



# Individual and community-level risk for COVID-19 mortality in the United States

Jin Jin<sup>1,6</sup>, Neha Agarwala<sup>1,6</sup>, Prosenjit Kundu<sup>1,6</sup>, Benjamin Harvey<sup>1</sup>, Yuqi Zhang<sup>1,6</sup>, Eliza Wallace<sup>4</sup> and Nilanjan Chatterjee<sup>1,5</sup>✉

**Reducing COVID-19 burden for populations will require equitable and effective risk-based allocations of scarce preventive resources, including vaccinations<sup>1</sup>. To aid in this effort, we developed a general population risk calculator for COVID-19 mortality based on various sociodemographic factors and pre-existing conditions for the US population, combining information from the UK-based OpenSAFELY study with mortality rates by age and ethnicity across US states. We tailored the tool to produce absolute risk estimates in future time frames by incorporating information on pandemic dynamics at the community level. We applied the model to data on risk factor distribution from a variety of sources to project risk for the general adult population across 477 US cities and for the Medicare population aged 65 years and older across 3,113 US counties, respectively. Validation analyses using 54,444 deaths from 7 June to 1 October 2020 show that the model is well calibrated for the US population. Projections show that the model can identify relatively small fractions of the population (for example 4.3%) that might experience a disproportionately large number of deaths (for example 48.7%), but there is wide variation in risk across communities. We provide a web-based risk calculator and interactive maps for viewing community-level risks.**

The first case of SARS-CoV-2 infection in the United States was reported on 20 January 2020 in the state of Washington<sup>2</sup> and to date, the pandemic has led to more than 240,000 COVID-19 deaths, making the United States by far the most affected country globally. However, there is high variation in rates of infections and underlying deaths across US states, counties and cities. Local population characteristics, such as population density<sup>3</sup> and mobility patterns<sup>4</sup>, as well as mitigation measures<sup>5,6</sup>, define background risk of illness and death across the regions. Further, predisposing factors, including age, sex, ethnicity and racial background, social conditions and pre-existing conditions, put individuals within the same community at differential risk of serious illness and mortality<sup>7–15</sup>.

To date, the United States and other countries have mostly relied on community-based intervention measures, such as lockdowns, social distancing and guidance on mask wearing, for mitigating the worst effects of the pandemic. A variety of pandemic scenario models with increasing sophistication are available for forecasting future trends in infection, hospitalizations and deaths at the population level (<https://www.cdc.gov/coronavirus/2019-ncov/covid-data/forecasting-us.html>). Although a variety of predisposing factors are known, there has been limited effort to incorporate these

factors into prevention strategies and/or forecasting models. In the future, however, as the United States and other countries continue to face increasing societal and economic pressure for relaxing some of the broad intervention measures, consideration of risk associated with predisposing factors for individuals and at the population level will be important for developing more equitable strategies for prevention<sup>16–18</sup>. Promising results from early phases of a number of vaccine trials (<https://www.businesswire.com/news/home/20201109005539/en/>, <https://investors.modernatx.com/news-releases/news-release-details/modernas-covid-19-vaccine-candidate-meets-its-primary-efficacy> and <https://www.pfizer.com/news/press-release/press-release-detail/pfizer-and-biontech-conclude-phase-3-study-covid-19-vaccine>)<sup>19–22</sup> have raised the likelihood of available vaccines by the end of 2020 and a number of national and international bodies (<https://www.csis.org/analysis/advancing-research-and-planning-equitable-distribution-covid-19-vaccine>, <https://www.nationalacademies.org/news/2020/07/national-academies-launch-study-on-equitable-allocation-of-a-covid-19-vaccine-first-meeting-july-24>)<sup>23</sup> have been developing frameworks that will allow equitable distribution of vaccines, taking into account differential risk for individuals and communities.

We describe the development and validation of a COVID-19 mortality risk calculator for the US adult (aged 18 years and older) population, integrating information from a variety of datasets for the estimation of risk associated with predisposing factors. We further extend the calculator to integrate information from pandemic forecasting models so that an individual's absolute risk can be informed based not only on their underlying risk factors, but also on community-level risk due to underlying pandemic dynamics. We use the information on the prevalence and co-occurrence of risk factors from various national databases to make population-level projections of risk associated with these predisposing factors for the adult population across 477 US cities and for the 65-years-and-older population enrolled in Medicare across 3,113 US counties. We provide national-, state- and city and/or county- level estimates for the size of populations who are at or above different risk thresholds and can be gradually prioritized for vaccination and other preventive efforts. We also provide a web-based individual-level risk calculator and interactive maps for viewing population-level risk projections to facilitate future policy decisions. The main findings and limitations of the study are summarized in Table 1.

Risk of mortality associated with various age groups in the United States follows a comparable pattern to that reported by the UK OpenSAFELY study<sup>15</sup> (Extended Data Fig. 1). Relative to their

<sup>1</sup>Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, USA. <sup>2</sup>Department of Mathematics and Statistics, University of Maryland, Baltimore County, Baltimore, MD, USA. <sup>3</sup>Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA. <sup>4</sup>PolicyMap, Inc., Philadelphia, PA, USA. <sup>5</sup>Department of Oncology, School of Medicine, Johns Hopkins University, Baltimore, MD, USA.

<sup>6</sup>These authors contributed equally: Jin Jin, Neha Agarwala, Prosenjit Kundu. ✉e-mail: [nilanjan@jhu.edu](mailto:nilanjan@jhu.edu)

**Table 1 | Policy summary**

|                               |  |
|-------------------------------|--|
| Background                    | Reducing the burden of COVID-19 illness and mortality for societies requires development of risk-based strategies for prevention through shielding, early vaccination and distribution of other scarce preventive resources. Such efforts could be aided by the development of general population models for predicting COVID-19 serious outcomes, including mortality, incorporating information on both individual-level risk factors and community-level pandemic dynamics.   |
| Main findings and limitations | We develop a model for predicting individual-level risk for COVID-19 mortality in the United States by synthesizing information on risk associated with various sociodemographic factors and pre-existing health conditions from a large UK-based study <sup>15</sup> , rates associated with age and race available across US states and state-level projected death rates from pandemic forecasting models. Projections of risk across US cities and counties show wide variation in the number and percentage of high-risk individuals across US communities and validation analyses using mortality data show that projected risk is well calibrated. Projections also show that the underlying individual-level model is expected to have high discriminatory power for the US population and can effectively identify small groups of populations that will lead to a large majority of deaths. Limitations of the study include lack of individual-level population cohort data to train and validate models, exclusion of occupational exposures and reliance on various modeling assumptions. The flexible framework, however, will allow for rapid updating of models as new risk information emerges. |
| Policy implications           | Our risk model will allow for defining priority categories for future targeted prevention efforts in a risk-coherent fashion, taking into account age, sex, pre-existing health conditions, racial and other social disparity and local pandemic dynamics. Current projections for the numbers of individuals at different levels of risk can be readily used by national and local policy makers to plan for vaccine allocations.   |

respective white reference populations, African American populations in the United States (relative risk (RR) = 3.18, 95% confidence interval (CI): 3.02–3.36) are at a higher risk compared to the black population (RR = 1.88, 95% CI: 1.65–2.14) in the United Kingdom. Further, compared with non-Hispanic white people in the United States, the non-Hispanic Asian population are at an elevated risk (RR = 1.38, 95% CI: 1.25–1.53) and the Hispanic population (RR = 2.77, 95% CI: 2.60–2.94) and non-Hispanic American Indian/Alaskan Native population (RR = 1.72, 95% CI: 1.23–2.41) are at a substantially elevated risk. External covariate adjustment indicates that accounting for other risk factors, including social deprivation and pre-existing conditions, such as obesity, which are more prevalent in various minority groups than in white populations, only explains a small fraction (<16.0%) of the racial differences in mortality rates (Supplementary Table 1). Adjusted estimates associated with age and racial groups, together with estimates of risk associated with the other factors from an underlying fully adjusted model reported in the UK OpenSAFELY study<sup>15</sup>, are used to define a risk score for individuals in the United States (Supplementary Table 1).

The distribution of the risk score in the National Health Interview Survey (NHIS) sample indicates wide variation of risk across individuals in the United States (Extended Data Fig. 2). Overall, for the US adult population, we estimate that 18.1%, 11.0%, 4.3% and 1.6% of the individuals are at or above risk thresholds associated with elevated ( $\geq 1.2$ -fold), substantially elevated ( $\geq$  twofold), high ( $\geq$  fivefold) and very-high ( $\geq$  tenfold) risk categories, respectively (Table 2, with 95% CIs provided in Supplementary Table 2). The percentage of each population that exceeds these thresholds varies strongly by age. Only a small fraction (0.1%) of the individuals who are younger than 65 years exceed the threshold for high risk. We further examine the distribution of other risk factors among individuals in the defined high-risk groups for the general population (Extended Data Figs. 3 and 4) and the 65-years-and-older population (Extended Data Figs. 5 and 6). Males, Hispanic people, African American people and individuals with history of obesity, diabetes, cancers, high blood pressure, stroke, chronic heart disease, kidney disease, arthritis and respiratory diseases (excluding asthma) are more common in all of the high-risk groups compared with the general NHIS population. We observed a similar pattern for the 65-years-and-older population.

We observed a substantial variation in risk due to predisposing factors across US communities (Fig. 1). The Index of Excess Risk (IER; Methods) varies around tenfold and eightfold across cities and counties for the underlying adult and 65-years-and-older Medicare populations, respectively (Supplementary Tables 3 and 4).

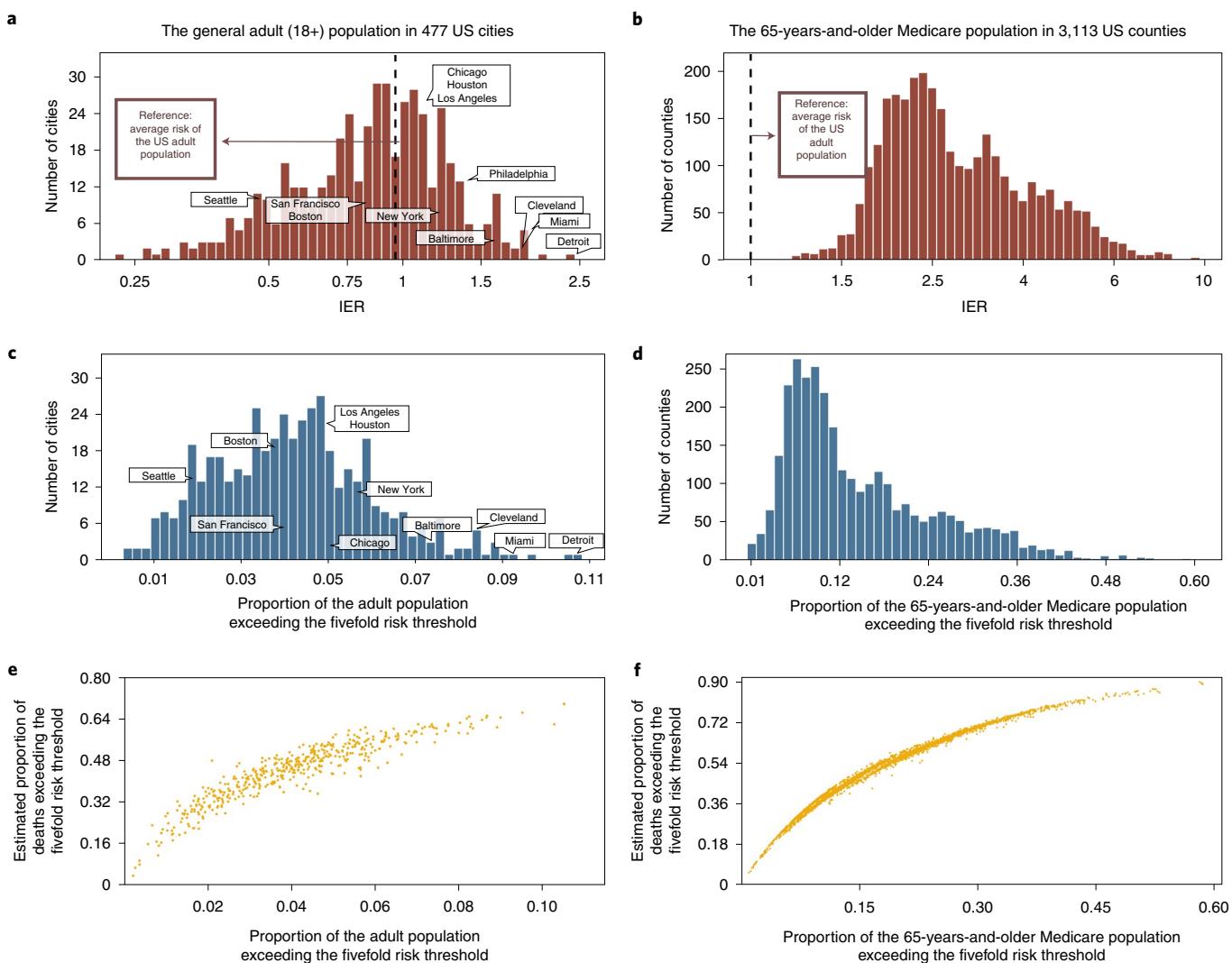
A number of major cities, including Detroit, Miami, Baltimore City, New Orleans and Philadelphia, rank very high according to this index. The proportion of individuals crossing various risk thresholds varies more widely across these communities (Supplementary Tables 3 and 4). For example, the percentage of the adult populations in cities that exceed the fivefold risk threshold varies from 0.4 (Layton, UT) to 10.7 (Detroit, MI). Similarly, the percentage of the 65-years-and-older Medicare population that exceeds the same threshold varies from <1.0% (multiple counties in CO) to >55.0% (multiple counties in TX). Risk distribution for the 65-years-and-older Medicare population also varies substantially across the states (Extended Data Fig. 7 and Supplementary Table 5). Our projections also show that high-risk groups will be disproportionately enriched for deaths across all communities (Fig. 1, Extended Data Fig. 7 and Supplementary Tables 3, 4 and 5). For example, the ratio of the proportion of deaths that are expected to occur in the  $\geq$  fivefold risk group to the proportion of the population at or above the same risk threshold ranges 6.4–32.9 across 477 US cities (Fig. 1).

