich, A. K. Venkatesh, Universal SARS-CoV-2 testing of emergency department admissions increases emergency department length of stay. *Ann. Emerg. Med.* **79**, 182–186 (2022).
56. S. R. Sastry, R. Pryor, J. E. Raybould, J. Reznicek, K. Cooper, A. Patrick, S. Knowlson, P. Bailey, E. Godbout, M. Doll, M. P. Stevens, G. Bearman, Universal screening for the SARS-CoV-2 virus on hospital admission in an area with low COVID-19 prevalence. *Infect. Control Hosp. Epidemiol.* **41**, 1231–1233 (2020).
57. S. Idicula-Thomas, U. Gawde, P. Jha, Comparison of machine learning algorithms applied to symptoms to determine infectious causes of death in children: National survey of 18,000 verbal autopsies in the Million Death Study in India. *BMC Public Health* **21**, 1787 (2021).
58. J. Kleinberg, J. Ludwig, S. Mullainathan, Z. Obermeyer, Prediction policy problems. *Am. Econ. Rev.* **105**, 491–495 (2015).
59. C. L. Ramspek, E. W. Steyerberg, R. D. Riley, F. R. Rosendaal, O. M. Dekkers, F. W. Dekker, M. van Diepen, Prediction or causality? A scoping review of their conflation within current observational research. *Eur. J. Epidemiol.* **36**, 889–898 (2021).
60. L. Einav, A. Finkelstein, S. Mullainathan, Z. Obermeyer, Predictive modeling of U.S. health care spending in late life. *Science* **360**, 1462–1465 (2018).
61. A. N. Angelopoulos, S. Bates, C. Fannjiang, M. I. Jordan, T. Zrnic, Prediction-powered inference. *Science* **382**, 669–674 (2023).
62. National Center for Health Statistics, Excess deaths associated with COVID-19: Provisional death counts for coronavirus disease (COVID-19) (Center for Disease Control and Prevention, 2023). www.cdc.gov/nchs/nvss/vsrr/covid19/excess_deaths.htm.
63. S. Gutierrez, M. M. Glymour, G. D. Smith, Evidence triangulation in health research. *Eur. J. Epidemiol.* **40**, 743–757 (2025).
64. S. Goggins, Contesting public forgetting: Memory and policy learning in the era of Covid-19. *Mem. Stud.* **17**, 1241–1258 (2024).
65. D. A. Weaver, The mortality experience of disabled persons in the United States during the COVID-19 pandemic. *Health Aff. Sch.* **2**, qxad082 (2024).
66. T. Keith, "As U.S. COVID deaths near 1 million, advocates press for a memorial day," *NPR*, 9 March 2022. www.npr.org/2022/03/09/1085243575/marked-by-covid-memorial-day-march.
67. D. Blumenthal, E. J. Fowler, M. Abrams, S. R. Collins, Covid-19 - implications for the health care system. *N. Engl. J. Med.* **383**, 1483–1488 (2020).
68. J. Melinek, L. C. Thomas, W. R. Oliver, G. A. Schmunk, V. W. Weedn, The National Association of Medical Examiners Ad Hoc Committee on Medical Examiner Independence, National Association of Medical Examiners position paper: Medical examiner, coroner, and forensic pathologist independence. *Acad. Forensic Pathol.* **3**, 93–98 (2013). https://journals.sagepub.com/doi/10.23907/2013.013.
69. National Research Council, *Strengthening Forensic Science in the United States: A Path Forward* (The National Academies Press, 2009).
70. J. R. Gill, M. E. DeJoseph, The importance of proper death certification during the COVID-19 pandemic. *JAMA* **324**, 27 (2020).
71. J. D. Stevens, S. D. Landes, Assessing state level variation in signature authority and cause of death accuracy, 2005–2017. *Prev. Med. Rep.* **21**, 101309 (2021).
72. V. A. Earnshaw, N. M. Brousseau, E. C. Hill, S. C. Kalichman, L. A. Eaton, A. B. Fox, Anticipated stigma, stereotypes, and COVID-19 testing. *Stig. Health* **5**, 390–393 (2020).

73. L. G. Conway III, S. R. Woodard, A. Zubrod, L. Chan, Why are conservatives less concerned about the coronavirus (COVID-19) than liberals? Comparing political, experiential, and partisan messaging explanations. *Pers. Individ. Dif.* **183**, 111124 (2021).
74. Y. Sun, S. M. Monnat, Rural-urban and within-rural differences in COVID-19 vaccination rates. *J. Rural Health* **38**, 916–922 (2022).
75. B. Neelon, F. Mutiso, N. T. Mueller, J. L. Pearce, S. E. Benjamin-Neelon, Associations between governor political affiliation and COVID-19 cases, deaths, and testing in the U.S. *Am. J. Prev. Med.* **61**, 115–119 (2021).
76. J. M. Turan, K. B. Crockett, Y.-I. Kim, M. Ryan, R. Ramachandran, L. B. Bateman, R. W. Durant, G. R. Oates, Adaption and validation of scales to measure COVID-19 stigma attitudes and perceptions. *Stigma Health* **10**, 577–585 (2024).
77. National Academies of Sciences, Engineering, and Medicine, Division of Behavioral and Social Sciences and Education, Committee on National Statistics, Committee on Population, Committee on Rising Midlife Mortality Rates and Socioeconomic Disparities, *High and Rising Mortality Rates Among Working-Age Adults* (National Academies Press, 2021).
78. Vital Strategies, *Lessons from Rapid Mortality Surveillance: Findings of a Multi-country Process Evaluation* (Vital Strategies and Resolve to Save Lives, 2023); www.vitalstrategies.org/resources/lessons-from-rapid-mortality-surveillance/.
79. R. M. Entress, J. Tyler, A.-A. Sadiq, Inequity after death: Exploring the equitable utilization of FEMA's COVID-19 funeral assistance funds. *Public Adm. Rev.*, 10.1111/puar.13572 (2022).
80. A. Denham, T. Vasu, P. Avendano, A. Boslett, M. Mendoza, E. L. Hill, Coroner county systems are associated with a higher likelihood of unclassified drug overdoses compared to medical examiner county systems. *Am. J. Drug Alcohol Abuse* **48**, 606–617 (2022).
81. A. C. Stokes, J. Weiss, D. J. Lundberg, W. Xie, J. K. Kim, S. H. Preston, E. M. Crimmins, Estimates of the association of dementia with US mortality levels using linked survey and mortality records. *JAMA Neurol.* **77**, 1543–1550 (2020).
82. S. A. M. Khatana, R. M. Werner, P. W. Groeneveld, Association of extreme heat with all-cause mortality in the contiguous US, 2008–2017. *JAMA Netw. Open* **5**, e2212957 (2022).
83. B. Teasdale, A. Narayan, S. Harman, K. A. Schulman, Place of death before and during the COVID-19 pandemic. *JAMA Netw. Open* **7**, e2350821 (2024).
84. S. Chuzi, R. Molsberry, M. E. McCabe, C. W. Yancy, A. Ogunseitan, N. B. Allen, S. S. Khan, Distribution in place of death for COVID-19-related mortality in the United States. *J. Am. Geriatr. Soc.* **68**, 1917–1918 (2020).
85. National Center for Health Statistics, Provisional Death Counts for Coronavirus Disease (COVID-19) Technical Notes (2022). www.cdc.gov/nchs/nvss/vsrr/covid19/tech_notes.htm.
86. F. B. Ahmad, J. A. Cisewski, J. Xu, R. N. Anderson, COVID-19 mortality update—United States, 2022. *MMWR Morb. Mortal. Wkly Rep.* **72**, 493–496 (2023).
87. J. M. Souch, J. S. Cossman, A commentary on rural-urban disparities in COVID-19 testing rates per 100,000 and risk factors. *J. Rural Health* **37**, 188–190 (2021).
88. S. H. Ali, Y. Tozan, A. M. Jones, J. Foreman, A. Capasso, R. J. DiClemente, Regional and socioeconomic predictors of perceived ability to access coronavirus testing in the United States: Results from a nationwide online COVID-19 survey. *Ann. Epidemiol.* **58**, 7–14 (2021).
89. S. Elezkurtaj, S. Greuel, J. Ihlow, E. G. Michaelis, P. Bischoff, C. A. Kunze, B. V. Sinn, M. Gerhold, K. Hauptmann, B. Ingold-Heppner, F. Miller, H. Herbst, V. M. Corman, H. Martin, H. Radbruch, F. L. Heppner, D. Horst, Causes of death and comorbidities in hospitalized patients with COVID-19. *Sci. Rep.* **11**, 4263 (2021).
90. L. Breiman, Random forests. *Mach. Learn.* **45**, 5–32 (2001).
91. T. Chen, C. Guestrin, XGBoost: A scalable tree boosting system, arXiv:1603.02754 [cs.LG] (2016).
92. G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, LightGBM: A highly efficient gradient boosting decision tree. *Neural Inf. Process. Syst.* , 3146–3154 (2017).
93. B. Q. Huynh, E. T. Chin, M. V. Kiang, Estimated childhood lead exposure from drinking water in Chicago. *JAMA Pediatr.* **178**, 473–479 (2024).
94. "Documenting COVID-19 project and USA TODAY Network, Uncounted: Inaccurate death certificates across the country hide the true toll of COVID-19," *USA Today*, 22 December 2021. www.usatoday.com/in-depth/news/nation/2021/12/22/covid-deaths-obscured-inaccurate-death-certificates/8899157002/.
95. E. Arias, M. Heron, National Center for Health Statistics, J. Hakes, US Census Bureau, The validity of race and Hispanic-origin reporting on death certificates in the United States: An update. *Vital Health Stat.* **2**, 1–21 (2016).
96. E. Arias, W. S. Schauman, K. Eschbach, P. D. Sorlie, E. Backlund, The validity of race and Hispanic origin reporting on death certificates in the United States. *Vital Health Stat.* **2**, 1–23 (2008).
97. A. P. Haas, A. D. Lane, J. R. Blosnich, B. A. Butcher, M. G. Mortali, Collecting sexual orientation and gender identity information at death. *Am. J. Public Health* **109**, 255–259 (2019).
98. J. K. Walters, M. C. Mew, K. K. Repp, Transgender and nonbinary deaths investigated by the state medical examiner in the Portland, Oregon, Metro Area and their concordance with vital records, 2011–2021. *J. Public Health Manag. Pract.* **29**, 64–70 (2023).
99. B. L. Rostron, J. L. Boies, E. Arias, Education reporting and classification on death certificates in the United States. *Vital Health Stat.* **2**, 1–21 (2010).
100. S. D. Landes, M. A. Turk, E. Bisesti, Uncertainty and the reporting of intellectual disability on death certificates: A cross-sectional study of US mortality data from 2005 to 2017. *BMJ Open* **11**, e045360 (2021).
101. B. Li, C. Gatsonis, I. J. Dahabreh, J. A. Steingrimsson, Estimating the area under the ROC curve when transporting a prediction model to a target population. *Biometrics* **79**, 2382–2393 (2023).
102. H. Lu, S. R. Cole, C. J. Howe, D. Westreich, Toward a clearer definition of selection bias when estimating causal effects. *Epidemiology* **33**, 699–706 (2022).
103. J. Friedman, T. Hastie, R. Tibshirani, Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1–22 (2010).
104. J. K. Tay, B. Narasimhan, T. Hastie, Elastic net regularization paths for all generalized linear models. *J. Stat. Softw.* **106**, 1 (2023).
105. P. I. Frazier, A tutorial on Bayesian optimization, arXiv:1807.02811 [stat.ML] (2018).
106. J. Fiksel, A. Datta, A. Amouzou, S. Zeger, Generalized Bayes quantification learning under dataset shift. *J. Am. Stat. Assoc.* **117**, 2163–2181 (2022).
107. G. Forman, Quantifying counts and costs via classification. *Data Min. Knowl. Discov.* **17**, 164–206 (2008).
108. M. M. Garrido, A. S. Kelley, J. Paris, K. Roza, D. E. Meier, R. S. Morrison, M. D. Aldridge, Methods for constructing and assessing propensity scores. *Health Serv. Res.* **49**, 1701–1720 (2014).