In a negative binomial regression analysis of deaths (between 7 June 2020 and 1 October 2020) in the 259 counties that contain the 477 cities, we found the coefficient of log(IER) to be statistically significant throughout moving windows of 2 weeks, with an average value of 0.94, close to its ideal value of 1.0, indicating excellent calibration of the underlying individual-level model for the population (Methods and Supplementary Table 6). Further, in a weighted least-squares analysis, we found that log(IER) explains on average 15.4% (95% CI: 12.7–17.3%) of the variation of death rates (on the logarithmic scale) over this time period across the underlying counties (Fig. 2). By comparison, population density and reported 2-week infection rate 3 weeks before the corresponding windows over which deaths were aggregated explain on average 2.4% (95% CI: 2.3–4.1%) and 20.0% (95% CI: 16.9–27.1%) of the variance of the underlying death rates, respectively. In a conditional analysis that accounts for major regional differences in pandemic dynamics, we found that IER explains as much or more of the variance of death rates as 3-weeks-before infection rates (Supplementary Table 6). In an additional validation analysis, we observed similar performance for an IER that was derived for the medication population aged 65 years and older for predicting death rates across 2,999 counties (Supplementary Table 7). In both analyses, we observed the relationship between IER and death rates to be close to linear throughout the range of risk (Extended Data Fig. 8). Finally, using the NHIS risk distribution, we projected that the underlying risk model is expected to have an area under the receiver operating

**Table 2 | Estimated percentages of the NHIS population and corresponding size of US population that exceed various risk thresholds, overall and within the 18–64-years or 65-years-and-older age group**

| Risk category  | Overall                      |   | Aged 18–64 years             |   | Aged 65 years and over       |   |
|----------------|------------------------------|---|------------------------------|---|------------------------------|---|
|                | Population size, million (%) | Deaths expected to arise from the risk category (%) | Population size, million (%) | Deaths expected to arise from the risk category (%) | Population size, million (%) | Deaths expected to arise from the risk category (%) |
| ≥tenfold risk  | 5.4 (1.6%)                   | 27.9%   | 0 (0%)                       | 0.05%   | 5.4 (7.7%)                   | 26.5%   |
| ≥fivefold risk | 14.2 (4.3%)                  | 48.7%   | 0.3 (0.1%)                   | 0.7%  | 13.9 (20.0%)                 | 54.1%   |
| ≥twofold risk  | 36.2 (11.0%)                 | 70.3%   | 2.4 (0.9%)                   | 9.3%  | 33.9 (48.4%)                 | 85.8%   |
| ≥1.2-fold risk | 59.7 (18.1%)                 | 80.0%   | 10.0 (3.8%)                  | 23.9%   | 49.7 (71.1%)                 | 94.6%   |

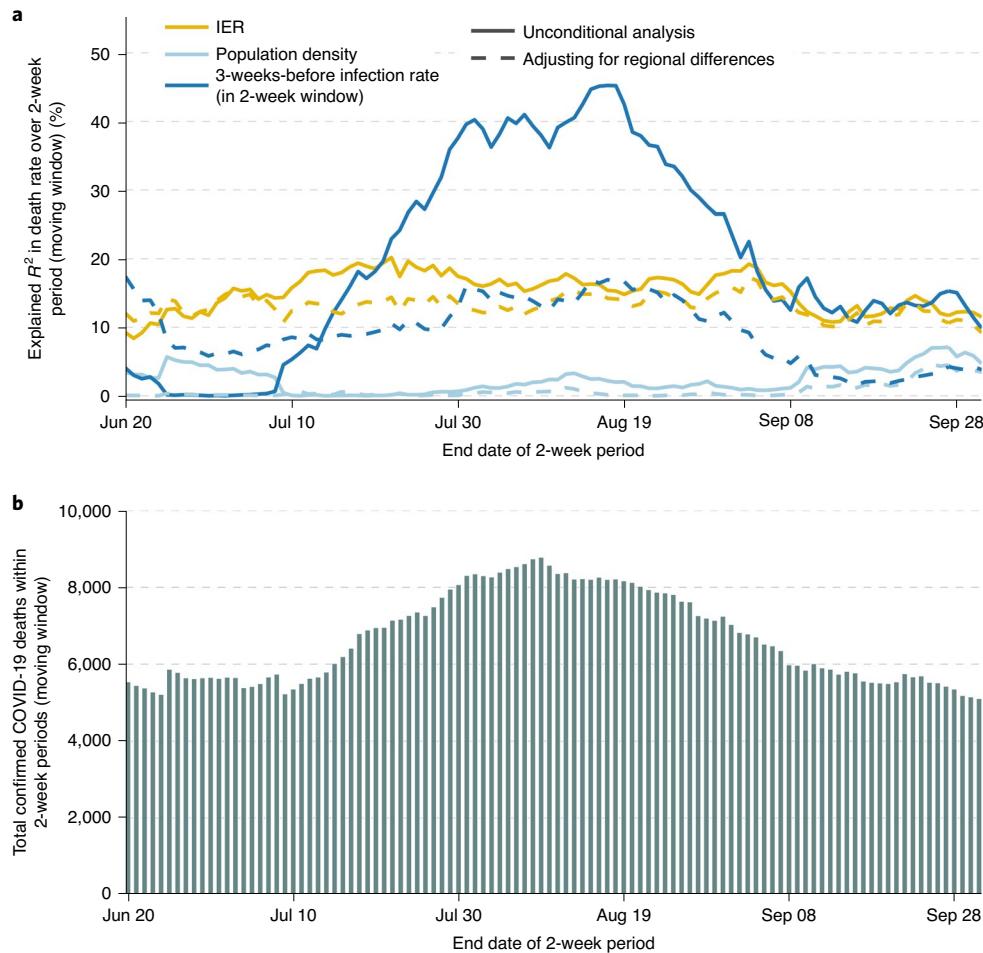
The age distribution was obtained from the most recent US Census Bureau 2019 data and information on other risk factors within each age group was obtained from the NHIS. Risk thresholds are evaluated in reference to the average risk over all subjects. Supplementary Table 1 details the 95% CIs.



**Fig. 1 | Distribution of the IER for COVID-19 mortality and projections for the proportion of high-risk populations across US communities.** Results for the general adult population across 477 US cities (left); results for the 65-years-and-older Medicare population across 3,113 US counties (right). **a,b**, Histograms of IER. **c,d**, Proportions of underlying populations exceeding fivefold risk threshold. **e,f**, Proportions of underlying populations exceeding fivefold risk threshold against proportions of total deaths in underlying populations that are expected to occur within the  $\geq$  fivefold risk group. Results for additional risk thresholds and corresponding 95% CIs are provided in Supplementary Tables 3 and 4. IER is an aggregated measure of risk for a community that accounts for prevalence and co-occurrences of the underlying risk factors and weights the contribution of the different factors according to their associated risk.

characteristic curve (AUC) of 0.90 for individual risk prediction in the United States (Methods), close to that reported for the UK population in the OpenSAFELY study<sup>15</sup>. Further, a simulation study

shows that the level of  $R^2$  achieved for community-level projections is consistent with the projected value of AUC for individual-level predictions (Extended Data Fig. 9).



**Fig. 2 | Validation of the risk model using county-level mortality data.** **a**, Explained variance ( $R^2$ ) of log death rate in a moving window of 2-week periods by IER, population density and 3-weeks-before infection rate (also aggregated over a 2-week window), across 259 counties representing 477 studied cities. All data are transformed to log scale. The county-level IER was calculated as the weighted average of city-level IERs with each city being weighted by its population size. The dashed curves show results from a conditional analysis, where region indicators for the Northeast, Midwest, South and West are introduced as additional covariates in the regression to account for major regional differences in pandemic dynamics. **b**, Cumulative number of confirmed COVID-19 deaths across 259 counties during corresponding 2-week periods. During each 2-week period, only the counties with nonzero deaths and infections were included in analyses.

We have made available a web-based risk calculator (<https://covid19risktools.com/riskcalculator>) that allows an individual to input information on risk factors and obtain estimates of individualized risk for COVID-19 mortality on both relative and absolute risk scales (Extended Data Fig. 10). The relative risk for individuals are reported based on the underlying risk score benchmarked with respect to a ‘population average risk’ defined as population-weighted average risk (IER) across the cities. The calculator returns a numerical value for relative risk and a color-coded categorization of risk into five categories: <1.2 (close to or less than average risk), 1.2–2.0 (moderately elevated risk), 2.0–5.0 (substantially elevated risk), 5.0–10.0 (high risk) and >10 (very high risk). Further, for each person, information on risk score is combined with projections available from the Ensemble pandemic forecasting models (<https://covid-19forecasthub.org/doc/ensemble/>, <https://covid-19forecasthub.org/doc/ensemble/> and <https://github.com/reichlab/covid19-forecast-hub#ensemble-model>) in their state of residence to report an absolute rate of mortality over a specified period of time. Both types of risk are provided with 95% CIs to reflect various sources of uncertainty in the projections. We have also made available interactive maps for viewing projections of the number and proportion of

individuals at different risk categories across US cities, counties and states (<https://jhu covid19.policymap.com/app>).

We developed a COVID-19 mortality risk model for the general US population by combining information across multiple data sources. We believe that the model is unique in that it can be used to project absolute rate of mortality for individuals with different risk profiles by combining information on individual-level risk factors, as well as on changing dynamics in the epidemic at the community-level captured through available forecasting models. We applied the model to data available from US national databases to identify high-risk cities and counties and estimate the size of populations at risk within these communities. Our findings could inform policy developments for equitable distribution of early vaccines and other scarce preventive resources. Further, the proposed methodological framework could aid flexible development and updating of other risk models and subsequently use them to perform population-level risk projections.

The US National Academy of Sciences, Engineering and Medicine (NASEM) has released a comprehensive report on guidelines, and underlying principles, for equitable vaccine allocations<sup>23</sup>. The report recommends a four-phase plan with different populations being

prioritized on the basis of evidence on risk of infection, severity and/or mortality following infection, negative societal impact such as those on national security and public services and transmission of infection to others. Although risk factors such as pre-existing conditions and living in overcrowded settings are considered, it is not clear how age and pre-existing conditions are relatively weighted to derive priority categories. Further, while the report acknowledges the excess burden of COVID-19 for various minority populations and areas of high social vulnerability, the development of actual criteria based on these factors has been left to be determined by local governments. Thus, we believe that while the NASEM report provides a strong set of ethical and procedural principles and a broad framework for vaccine prioritization, more quantitative analyses are needed so that priority categories can be better aligned with underlying levels of risk of individuals and populations.

In principle, similar analyses to ours could also be useful for developing strategies for the global distribution of vaccines and preventive therapeutics<sup>24</sup>. We are working with investigators from the Pan American Health Organization to identify risk factor surveillance datasets and provide estimates of the size of high-risk populations across countries in South America. Development of policies for the efficient and impactful distribution of vaccines, however, also depends on many other factors, including overall social benefit, ease of implementation and available infrastructures. In particular, using priority categories based on a simple set of rules, as opposed to individual risk calculations, might be desirable for the ease of implementation. Risk models, nevertheless, can be useful to evaluate effectiveness of any proposed prioritization plan for its effectiveness in reducing population burden of serious illness or deaths. Further, when data from vaccine trials become available, it will be important to explore possible heterogeneity in vaccine efficacy by risk groups, defined by individual and combinations of risk factors and appropriately modify allocation strategy to maximize population benefit.