Acknowledgments

Funding: The content is solely the responsibility of the authors and does not necessarily represent the official views of the study sponsors. We acknowledge receiving financial support from the W.K. Kellogg Foundation (to A.C.S.) and the Eunice Kennedy Shriver National Institute of Child Health and Human Development (P2C HD041023 to E.W.-F.). **Author contributions:** M.V.K. and A.C.S. conceptualized the study, acquired the data, supervised the study, and acquired funding for the research. M.V.K. led the data analysis, with contributions from Z.R.L., R.V.R., and B.H. All authors contributed to interpretation of the data and writing of the manuscript. **Competing interests:** The authors declare that they have no competing interests. **Data, code, and materials availability:** All data and code needed to evaluate and reproduce the results in the paper are present in the paper and/or the Supplementary Materials. There are no new materials generated in the study. Replication data and code are available at https://zenodo.org/records/18612091. (DOI: 10.5281/zenodo.18612091) and https://github.com/mkiang/unclassified_covid_deaths. Because of a data use agreement with the CDC, individual-level data cannot be shared; however, researchers can request the individual-level data from the National Center for Health Statistics at www.cdc.gov/nchs/nvss/nvss-restricted-data.htm.

Submitted 18 January 2026
Accepted 13 February 2026
Published 18 March 2026
10.1126/sciadv.aef5697

ScienceAdvances

Applying machine learning to identify unrecognized COVID-19 deaths recorded as other causes of death in the United States

Mathew V. Kiang, Zehang Richard Li, Elizabeth Wrigley-Field, Rafeya V. Raquib, Dielle J. Lundberg, Eugenio Paglino, Benjamin Huynh, Kirsten Bibbins-Domingo, M. Maria Glymour, and Andrew C. Stokes

Sci. Adv. **12** (12), eaef5697. DOI: 10.1126/sciadv.aef5697

View the article online
https://www.science.org/doi/10.1126/sciadv.aef5697
Permissions
https://www.science.org/help/reprints-and-permissions

Use of this article is subject to the Terms of service

Science Advances (ISSN 2375-2548) is published by the American Association for the Advancement of Science. 1200 New York Avenue NW, Washington, DC 20005. The title *Science Advances* is a registered trademark of AAAS.

Copyright © 2026 The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original U.S. Government Works. Distributed under a Creative Commons Attribution NonCommercial License 4.0 (CC BY-NC).