Our risk tools and projections can be useful for identifying high-risk groups that would benefit most from ‘shielding’ efforts until they can be vaccinated. In the beginning of the pandemic, the National Health Service of the United Kingdom identified about 1.5 million individuals to be at extremely high risk due to selected conditions and provided them with assistance for food delivery and medical services<sup>25</sup>. In California, local and state governments developed Project Roomkey (<https://covid19.lacounty.gov/project-roomkey/>) to provide free-of-charge hotel rooms, meals and other services to asymptomatic homeless people who are at high risk due to their age or/and underlying health conditions. In the future, as businesses, schools and higher education institutes reopen, strategies need to be in place to identify and shield high-risk individuals. Finally, general population risk tools can also help healthy individuals to understand future risk for serious outcomes, not only for themselves, but also for family members and friends and thus could better motivate them to adhere to standard guidelines for infection prevention, such as through handwashing and mask wearing.

A few studies in the past have investigated the proportions of ‘high-risk’ individuals for COVID-19-related serious illness or mortality in the United Kingdom, the United States and across nations globally (<https://www.nytimes.com/interactive/2020/05/18/us/coronavirus-underlying-conditions.html?auth=link-dismiss-google1tap>)<sup>25–27</sup>. These studies have defined high-risk individuals based on the prevalence of one or more risk factors without taking into account the relative contribution of these factors. Further, because of the broad definition used, they estimate that a very large fraction of the populations, 20% in the United Kingdom and 16–31% globally, are at high risk. By contrast, we have defined different risk categories based on an underlying score that allows one to assign a more precise magnitude of risk to these categories. Further, our framework allows evaluation of future absolute

risk for individuals and communities, incorporating information from pandemic forecasting models and thus is uniquely suitable for planning vaccination and other prevention efforts across regions that may have wide variation in the infection dynamics.

Our study has several limitations. First and foremost, information on risk for the majority of risk factors was derived from the UK-based OpenSAFELY study<sup>15</sup>. We have modified the model to make it suitable for the US population by incorporating population-based information on age- and race-associated rate of mortality. Further, we have empirically shown through independent validation analyses that the projected risk is well calibrated for the general US adult population and correlates strongly with death rates across counties in the United States. There is, however, an urgent need for individual-level data from large population-based studies, akin to the UK OpenSAFELY study, for building and validating general population risk models in the US setting.

Another limitation of our study is that we have not incorporated information associated with front-line occupations that pose higher risk of infection (<https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/causesofdeath/bulletins/coronaviruscovid19relateddeathsbyoccupationenglandandwales/deathsregistereduptoandincluding20april2020>)<sup>28</sup>. We have mapped individuals in the NHIS study to various high-risk occupation categories (Supplementary Table 8) and have observed overrepresentation of minority populations in highest risk categories, such as for black and Hispanic individuals in the low-skilled elementary occupation category. Further, a post-publication report from the UK OpenSAFELY study (<https://doi.org/10.1038/s41586-020-2521-4>) indicates the presence of interactions between age and other risk factors, although required information on all parameters from a fully multivariate model is not available yet. We plan to continually update our risk model, and the corresponding community-level risk projections, through incorporation of emerging information on risk associated with occupations, age interactions and other new risk factors.

We present a comprehensive and flexible framework for assessing general population risk of COVID-19 mortality incorporating individual-level risk profiles, population-level risk factor distribution and time-varying pandemic dynamics. Our risk projections for US cities and counties might be useful for guiding strategies for equitable allocation of early vaccines and other preventive resources in the coming months. Our risk tool and the underlying statistical methodologies can be applied to carry out similar analyses internationally and thus to inform prevention efforts globally.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41591-020-01191-8>.

Received: 16 August 2020; Accepted: 24 November 2020;

Published online: 11 December 2020

## References

- Emanuel, E. J. et al. Fair allocation of scarce medical resources in the time of COVID-19. *N. Engl. J. Med.* <https://doi.org/10.1056/NEJMsb2005114> (2020).
- Link, A. & Hold, G. First case of COVID-19 in the United States. *N. Engl. J. Med.* **382**, e53 (2020).
- Chinazzi, M. et al. The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. *Science* **368**, 395–400 (2020).
- Kraemer, M. U. G. et al. The effect of human mobility and control measures on the COVID-19 epidemic in China. *Science* **368**, 493–497 (2020).
- Giordano, G. et al. Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy. *Nat. Med.* **26**, 855–860 (2020).

6. Pan, A. et al. Association of public health interventions with the epidemiology of the COVID-19 outbreak in Wuhan, China. *JAMA* **323**, 1915–1923 (2020).
7. Deng, G., Yin, M., Chen, X. & Zeng, F. Clinical determinants for fatality of 44,672 patients with COVID-19. *Crit. Care* **24**, 179 (2020).
8. Docherty, A. B. et al. Features of 20,133 UK patients in hospital with covid-19 using the ISARIC WHO clinical characterisation protocol: prospective observational cohort study. *BMJ* **369**, m1985 (2020).
9. Guan, W. J. et al. Clinical characteristics of coronavirus disease 2019 in China. *N. Engl. J. Med.* **382**, 1708–1720 (2020).
10. Khunti, K., Singh, A. K., Pareek, M. & Hanif, W. Is ethnicity linked to incidence or outcomes of COVID-19? *BMJ* **369**, m1548 (2020).
11. Parohan, M. et al. Risk factors for mortality in patients with coronavirus disease 2019 (COVID-19) infection: a systematic review and meta-analysis of observational studies. *Aging Male* <https://doi.org/10.1080/13685538.2020.1774748> (2020).
12. Richardson, S. et al. Presenting characteristics, comorbidities, and outcomes among 5700 patients hospitalized with COVID-19 in the New York City area. *JAMA* **323**, 2052–2059 (2020).
13. Wenham, C., Smith, J. & Morgan, R. COVID-19: the gendered impacts of the outbreak. *Lancet* **395**, 846–848 (2020).
14. Yancy, C. W. COVID-19 and African Americans. *JAMA* **323**, 1891–1892 (2020).
15. Williamson, E. J. et al. Factors associated with COVID-19-related death using OpenSAFELY. *Nature* **584**, 430–436 (2020).
16. Rasmussen, S. A., Khoury, M. J. & Del Rio, C. Precision public health as a key tool in the COVID-19 response. *JAMA* **324**, 933–934 (2020).
17. Smith, G. D. & Spiegelhalter, D. Shielding from COVID-19 should be stratified by risk. *BMJ* **369**, m2063 (2020).
18. van Bunnik, B. A. D. et al. Segmentation and shielding of the most vulnerable members of the population as elements of an exit strategy from COVID-19 lockdown. Preprint at *medRxiv* <https://doi.org/10.1101/2020.05.04.20090597> (2020).
19. Folegatti, P. M. et al. Safety and immunogenicity of the ChAdOx1 nCoV-19 vaccine against SARS-CoV-2: a preliminary report of a phase 1/2, single-blind, randomised controlled trial. *Lancet* **396**, 467–78. (2020).
20. Jackson, L. A. et al. An mRNA vaccine against SARS-CoV-2: preliminary report. *N. Engl. J. Med.* **383**, 1920–1931 (2020).
21. Sahin, U. et al. COVID-19 vaccine BNT162b1 elicits human antibody and T<sub>H</sub>1 T cell responses. *Nature* **586**, 594–599 (2020).
22. Walsh, E. E. et al. Safety and immunogenicity of two RNA-based COVID-19 vaccine candidates. *N. Engl. J. Med.* <https://doi.org/10.1056/NEJMoa2027906> (2020).
23. National Academies of Sciences Engineering Medicine. *Framework for Equitable Allocation of COVID-19 Vaccine* (National Academies Press, 2020).
24. Emanuel, E. J. et al. An ethical framework for global vaccine allocation. *Science* **369**, 1309–1312 (2020).
25. Banerjee, A. et al. Estimating excess 1-year mortality associated with the COVID-19 pandemic according to underlying conditions and age: a population-based cohort study. *Lancet* **395**, 1715–1725 (2020).
26. Clark, A. et al. Global, regional, and national estimates of the population at increased risk of severe COVID-19 due to underlying health conditions in 2020: a modelling study. *Lancet Glob. Health* **8**, e1003–e1017 (2020).
27. Adams, M. L., Katz, D. L. & Grandpre, J. Population-based estimates of chronic conditions affecting risk for complications from coronavirus disease, United States. *Emerg. Infect. Dis.* **26**, 1831–1833 (2020).
28. Nguyen, L. H. et al. Risk of COVID-19 among front-line health-care workers and the general community: a prospective cohort study. *Lancet Public Health* **5**, e475–e483 (2020).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2020

## Methods

**Definition of COVID-19 mortality risk score.** The risk score for an individual is defined as a weighted combination of various sociodemographic characteristics and predisposing health conditions, with weights defined by the relative magnitude of the contribution of these factors to the risk of death due to COVID-19 in the adult population. We use two sources of information to build the risk score: (1) multivariate-adjusted estimate of risk associated with sex, quintiles for social deprivation index (SDI), which is an approximation for the Index of Multiple Deprivation (IMD), and 12 pre-existing conditions, including body mass index (BMI), smoking status, blood pressure, respiratory disease excluding asthma, asthma, chronic heart disease, diabetes, nonhematological cancer, hematological cancer, stroke, kidney disease and rheumatoid arthritis, from the recently published UK-based OpenSAFELY study<sup>15</sup> and (2) death rates associated with different age and racial/ethnic groups in the United States published by the Centers for Disease Control and Prevention (CDC), after performing external covariate adjustment accounting for the correlation of these factors with other risk factors in the model (described below).

**Estimation of US-specific risk associated with age and racial/ethnic groups and external covariate adjustments.** We used data made available by the CDC on reported COVID-19 deaths as of 6 June 2020, by race, age and state. We fitted a Poisson regression to observe death counts across strata defined by age, race/ethnicity and state. Specifically, the model takes the following form:  $\log(D_{ijk}) = \log(P_{ijk}) + \sum_i \gamma_i State_i + \sum_j \beta_j Race_j + \sum_k \delta_k Age_k + \epsilon_{ijk}$ , where  $D_{ijk}$  and  $P_{ijk}$  denote the death count and population total in state  $i$ , race category  $j$  and age category  $k$ , respectively. The corresponding error term is denoted by  $\epsilon_{ijk}$ . The model accounts for the underlying population size by an offset term and assumes additive effects of age and race/ethnicity categories on log death rate after adjusting for states as fixed effects in the model. While fitting the model, we excluded the race categories: 'more than one race' and 'Native Hawaiian or Other Pacific Islander'. We subsequently performed external covariate analysis to account for possible correlation between age and race/ethnicity with other risk factors in the model. We used the NHIS dataset as a reference dataset for estimation of joint distribution of all risk factors in the US setting. We assumed that the association of all risk factors, except age and race, in the fully adjusted model was the same across the United Kingdom and United States. We used generalized method of moment techniques that we developed earlier<sup>29</sup> to obtain estimates for the effects of age and race adjusted for other risk factors, with their effects being fixed at those available from the fully adjusted model from the UK OpenSAFELY study, based on corresponding unadjusted estimates available from CDC data and correlations of these risk factors with others as observed in the NHIS dataset. The final set of parameters used for the US fully adjusted model is shown in Supplementary Table 1. We observed that while adjustment for other risk factors attenuates the risk associated with some of the ethnic/racial groups, to a large extent the excess risk remains.

**Data sources and processing.** We utilized a variety of data sources to obtain the latest information on the prevalence of demographic variables and health conditions across US cities. The resources include American Community Survey (ACS) of the US Census Bureau for age, sex and race (2017/2018 Table, 1-year estimates), Behavioral Risk Factor Surveillance System (BRFSS) 2017 survey of the CDC for various health conditions and smoking, National Health and Nutrition Examination Survey (NHANES) for estimating relative proportions of certain subcategories of conditions that were not available in BRFSS, US Cancer Statistics (2012–2016 incidence rates) maintained by the National Cancer Institute (NCI) and the CDC to derive prevalence of hematological and nonhematological malignancies and a database available from the Robert Graham Center on SDI derived from data available from the ACS (2011–2015, 5-year estimates). In addition, we accessed individual-level data from the NHIS of CDC. We extracted individual-level data on the risk factors of 22,109 adults from the 2017 NHIS study. All the required variables, except SDI, were available for individuals in the NHIS. For projections of risk for the United States overall, we applied the most recent age distribution from the US Census Bureau 2019 data and information on other risk factors within age groups from the NHIS.

We used 2018 data from the Centers for Medicare and Medicaid Services (CMS) to obtain the latest information on prevalence of chronic health conditions across US counties for the 65-years-and-older Medicare population (the <65-year-old Medicare population has missing information on age and thus was not included in the analyses). We utilized data for individuals aged 65 years or older from 2017–2018 NHANES and 2017 NHIS to estimate relative proportions of certain subcategories of health conditions that were not available from the CMS. Prevalence of age, race and sex variables for the Medicare population were obtained from the ACS.

Definitions of categories of each risk factor, the corresponding data sources and weights ( $\beta_k$ s) are summarized in the following sections.

**Behavioral Risk Factor Surveillance System.** We used the BRFSS '500 Cities: Local Data for Better Health, 2019 release'. This dataset provides information on estimates of crude prevalence for covariates related to unhealthy behaviors and

health outcomes among adult US residents in 2017. As the 2019 release is from the 2017 data, which were based on a telephone survey across the states, the total sample size (landline and cellphone) varied from 39,510 to 754,950 with a mean response rate of 44.9%.

In the UK analysis, BMI in the obese group has three categories: Obese class I ( $30\text{--}34.9\text{ kg m}^{-2}$ ), Obese class II ( $35\text{--}39.9\text{ kg m}^{-2}$ ) and Obese class III ( $>40\text{ kg m}^{-2}$ ). From the BRFSS data, we obtained estimates of total prevalence for these three obese groups for each city, which is defined as the proportion of respondents aged  $\geq 18$  years who have a  $\text{BMI} \geq 30.0\text{ kg m}^{-2}$ . We then obtained the relative prevalence of Obese I, Obese II and Obese III subcategories from the NHIS 2017 population. Assuming that among obese individuals, the relative proportion of the three obese categories was the same across the NHIS population and cities, we applied the NHIS estimate of relative proportions to city-level information on overall prevalence to obtain prevalence estimates for each of the three categories of obesity across different cities.

Regarding smoking status, each individual was defined as either nonsmoker, former smoker or current smoker. We obtained information on the estimate of crude prevalence of smokers (former and current smokers combined) for each city from BRFSS data. This was defined as the proportion of respondents aged  $\geq 18$  years who reported having smoked  $\geq 100$  cigarettes in their lifetime and currently smoke every day or some days. Using the relative proportion of current to former smokers from the NHIS, we further allocated smokers into current and former smokers and obtained the respective prevalence for each city.

Blood pressure is categorized as either normal or high/diagnosed hypertension, where hypertension is defined as systolic BP  $\geq 140\text{ mm Hg}$  or diastolic BP  $\geq 90\text{ mm Hg}$ . We used estimates of crude prevalence for this group, defined as the group of respondents aged  $\geq 18$  years who reported ever having been told by a doctor, nurse or other health professional that they have high blood pressure, from the BRFSS data. Estimates of city-level crude prevalence for respiratory disease other than asthma, chronic heart disease, stroke/dementia, kidney disease and rheumatoid arthritis/lupus/psoriasis were directly obtained from BRFSS data.

Diabetes is quantified based on glycated hemoglobin (HbA1c) measurement in  $\text{mmol mol}^{-1}$  into three categories: controlled ( $\text{HbA1c} < 58\text{ mmol mol}^{-1}$ ), uncontrolled ( $\text{HbA1c} \geq 58\text{ mmol mol}^{-1}$ ) and no recent HbA1c measure. From BRFSS we obtained the crude prevalence of diabetes, which is the proportion of respondents aged  $\geq 18$  years who reported ever been told by health professional that they have diabetes other than diabetes during pregnancy. We then estimated the relative proportion of controlled and uncontrolled diabetes using information on HbA1c in the 2016–2017 NHANES and applied this proportion to the prevalence of overall diabetes, information on which was available for each city from BRFSS, and thus obtained estimates of prevalence of controlled and uncontrolled diabetes across cities.

Asthma was grouped by status of recent use of oral corticosteroids (OCSs). The BRFSS defines crude prevalence of asthma as the ratio of weighted number of respondents who answered 'yes' to both of the following questions: 'Have you ever been told by a doctor, nurse or other health professional that you have asthma?' and 'Do you still have asthma?' to the weighted number of respondents. The OpenSAFELY study reported about 10.69% of patients with asthma were recent OCS users in the United Kingdom and we used this same proportion to allocate the proportion of patients with asthma for each city to recent OCS users.

For the United States, we have not found available data for city-level crude prevalences of several risk factors considered in the UK study<sup>15</sup>, including liver disease, organ transplant, spleen diseases and other neurological and immunosuppressive conditions and hence did not take these comorbidities into account in the risk score calculation. As these conditions are generally rare, we do not expect that exclusion of these variable from the risk score will lead to a change in risk coefficients for the other variables.

**US Census Bureau.** We used the publicly available ACS data to obtain the prevalence of demographic variables including age, sex and ethnicity at city level ('census-designated place').

Distribution of age and sex for each city were obtained from the 2017 table. The number of individual interviews were around 2,303,000. The age categories available from the census data could be collapsed into age groups 15–<45, 45–<55, 55–<65, 65–<75, 75–<85 and 85+. Sex was categorized into male and female.

The city-wide information on ethnicity was extracted from the latest available 2018 data. The number of individual interviews were around 2,300,000. Ethnicity included in the analysis has the following categories: white (reference), black, Asian, American Indian/Alaskan Native and Hispanic.

**Centers for Disease Control and Prevention.** We used state-wide COVID-19 death rates reported till 6 June 2020 to estimate risk of COVID-19 mortality associated with age and ethnicity/race categories in the United States. The age categories were 15–<45, 45–<55, 55–<65, 65–<75, 75–<85 and 85+ and the race categories were white, black, Asian, American Indian/Native Alaskan and Hispanic.

**United States Cancer Statistics.** The United States Cancer Statistics data are a combined cancer data source from the CDC and the NCI that provides statistics

on the cancer incidence (newly diagnosed cases) across different cancer sites. We used the most recent county-level incidence data available for 5 years combined from 2012 to 2016. In the UK analysis, the cancer-related risk factors include hematological malignancies and nonhematological cancer, each of which was grouped by time since first diagnosis (within the last year, 2–<5 years or >5 years). To use this categorization for cancer comorbidities, we carried out the following steps.

We first classified the incidence data for various cancer sites into hematological cancer and nonhematological cancer for each county. In particular, 'Hodgkin lymphoma', 'non-Hodgkin lymphoma', 'leukemias' and 'myeloma' were grouped into hematological cancer and the rest were grouped into nonhematological cancer. For each of the two cancer groups, we adjusted the county-level annual incidence rates with corresponding overall annual US survival rates (2017 Leading Cancer Cases and Deaths) to obtain county-level prevalence of cancer groups across categories based on diagnosis years. We conducted the following steps.

We first obtained 5-year US survival rates for each cancer group using a weighted combination of annual US survival rates of cancer sites belonging to that specific group. The weights are the proportion of incident cases of cancer sites among incident cases of the particular group. We then computed annual death rate from the 5-year survival rate.

We show the calculations for hematological cancer as an example. Let us denote the obtained annual US death rate for hematological cancer by  $D_{hema}$ . Next, we denote the annual incidence rates of the hematological cancer group of the  $k$ th county by  $I_{hema,k}$ . We calculate the prevalence in each of the following categories:

- (1) Diagnosed in <1 year. We adjust 1-year incidence rates using 6-month survival rates assuming during that 1-year period, people on average had cancer 6 months ago. The prevalence of this category is computed by  $I_{hema,k} \times (1 - \frac{D_{hema}}{2})$ .
- (2) Diagnosed in  $\geq 1$  year and <5 years. We adjust 4-year incidence rates with 2-year survival rates. Using a similar idea, we calculate the prevalence using the formula given by  $4I_{hema,k} \times (1 - 2D_{hema})$ .
- (3) Diagnosed in  $\geq 5$  year. We consider the upper bound for time since diagnosis to be 20 years and hence, we adjust 15-year incidence rates with 7.5-year survival rates in a similar way as for steps (1) and (2).

We repeat the above steps to obtain the cancer prevalence of nonhematological cancer in the three categories.

For city-level projections, we assumed the prevalence for different cancer in a city to be the same as that of the underlying counties because city-level cancer incidence rates were unavailable. The approximation is expected to be reasonable for major metropolitan cities that tend to contain the majority of the population for the underlying county. However, for smaller cities that represent a small proportion of the population for the underlying counties, there is likely to be some bias due to the difference in sociodemographic characteristics of individuals living in and outside cities.

**Robert Graham Center and American Community Survey.** One of the upstream risk factors considered in the UK analysis was social deprivation, quantified using the IMD in quintiles, which is not available in the US setting. It is a geographic level measure with higher values indicating greater deprivation. In our study, we considered an approximate measure that is available in the United States, the SDI. SDI is a composite measure of seven demographic characteristics, developed by the Robert Graham Center from the ACS. We obtained the latest available data for county-level SDI quintiles in 2015 and for our city-level analysis, we defined the SDI quintile for each city as the SDI quintile of the corresponding county to which the city belongs to. The major components of IMD and SDI are similar and capture income, education, employment and housing condition. The seven factors used for calculating the SDI include the percentage population <100% Federal Poverty Line, percentage population aged 25 years or more with <12 years of education, percentage nonemployed, percentage population living in renter-occupied housing units, percentage population living in crowded housing units, percentage single-parent households with dependents <18 years and percentage population with no car.

**National Health Interview Survey.** The NHIS data are based on in-person interviews conducted by the National Center for Health Statistics. A sample weight is assigned to each participant to ensure that the NHIS data are representative of civilian noninstitutionalized US population. We used the 2017 NHIS data to obtain individual-level information on various risk factors.

**Center for Medicare and Medicaid Services Office of Minority Health.** The CMS data are based on the administrative claims of the Medicare beneficiaries obtained from the Chronic Conditions Warehouse. The data provide extensive information on county-wise prevalence of chronic and disabling health conditions (2018) for the 65-years-and-older population enrolled through the Free-for-Service Program. The health conditions include liver disease, cerebral palsy and HIV/AIDS in addition to the ones in BRFSS. We obtained county-level prevalence of race and sex categories for the Medicare population from the US census. We excluded 11 counties that had missing data for multiple risk factors. All remaining

3,113 counties had prevalence data available for hematological/nonhematological cancers, but some had missing information on subcategories of cancer prevalence, defined by number of years since cancer onset. For the 966 and 196 counties that had missing information on subcategories of hematological and nonhematological cancers, respectively, we conducted imputation conditional on age groups, considering the three subcategories of hematological cancer as an example. We first used NHIS data to estimate age-stratified (65–74, 75–84 and 85+ age groups) proportions of each subcategory of hematological cancer (diagnosed within 1 year, 1–5 years ago or >5 years ago) among the 65-years-and-older NHIS individuals with a history of hematological cancer. For each county, we then allocated the prevalence of age groups and the overall prevalence of hematologic cancer for the county to these different subcategories according to their relative proportions observed in the NHIS sample.

**Statistical models and methods. A framework for integrating population- and individual-level risk.** Similar to the OpenSAFELY study, we assumed that the risk of COVID-19 death at time  $t$  for an individual  $i$  residing in location  $l$ , for example a city or a county, can be described by the proportional risk model

$$\lambda_{il}(X, t) = \lambda_l(t) \exp\left(\sum_{k=1}^K \beta_k X_{ik}\right) = \lambda_l(t) R_i(\beta)$$

where  $\lambda_l(t)$  denotes the baseline risk for location  $l$  due to underlying pandemic characteristics, such as social distance measures, population density and mobility patterns and  $R_i(\beta) = \exp\left(\sum_{k=1}^K \beta_k X_{ik}\right)$  denotes a multiplicative factor associated with risk due to various predisposing factors. Here,  $t$  refers to calendar time since some landmark, such as the day when cumulative death reaches some minimum threshold. The average risk of the population at location  $l$  can be defined as

$$\lambda_l^A(t) = \lambda_l(t) E_l \left\{ \exp\left(\sum_{k=1}^K \beta_k X_{ik}\right) \right\},$$

where  $E_l$  denotes the expectation (average) with respect to distribution of the individual-level risk factors in location  $l$ . The above formula allows linking individual-level relative risk models to pandemic scenario models and hence can produce estimates of absolute risk of individuals, taking into account both individual-level risk factors and community-level risk due to pandemic dynamics. In particular, there are a variety of pandemic models available to produce estimates of population-level risk  $\lambda_l^A(t)$ , for example in the state of residence of an individual, over the course of a period of time in the future, which could take into account local characteristics, such as reproduction rate, population density and mobility patterns and such information can be used to calculate baseline risk  $\lambda_l(t)$  through the equation:

$$\lambda_l(t) = \frac{\lambda_l^A(t)}{E_l \left\{ \exp\left(\sum_{k=1}^K \beta_k X_{ik}\right) \right\}},$$

and hence the absolute risk denoted as  $\lambda_l(X, t)$ . Here we consider the Ensemble model developed by the Reich Laboratory at the University of Massachusetts Amherst (<https://covid19forecasthub.org/doc/ensemble/> and <https://github.com/reichlab/covid19-forecast-hub#ensemble-model>)<sup>30</sup>, which produces projection of deaths by taking the arithmetic average of the projections from up to 61 individual forecasting models. The list and detailed descriptions of the individual forecasting models, including assumptions about changing dynamics in the epidemic, are provided at the Reich Laboratory COVID-19 Forecast Hub (<https://github.com/reichlab/covid19-forecast-hub#ensemble-model>). We download state-level projections and corresponding 95% CIs on a weekly basis.

**Approximate characterization for the distribution of risk score (risk on log scale).** Recall that we defined the risk score of COVID-19 death for an individual  $i$  as  $RS_i = \sum_{k=1}^K \beta_k X_{ik}$ . We observed that the distribution of the calculated individual-level risk score for the NHIS population could be approximated by a mixture-normal distribution with three components that correspond to the 18–44-year age group, 45–74-year age group and 75+ age group, respectively (Extended Data Fig. 2). Under the mixture-normal approximation, we can characterize the distribution of risk scores for each location by evaluating its mean and variance within each of the three age groups:

$$E_{l,a}\{RS\} = \sum_{k=1}^K \beta_k E_{l,a}\{X_k\}, \quad (1)$$

$$V_{l,a}\{RS\} = \sum_{k=1}^K \beta_k^2 Var_{l,a}\{X_k\} + \sum_{k=1}^K \sum_{k'=1}^K \beta_k \beta_{k'} Cov_{l,a}(X_k, X_{k'}), \quad (2)$$

where  $a$  is the indicator for age group ( $a = 1$ : aged 18–44 years,  $a = 2$ : aged 45–74 years,  $a = 3$ : aged 75+ years),  $E_{l,a}$  and  $V_{l,a}$  denote the expectation

and variance of the risk score with respect to distribution of the risk factors among age group  $a$  in location  $l$ .

In the following paragraphs, we will focus on the discussion of approximating characteristics of the risk score distribution for the age 15+ years general population in the 477 cities by the mixture-normal distribution. For the 65-years-and-older Medicare population in each county, we observed that the distribution of risk score could be adequately approximated by a single normal distribution as shown by the NHIS 65-years-and-older population (Extended Data Fig. 2). The corresponding calculations were thus performed based on a single normal distribution.

For simplicity, we denote the mean and variance of the risk score within age group  $a$  in location  $l$  as  $(\mu_{l,a}, \sigma_{l,a}^2)$ ,  $a=1,2,3$ . In this particular analysis, the  $X_k$ s denote the dummy variables that indicate the levels of categorical risk factors and thus the mean and variance of the risk score can be characterized by the estimates of prevalence of the individual risk factor categories and joint prevalence of pairs of categorical variables. Briefly, for a location  $l$ , within each age group, we first obtained the age-stratified prevalence of each risk factor category, then calculated the co-prevalence of each pair of categorical variables using the corresponding odds ratio estimated from NHIS individual-level data.

We first discuss how to calculate the prevalence of each categorical variable within each age group in a location using information on the overall prevalence for that location and relationship between age and that variable estimated from NHIS data. We denote the age group by an indicator variable  $X_1$  ( $X_1=1$  denotes the 18–44 age group and 0 denotes the 45+ age group), and the categorical variable by  $X_2$ . In the scenarios where  $X_2$  has more than two categories, we re-classified  $X_2$  into two categories, one for which we computed the prevalence and the other where we collapsed the rest. We estimated the odds ratio, a measure of association, between  $X_1$  and  $X_2$  using NHIS data after accounting for sampling weights. The odds ratio is defined as  $(P(X_1=1, X_2=1)P(X_1=0, X_2=0))/(P(X_1=1, X_2=0)P(X_1=0, X_2=1))$  and, under a logistic regression framework with  $X_1$  being the outcome and  $X_2$  being the exposure, it is equal to  $e^\theta$ , where  $\theta$  is the regression parameter associated with  $X_2$ . Thus, we fit a logistic regression model with  $X_1$  as the outcome and  $X_2$  as the exposure and estimate  $\theta$  by taking into account the individual weights  $\theta$  in NHIS data. We assumed that the odds ratio estimates obtained from the NHIS could be generalized to underlying populations of cities. We plugged in the odds ratio estimate and the city-specific marginal prevalence of the two groups in equation (3), described below, to derive the proportions of cells in a  $2\times 2$  contingency table defined by  $Y$  and  $X$ . We then evaluated age-group-specific prevalence of the required category by taking the ratio of the corresponding cell proportion and city-specific marginal prevalence of corresponding age groups.

**Obtaining joint prevalence of two factors from marginal prevalence and odds ratios.** Suppose that we have two binary variables for two different risk factors, which, for city  $l$ , are denoted as  $X_1^l$  and  $X_2^l$  with marginal prevalence  $p_1^l$  and  $p_2^l$ , respectively and odds ratio  $r = p_{11}^l p_{00}^l / (p_{01}^l p_{10}^l)$ , which is estimated based on NHIS individual-level data and is assumed to be constant across all cities. We will use these notations without  $l$  for simplicity. The  $2\times 2$  table for the city-specific  $X_1$  and  $X_2$  is

| $X_1=1$ | $X_1=0$  | Total    |         |
|---------|----------|----------|---------|
| $X_2=1$ | $p_{11}$ | $p_{10}$ | $p_2$   |
| $X_2=0$ | $p_{01}$ | $p_{00}$ | $1-p_2$ |
| Total   | $p_1$    | $1-p_1$  | 1       |

Given  $p_1$ ,  $p_2$  and  $r$ , we have the following equations:

$$\begin{aligned} p_{11} + p_{10} &= p_2, \\ p_{11} + p_{01} &= p_1, \\ \frac{p_{11}p_{01}}{p_{01}p_{10}} &= r, \\ p_{11} + p_{10} + p_{01} + p_{00} &= 1. \end{aligned}$$

It is easy to see that  $p_{00}=1-p_1-p_2+p_{11}$ . We then have

$$\begin{aligned} \frac{p_{11}(1-p_1-p_2+p_{11})}{(p_1-p_{11})(p_2-p_{11})} &= r \\ \Rightarrow (1-r)p_{11}^2 + [1+(r-1)(p_1+p_2)]p_{11} - rp_1p_2 &= 0. \end{aligned}$$

We define  $a=1-r$ ,  $b=1+(r-1)(p_1+p_2)$  and  $c=-rp_1p_2$ , then the solution is given by the quadratic formula:

$$p_{11} = \frac{-b + \sqrt{b^2 - 4ac}}{2a}. \quad (3)$$

After obtaining the marginal prevalence of the variables and their co-prevalence,  $p_{11}$ , we can then estimate the expectation, variance of the risk factors and the covariance across them using the following formulas

$$\begin{aligned} E(X_i) &= p_i, \quad i = 1, 2, \\ Var(X_i) &= p_i(1-p_i), \quad i = 1, 2, \\ Cov(X_1, X_2) &= p_{11} - p_1p_2. \end{aligned}$$

For the risk factors that have multiple categories, such as hematological cancer status, we denote dummy variables for the  $j$ th and  $k$ th categories as  $X_j$  and  $X_k$ , respectively. Given marginal probabilities  $p_j$  and  $p_k$ , we can calculate the expectation and variance of  $X_j$  and  $X_k$  based on the properties of multinomial distribution:

$$\begin{aligned} E(X_i) &= p_i \quad i = j, k, \\ Var(X_i) &= p_i(1-p_i), \quad i = j, k, \\ Cov(X_j, X_k) &= -p_jp_k. \end{aligned}$$

We apply these formulas repeatedly to calculate the age-stratified individual and joint prevalence of different risk factors and thus calculate the mean and variance for the risk score for each of the three components of the normal mixture model.

**The IER of a city/county.** We define the quantity  $\bar{R}_l(\beta) = E_l\{\exp(\sum_{k=1}^K \beta_k X_{ik})\}$  as an IER for the population associated with the underlying risk factor distribution in location  $l$ , and present the scaled version of IER as  $\bar{R}_l(\beta)/\bar{R}$ , where  $\bar{R}$  denotes the weighted average of  $\bar{R}_l(\beta)$  across cities/counties with population sizes as the weights, to rank cities/counties for their excess risk due to the underlying distribution of risk factors in the populations. As the individual-level data from each location are unavailable, we estimated IER using the available city- and county-level data for the prevalence of  $X_k$ s, and individual-level data from a representative sample of the US population from the NHIS. Taking city-level analysis as an example, we first obtained the mean and variance of individual-level risk score (RS) within each age group  $a$  ( $\mu_{l,a} = E_{l,a}\{RS\}$  and  $\sigma_{l,a}^2 = V_{l,a}\{RS\}$ , respectively,  $a=1,2,3$ ). Note that IER can be written as  $IER_l(\beta) = E_l\{\exp(RS)\}/\bar{R}$ . Given the mixture-normal assumption for RS, we have

$$E_l\{\exp(RS)\} = \sum_{a=1}^3 p_{l,a} \exp\left(\mu_{l,a} + 0.5\sigma_{l,a}^2\right)$$

where  $p_{l,a}$  denotes the prevalence of age group  $a$  at location  $l$ , which is available from ACS. We defined an average risk of the US population ( $\bar{R}$ ) based on a weighted average of IER across 477 cities with weights proportional to population size. For the county-level analysis of the 65-years-and-older Medicare population, similar calculations were performed based on a single normal distribution for underlying risk scores.

**Model validation.** We correlated IER with recent death rates across US cities to validate the underlying model.

Specifically, we conducted independent validation analyses using county-level mortality information from the CDC between 7 June and 1 October 2020, which did not contribute to the model development. In one analysis, for each of the 259 counties that contained the 477 studied cities, we calculated a weighted IER with each city being weighted by its population size. We then examined how strongly the county-level IER predicted the underlying death rates using two approaches. First, we fitted a negative binomial regression of the death counts on  $\log(IER)$ , where the underlying population sizes were used as offset terms for modeling rates and residual heterogeneity in the model was accounted for using an underlying Poisson-Gamma random effects model. If the underlying individual-level risk model is correctly specified, then in this group-level model, one would expect the slope of  $\log(IER)$  to be close to 1.0. We further modeled the log of death rates across the counties as a linear function of  $\log(IER)$  and used weighted least squares to estimate a measure of explained variance ( $R^2$ ) of  $\log(\text{death rate})$  associated with  $\log(IER)$ . As a benchmark, we estimated similar measures for two other likely predictors,  $\log$  of population density ([https://covid19.census.gov/datasets/21843f238ccb46b08615fc53e19e0daf\\_1?geometry=-168.434%2C28.795%2C169.066%2C67.148&selectedAttribute=B01001\\_calc\\_PopDensity](https://covid19.census.gov/datasets/21843f238ccb46b08615fc53e19e0daf_1?geometry=-168.434%2C28.795%2C169.066%2C67.148&selectedAttribute=B01001_calc_PopDensity)) and  $\log$  of 3-weeks-before infection rate (<https://github.com/CSSSEGISandData/COVID-19>). The 95% CIs for  $R^2$  were calculated based on 1,000 bootstrap replicates of county-level data. We also conducted a conditional analysis by incorporating region indicators for the Northeast, Midwest, South and West as additional covariates in the regression to account for major regional differences in pandemic dynamics between regions ([https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us\\_regdiv.pdf](https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf)).

Additionally, we conducted a validation analysis using a much expanded set of 2,999 counties, but with the IER derived based on the underlying Medicare population aged 65 and older, which was expected to lead to the majority of the observed deaths, as information on risk factor prevalence for the general population is not available across US counties. For each county, we derived a conditional IER based on the risk-factor prevalence for the underlying 65-years-and-older population and then multiplied it by the proportion of population who are 65 years and older in that county to capture its associated risk. All counties with zero death or zero infection over the 2-week period were excluded from the analyses. All analyses were performed using information on deaths over a moving window of 2-week periods for the detection of potential temporal effects. In both analyses, we observed that the relationship between IER and death rates were fairly linear throughout the range of risk.

Further, we use the above calculations to project expected discriminatory performance of the underlying risk model for the US population. We calculated AUC, which is the probability that risk score value is higher for a randomly selected case compared to that for a randomly selected control, based on the

observed distribution of risk scores in the NHIS sample (treated as controls) and the projected distribution of risk scores among cases ( $D = 1$ ).

**Calculating the proportion and size of the vulnerable population within a city/county.** We further examined the distribution of  $R_i(\beta)$  across individuals within a location to identify the size of the underlying most ‘vulnerable’ populations. For these evaluations, ideally one would require individual-level data for a representative sample of individuals from each location. However, in the absence of such data, we developed a framework to approximate the distributions using city/county-specific information on prevalence and information on the co-occurrence of these factors, captured through the underlying odds ratio parameters, from the NHIS (Extended Data Fig. 2 and details below).

We used  $\bar{R}$ , the average risk for the US population associated with these risk factors, as a reference risk. The proportion of individuals who are at  $k$ -fold or higher risk compared to the reference risk can be defined as

$$\Pr(R_{il} > k \times \bar{R}) = \Pr(\log(R_{il}) > \log(k) + \log(\bar{R})).$$

We first discuss calculation of the proportion and number of high-risk individuals among the general 18-years-and-older population in each of the 477 US cities. As mentioned previously, the distribution of the risk scores of the 18-years-and-older US population shows a mixture-normal pattern due to the substantial difference in risk of mortality across different age groups. As shown in Extended Data Fig. 2, the distribution of the risk scores of the general adult population in NHIS population can be approximated by a mixture of three normal distributions that correspond to the 18–44 age group, 45–74 age group and 75+ age group, respectively. We therefore assumed that  $\log(R_{il})$ , the risk score of an individual  $i$  in city  $l$ , follows a mixture-normal distribution of three components,  $N(\mu_{l,a}, \sigma_{l,a}^2)$  with weight  $p_{l,a}$  being to the prevalence of age group  $a$ ,  $a = 1, 2, 3$ , in city  $l$ . The proportion of the population in city  $l$  that exceed  $k$ -fold of the reference risk is then

$\Pr(R_{il} > k \times \bar{R}) = 1 - \text{CDF}_{N(\mu_{l,a}, \sigma_{l,a}^2), p_{l,a}}, a=1,2,3 (\log\{k\} + \log\{\bar{R}\})$ , where  $\text{CDF}_{N(\mu_{l,a}, \sigma_{l,a}^2), p_{l,a}}, a=1,2,3$  denotes the cumulative density function (CDF) of the assumed mixture-normal distribution for the individual-level risk score in city  $l$ . Finally, the actual size of the vulnerable population that exceed  $k$ -fold of average risk in city  $l$  is estimated by multiplying the proportion of vulnerable population by total population size  $M_l$ , that is  $M_l(1 - \text{CDF}_{N(\mu_{l,a}, \sigma_{l,a}^2), p_{l,a}}, a=1,2,3 (\log\{k\} + \log\{\bar{R}\}))$ .

For the 65-years-and-older Medicare population, we expected that the risk score  $R_{il}$ s has an approximately normal distribution, which was validated using the NHIS 65-years-and-older individuals (Extended Data Fig. 2). Under the normal assumption,  $R_{il} \sim N(\mu_l, \sigma_l^2)$ , we estimated the mean  $\mu_l$  and variance  $\sigma_l^2$  of the risk score  $R_{il}$  using previously described methods and the proportion of the 65-years-and-older Medicare population in location  $l$  that exceed  $k$ -fold of the reference risk is

$$\Pr(R_{il} > k \times \bar{R}) = 1 - \Phi_{\mu_l, \sigma_l^2}(\log\{k\} + \log\{\bar{R}\}),$$

where  $\Phi_{\mu, \sigma^2}(\cdot)$  denotes the CDF of  $N(\mu, \sigma^2)$ .

**Expected proportion of deaths at high risk within a location.** As validation analysis show that the underlying model is well calibrated for the US population, the distribution of risk score among individuals who are expected to die can be derived based on that for the general population and the underlying risk model. Thus, we make a series of projections for the proportion of total deaths that are expected to occur within various risk categories based on the model. Specifically, we calculate the proportion of deaths expected to occur at or higher than the  $k$ -fold risk threshold as

$$\Pr(R_{il} > k \times \bar{R} | D_{il} = 1) = \frac{\Pr(D_{il} = 1, R_{il} > k \times \bar{R})}{\Pr(D_{il} = 1, R_{il} > k \times \bar{R}) + \Pr(D_{il} = 1, R_{il} \leq k \times \bar{R})},$$

where  $\Pr(D_{il} = 1 | R_{il})$ , the probability of death given risk  $R_{il}$  is proportional to  $R_{il}$ .

For the city-level analysis of the general 18-years-and-older population in the 477 US cities, we have

$$\begin{aligned} \Pr(D_{il} = 1, R_{il} > k \times \bar{R}) &= \int_{R_{il} > k \times \bar{R}} \Pr(D_{il} = 1 | R_{il}) f(R_{il}) dR_{il} \\ &= \sum_{j=1}^3 \left\{ \int_{\log(k \times \bar{R})}^{\infty} \Pr(D_{il} = 1 | RS_{il}) f(RS_{il} | a_{il} = j) dRS_{il} \right\} \times \Pr(a_{il} = j) \\ &\propto \sum_{j=1}^3 p_{l,j} \int_{\log(k \times \bar{R})}^{\infty} e^{RS_{il}} e^{-\frac{(RS_{il} - \mu_{l,j})^2}{2\sigma_{l,j}^2}} dRS_{il} \\ &\propto \sum_{j=1}^3 p_{l,j} e^{\mu_{l,j}^2 + 0.5\sigma_{l,j}^2} \left[ 1 - \Phi_{\mu_{l,j} + \sigma_{l,j}^2, \sigma_{l,j}^2}(\log(k \times \bar{R})) \right], \end{aligned}$$

similarly,

$$\Pr(D_{il} = 1, R_{il} \leq k \times \bar{R}) \propto \sum_{j=1}^3 p_{l,j} e^{\mu_{l,j}^2 + 0.5\sigma_{l,j}^2} \Phi_{\mu_{l,j} + \sigma_{l,j}^2, \sigma_{l,j}^2}(\log(k \times \bar{R})),$$

where  $RS_{il} = \log(R_{il})$ ,  $p_{l,j}$  denotes the prevalence of age group  $j$ ,  $a_{il}$  is the indicator of age group ( $a_{il} = 1$ : age 15–44,  $a_{il} = 2$ : age 45–74,  $a_{il} = 3$ : age 75+). The proportion of deaths in location  $l$  that exceed  $k$ -fold of the reference risk is then

$$\begin{aligned} \Pr(R_{il} > k \times \bar{R} | D_{il} = 1) &= \frac{\sum_{j=1}^3 p_{l,j} e^{\mu_{l,j}^2 + 0.5\sigma_{l,j}^2} \left[ 1 - \Phi_{\mu_{l,j} + \sigma_{l,j}^2, \sigma_{l,j}^2}(\log(k \times \bar{R})) \right]}{\sum_{j=1}^3 p_{l,j} e^{\mu_{l,j}^2 + 0.5\sigma_{l,j}^2} \left[ 1 - \Phi_{\mu_{l,j} + \sigma_{l,j}^2, \sigma_{l,j}^2}(\log(k \times \bar{R})) \right] + \sum_{j=1}^3 p_{l,j} e^{\mu_{l,j}^2 + 0.5\sigma_{l,j}^2} \Phi_{\mu_{l,j} + \sigma_{l,j}^2, \sigma_{l,j}^2}(\log(k \times \bar{R}))} \end{aligned}$$

For the county-level analysis of the 65-years-and-older Medicare population in the 3,113 US counties, we have

$$\begin{aligned} \Pr(R_{il} > k \times \bar{R} | D_{il} = 1) &\propto \int_{R_{il} > k \times \bar{R}} \Pr(D_{il} = 1 | R_{il}) f(R_{il}) dR_{il} \\ &= \int_{\log(k \times \bar{R})}^{\infty} \Pr(D_{il} = 1 | RS_{il}) f(RS_{il}) dRS_{il} \\ &= 1 - \Phi_{\mu_l + \sigma_l^2, \sigma_l^2}(\log(k \times \bar{R})), \end{aligned}$$

where  $\mu_l$  and  $\sigma_l^2$  are the mean and variance of the risk score  $RS_{il}$  in county  $l$ .

**Uncertainty in community-level projections.** We provided estimates of uncertainties for community-level projections, taking into account various sources of uncertainties in the estimates of risk parameters and random variations associated with various survey datasets (such as BRFSS, NHANES, NHIS). We created bootstrap samples for all inputs that went into the risk calculations, including parameter values, risk factor prevalence and survey datasets and carried out all the calculations on these replicated datasets as we did for our original analyses. We report 95% empirical CI for the final risk estimate for all community-level projections based on their underlying bootstrap distributions.

We first simulated 1,000 bootstrap replicates for the set of coefficients for the UK OpenSAFELY model by simulating them from a multivariate normal distribution with mean and dispersion matrix fixed at the set of original parameter estimates and underlying variance-covariance matrix (obtained by personal communication), respectively. We created 1,000 bootstrap replicates for the NHIS and NHANES datasets by sampling subjects with replacement. For the BRFSS study, for which individual-level data were not available, we obtained bootstrap replicates for the estimates of the risk factor prevalence through simulations. Let  $X_{il}$  denote the indicator variable for a risk factor category for individual  $i$  in location  $l$ . We simulated  $X_{il}$  from  $Ber(np_l)$ , where  $p_l$  denotes the original estimate of risk-factor prevalence in location  $l$  reported in BRFSS and  $n_l$  denote the sample size in BRFSS for location  $l$ . As the city-wise sample sizes were not reported in BRFSS, we used the total sample size across cities within a state and distributed it to the underlying cities proportionately to their population sizes. We did not consider uncertainty in the estimates associated with the risk factors: sex, ethnicity and age, as the information came from the entire US population collected through census data.

**Uncertainty in individual-level risk estimation.** We also provided estimates of uncertainties for individual-level projections. For projections of individual-level relative risk, we took into account uncertainty in the estimates of risk parameters. For the projections of individual-level absolute risk, we further incorporated uncertainties associated with state-level projections across up to 61 different forecasting models included in the Ensemble estimator (<https://covid19forecasthub.org/doc/ensemble/>, <https://github.com/reichlab/covid19-forecast-hub/ensemble-model>)<sup>30</sup>.

Denoting  $X$  as the vector of risk factors of an individual,  $\hat{\beta}$  as the vector of the estimated fully adjusted log hazard ratios of the risk factors and  $\bar{R}(\hat{\beta}) = (\sum_{l=1}^L w_l \bar{R}_l(\hat{\beta})) / (\sum_{l=1}^L w_l)$  the weighted average of the estimated mean risk in each city (that is  $\bar{R}_l(\hat{\beta})$ ), with weight  $w_l$  denoting the population size of city  $l$ , the estimated relative risk of the individual is reported as

$$\widehat{RR} = \frac{e^{X^T \hat{\beta}}}{\bar{R}(\hat{\beta})}.$$

The sources of uncertainty in  $\widehat{RR}$  include the variation in  $\hat{\beta}$ ,  $\bar{R}(\hat{\beta})$  and the covariance of  $\hat{\beta}$  and  $\bar{R}(\hat{\beta})$ . We quantified these sources of uncertainty by constructing a 95% bootstrap CI for  $\widehat{RR}$ . Specifically, we have

$$\begin{aligned} \text{Var}\{\log(\widehat{RR})\} &= \text{Var}[X_i^T \hat{\beta} - \log(\bar{R}(\hat{\beta}))] \\ &= X^T \text{Cov}(\hat{\beta}) X + \text{Var}[\log(\bar{R}(\hat{\beta}))] - 2X^T \text{Cov}[\hat{\beta}, \log(\bar{R}(\hat{\beta}))], \end{aligned}$$

where  $\text{Cov}(\hat{\beta})$ ,  $\text{Var}[\log(\bar{R}(\hat{\beta}))]$  and  $\text{Cov}[\hat{\beta}, \log(\bar{R}(\hat{\beta}))]$  are estimated from 1,000 bootstrap samples of  $\hat{\beta}$ , city-level BRFSS risk factor prevalence and NHIS and NHANES datasets (described above).

The 95% CI of an individual's relative risk is computed as

$$\left[ e^{\log \widehat{RR} - z_{1-\frac{\alpha}{2}} \times \text{sd}\{\log(\widehat{RR})\}}, e^{\log \widehat{RR} + z_{1-\frac{\alpha}{2}} \times \text{sd}\{\log(\widehat{RR})\}} \right],$$

where  $\alpha = 0.05$  and  $z_{1-\frac{\alpha}{2}}$  denotes the  $(1 - \frac{\alpha}{2})$  quantile of the standard normal distribution.

Recall that the estimate of an individual's absolute risk is calculated by combining the estimated relative risk that quantifies the risk due to a set of predisposing factors with the estimated baseline risk in the individual's state of residence that quantifies the risk due to the underlying pandemic characteristics over time:

$$\hat{\lambda}_s(X, t) = \hat{\lambda}_s(t)e^{X^T\hat{\beta}} \\ = \frac{\hat{\lambda}_s^A(t)}{R_s(\hat{\beta})} e^{X^T\hat{\beta}},$$

where  $\hat{\lambda}_s^A(t)$  and  $R_s(\hat{\beta})$  denote the projected mortality rate and the average relative risk of the population, respectively, in state  $s$  (the individual's state of residence). Taking log on both sides of the above equation, we get  $\log \hat{\lambda}_s(X, t) = \log \hat{\lambda}_s^A(t) - \log R_s(\hat{\beta}) + X^T\hat{\beta}$ . We assume that  $\log \hat{\lambda}_s(X, t)$  is approximately normally distributed and derive its variability as

$$\begin{aligned} \text{Var}\{\log \hat{\lambda}_s(X, t)\} &= \text{Var}\{\log \hat{\lambda}_s^A(t) - \log R_s(\hat{\beta}) + X^T\hat{\beta}\} \\ &= \text{Var}\{\log \hat{\lambda}_s^A(t)\} + \text{Var}\{\log R_s(\hat{\beta})\} + \text{Var}(X^T\hat{\beta}) \\ &\quad - 2\text{Cov}(\log \hat{\lambda}_s^A(t), \log R_s(\hat{\beta})) \\ &\quad - 2\text{Cov}(X^T\hat{\beta}, \log R_s(\hat{\beta})) + 2\text{Cov}(\log \hat{\lambda}_s^A(t), X^T\hat{\beta}) \\ &= \text{Var}\{\log R_s(\hat{\beta})\} + \text{Var}\{\log \hat{\lambda}_s^A(t)\} + X^T\text{Cov}(\hat{\beta}, X) \\ &\quad + 2X^T\text{Cov}(\hat{\beta}, \log(R_s(\hat{\beta}))). \end{aligned}$$

The other two covariance terms are zero as  $\hat{\lambda}_s^A(t)$  and  $\hat{\beta}$  are independent. Here,  $\text{Cov}(\hat{\beta})\text{Var}(\log R_s(\hat{\beta}))$  and  $\text{Cov}(\hat{\beta}, \log R_s(\hat{\beta}))$  are estimated from 1,000 bootstrap samples as described above. We estimate  $\text{Var}\{\log \hat{\lambda}_s^A(t)\}$  based on the variation of death reported across the 58 models included in the Ensemble estimator<sup>30</sup>. Specifically, based on the reported CI  $([MR_L, MR_U])$  and normally, we calculate

$$\text{Var}\{\log \hat{\lambda}_s^A(t)\} = \left\{ \frac{\log(MR_U) - \log(MR_L)}{2(z_{1-\frac{\alpha}{2}})} \right\}^2.$$

The 100(1 -  $\alpha$ )% CI of an individual's absolute risk is then finally computed as

$$\left[ e^{\log \hat{\lambda}_s(X, t) - z_{1-\frac{\alpha}{2}} \times \text{sd}(\log \hat{\lambda}_s(X, t))}, e^{\log \hat{\lambda}_s(X, t) + z_{1-\frac{\alpha}{2}} \times \text{sd}(\log \hat{\lambda}_s(X, t))} \right].$$

**Reporting Summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

## Data availability

Interactive maps for viewing city-, county-, state- and national-level risk projections in the United States (with 95% CIs provided) and the web-based tool for the individualized risk calculator are available at <http://covid19risktools.com/>. All data used in the manuscript are publicly available, including NHIS 2017 Data Release ([https://www.cdc.gov/nchs/nhis/nhis\\_2017\\_data\\_release.htm](https://www.cdc.gov/nchs/nhis/nhis_2017_data_release.htm)), deaths associated with different age and racial/ethnic groups in the United States (as of 6 June 2020, <https://data.cdc.gov/NCHS/Deaths-involving-coronavirus-disease-2019-COVID-19/ks3g-spdg>), 2017 Leading Cancer Cases and Deaths (<https://gis.cdc.gov/Cancer/USCS/DataViz.html>), US Census Bureau ACS 1-year estimates for age, gender and race, 2017/2018 table (<https://www.census.gov/programs-surveys/acs>), US Census Bureau ACS 2017 age-sex table (<https://data.census.gov/cedsci/table?q=Age%20and%20Sex&hidePreview=true&t=Age%20and%20Sex&tid=ACSST1Y2017.S0101&vintage=2018&y=2017>), 2010–2019 state population by characteristics (<https://www.census.gov/data/tables/time-series/demo/popest/2010s-state-detail.html>), 2010–2019 county population

by characteristics (<https://www.census.gov/data/tables/time-series/demo/popest/2010s-counties-detail.html>), US Census Bureau ACS sample size (<https://www.census.gov/acs/www/methodology/sample-size-and-data-quality/sample-size/>), 2018 city-wide information on Hispanic or Latino origin by race (<https://data.census.gov/cedsci/table?q=hispanic&hidePreview=true&tid=ACSDT1Y2018.B03002&t=Hispanic%20or%20Latino&vintage=2018>), 500 Cities: Local Data for Better Health, 2019 release, BRFSS, CDC (<https://chronicdata.cdc.gov/500-Cities/500-Cities-Local-Data-for-Better-Health-2019-release/6vp6-wxuq>), BRFSS 2017 Summary Data Quality Report ([https://www.cdc.gov/brfss/annual\\_data/2017/pdf/2017-sdqr-508.pdf](https://www.cdc.gov/brfss/annual_data/2017/pdf/2017-sdqr-508.pdf)) and 2017–2018 NHANES (<https://wwwn.cdc.gov/nhanes/continuousnhanes/default.aspx?BeginYear=2017>). All data used in the analyses can be accessed at <https://github.com/nchatterjeelab/COVID19Risk/tree/master/data>.

## Code availability

The R codes for data management and analyses in this article can be accessed at <https://github.com/nchatterjeelab/COVID19Risk>.

## References

29. Kundu, P., Tang, R. & Chatterjee, N. Generalized meta-analysis for multiple regression models across studies with disparate covariate information. *Biometrika* **106**, 567–585 (2019).
30. Ray, E. L. et al. Ensemble forecasts of coronavirus disease 2019 (COVID-19) in the US. Preprint at *medRxiv* <https://doi.org/10.1101/2020.08.19.20177493> (2020).

## Acknowledgements

We thank A. Meisner from the John Hopkins University, Biostatistics Department and M. García-Closas from the Division of Cancer Epidemiology and Genetics at the NCI for their comments on a previous version of the manuscript. This research was funded by the Bloomberg Distinguished Professorship endowment.

## Author contributions

J.J., N.A., P.K. and N.C. developed all methods. J.J., N.A., P.K. and Y.Z. conducted the data analyses. B.H. developed the web-based tool with assistance from J.J., N.A., P.K. and N.C. E.W. developed the interactive maps. E.W. created the maps with PolicyMap mapping tools. N.C., J.J., N.A. and P.K. wrote the first draft of the manuscript. All authors reviewed the final manuscript.

## Competing interests

E.W. is an employee of the business corporation PolicyMap, Inc. She was not involved in the design or analysis of the study.

## Additional information

**Extended data** is available for this paper at <https://doi.org/10.1038/s41591-020-01191-8>.

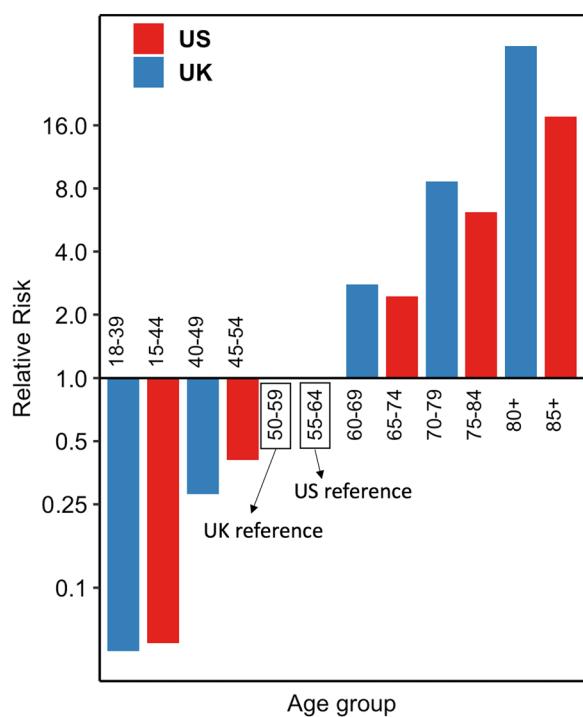
**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41591-020-01191-8>.

**Correspondence and requests for materials** should be addressed to N.C.

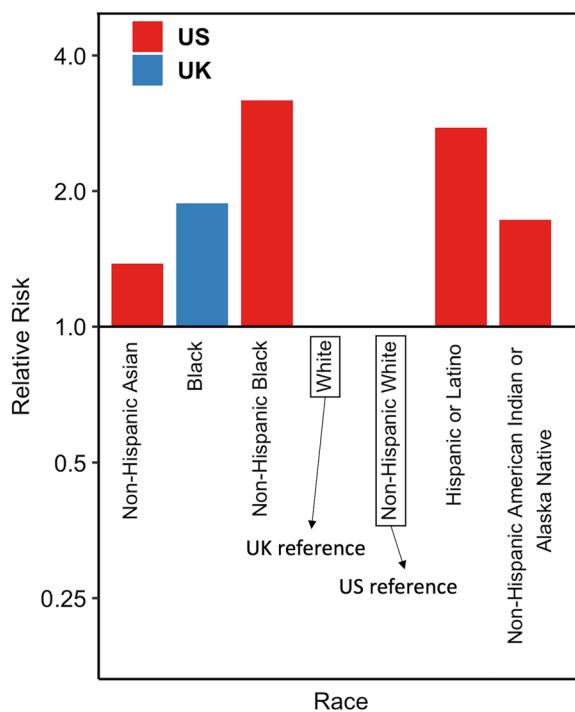
**Peer review information** Jennifer Sargent was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

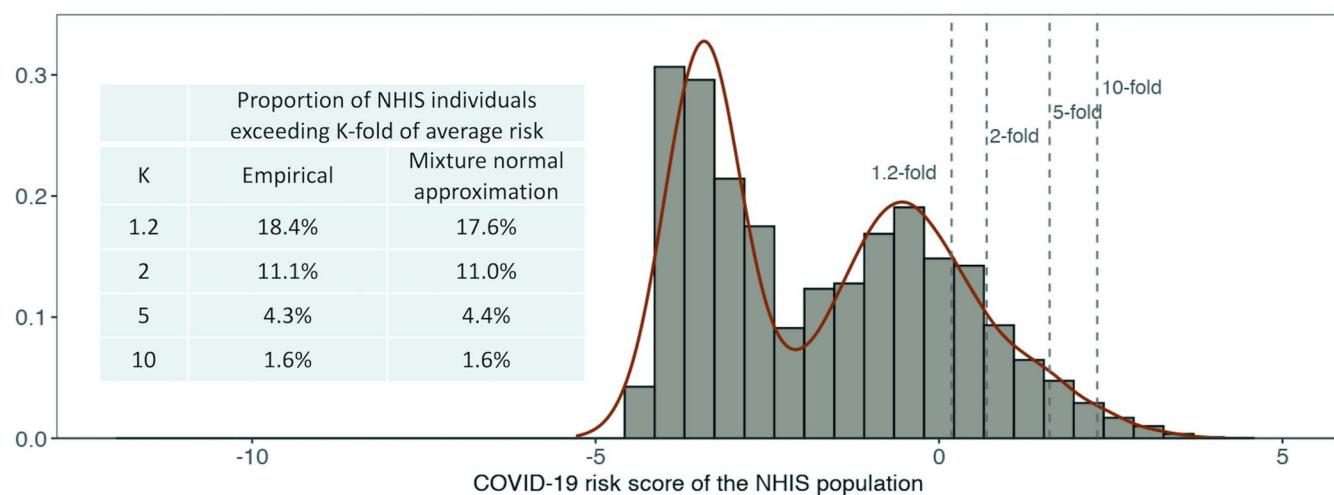
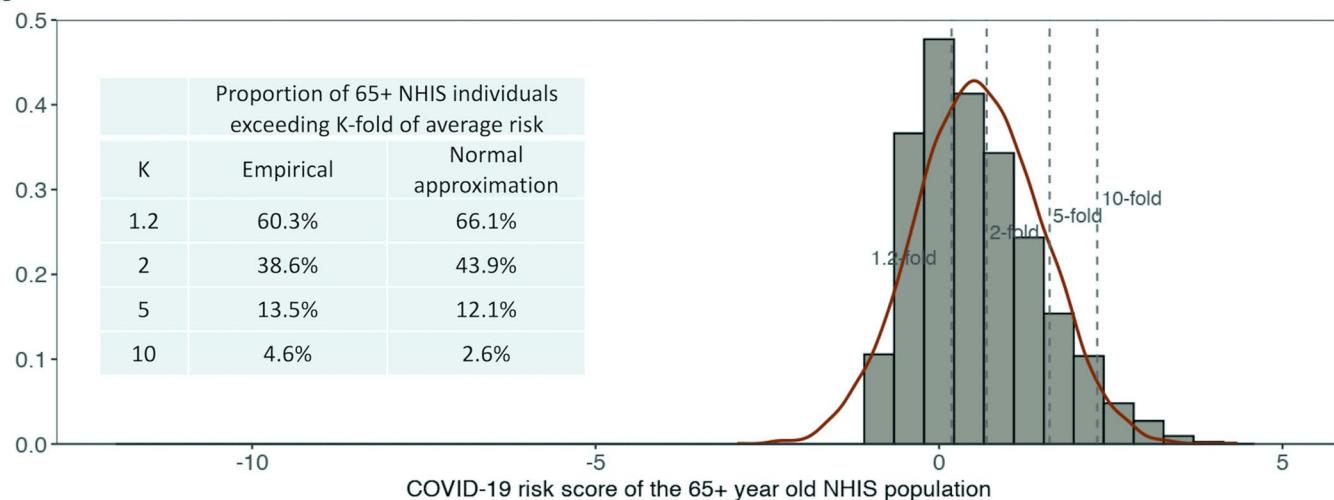
a



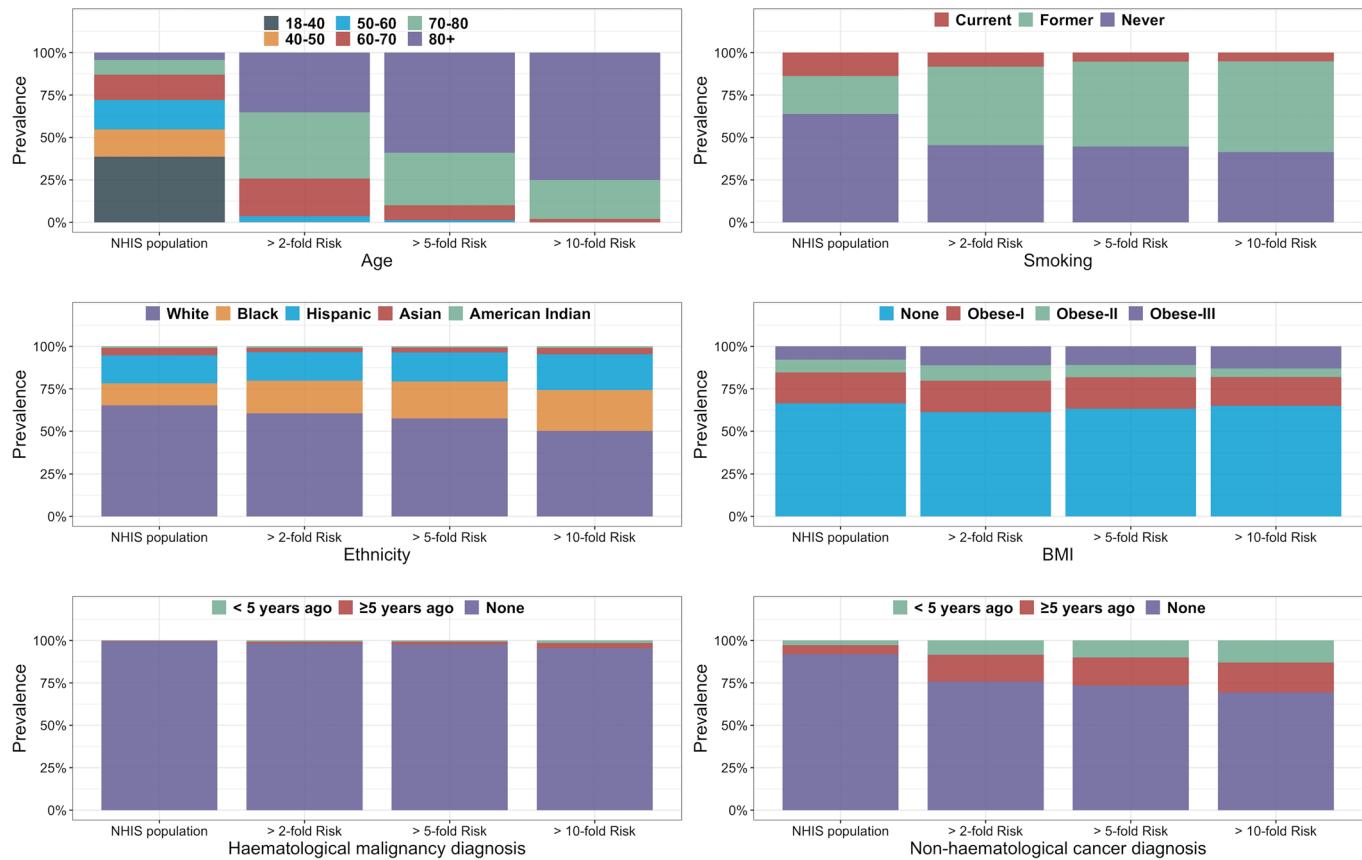
b


**Extended Data Fig. 1 | Comparison of COVID-19 mortality risk associated with various age and race and/or ethnic groups between the US and the UK.**

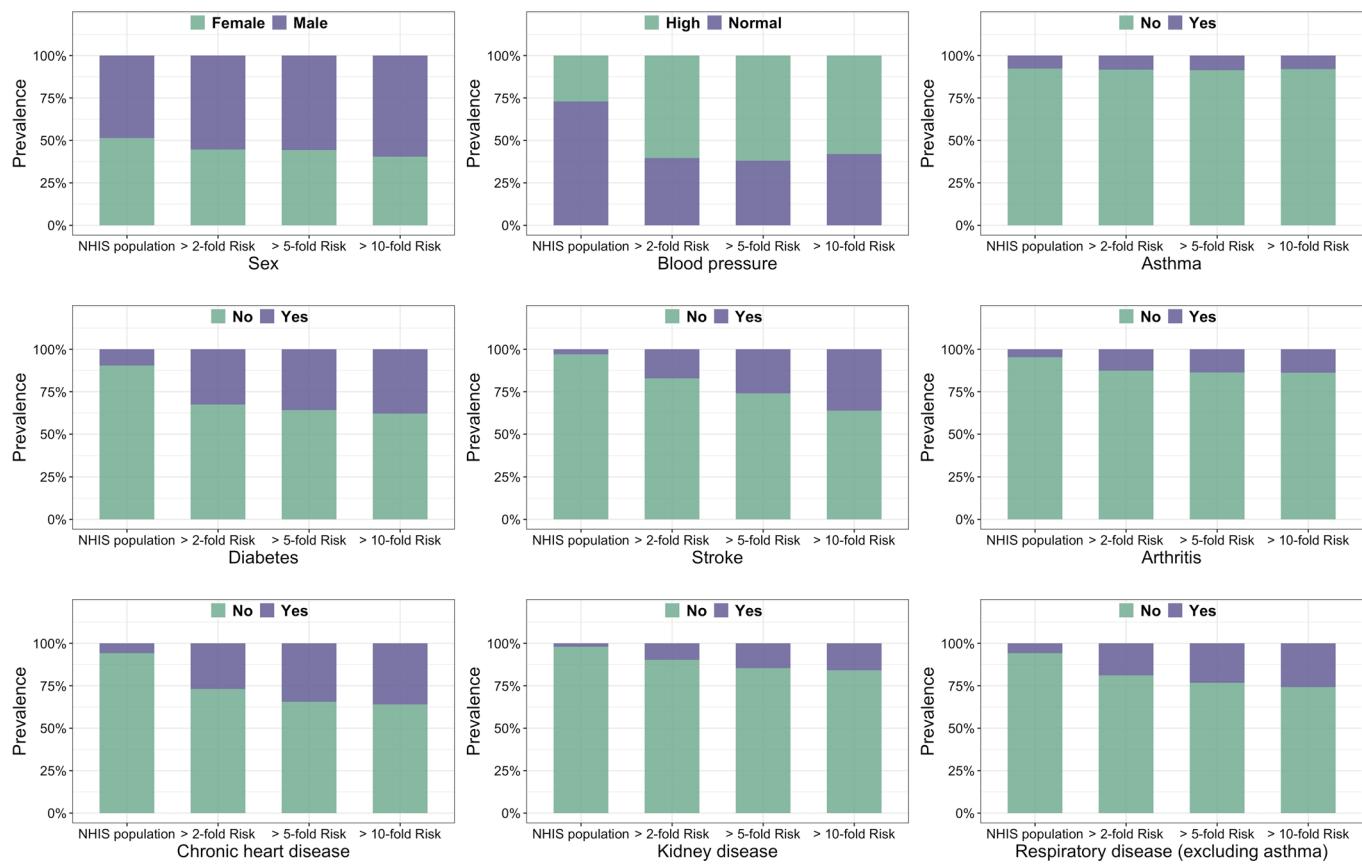
**a**, Relative risk of age groups in the UK versus that of in the US. **b**, Relative risk of ethnic groups in the UK versus that of in the US. For UK, the relative risk is from age-sex adjusted model in Table 1 of the UK OpenSAFELY study. For US, the relative risk associated with race and/or ethnic groups is adjusted for both age and state whereas the relative risk associated with age is adjusted for state using Poisson regression model fitted to the CDC state-level death count data. The UK estimates are based on study population of 17,278,392 adults with 10,926 COVID-19-related deaths whereas the estimates for US are based on the whole population across 51 states with a total of 99,866 deaths reported prior to 7 June 2020.

**a****b**

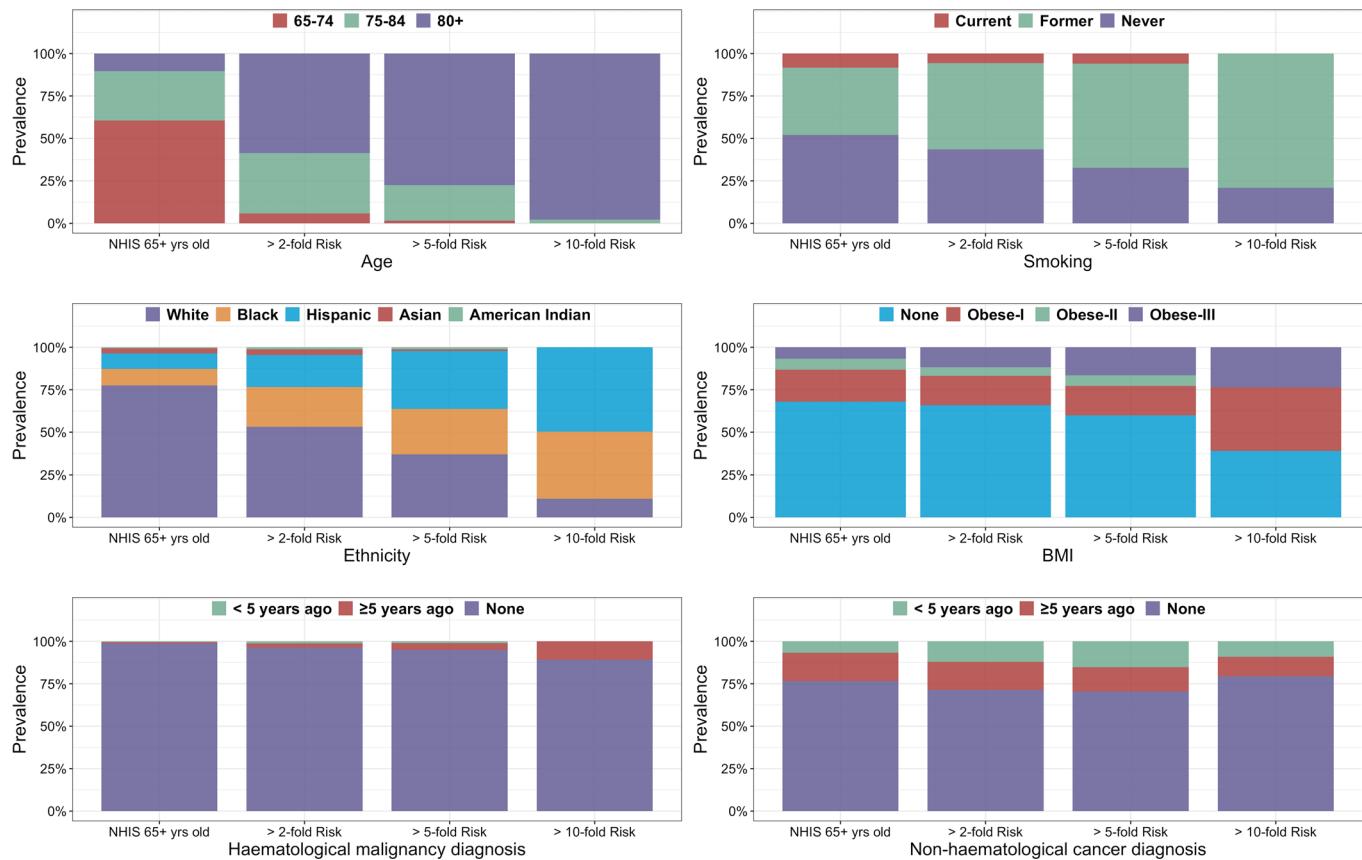
**Extended Data Fig. 2 | Distribution of risk score in NHIS population.** Results are shown separately for (a) the 18 years and older population ( $N=22,901$ ), and (b) the 65 years and older population ( $N=5,875$ ). Empirical distributions are compared with those based on mixture-normal or normal approximations. The risk scores in both sub-figures are calculated based on age, gender, ethnicity and 12 different health conditions, but not social deprivation index (SDI) due to the absence of the relevant data in NHIS. The risk score is centered using a reference value that corresponds to the average risk across the individuals in NHIS. NHIS: National Health Interview Survey.



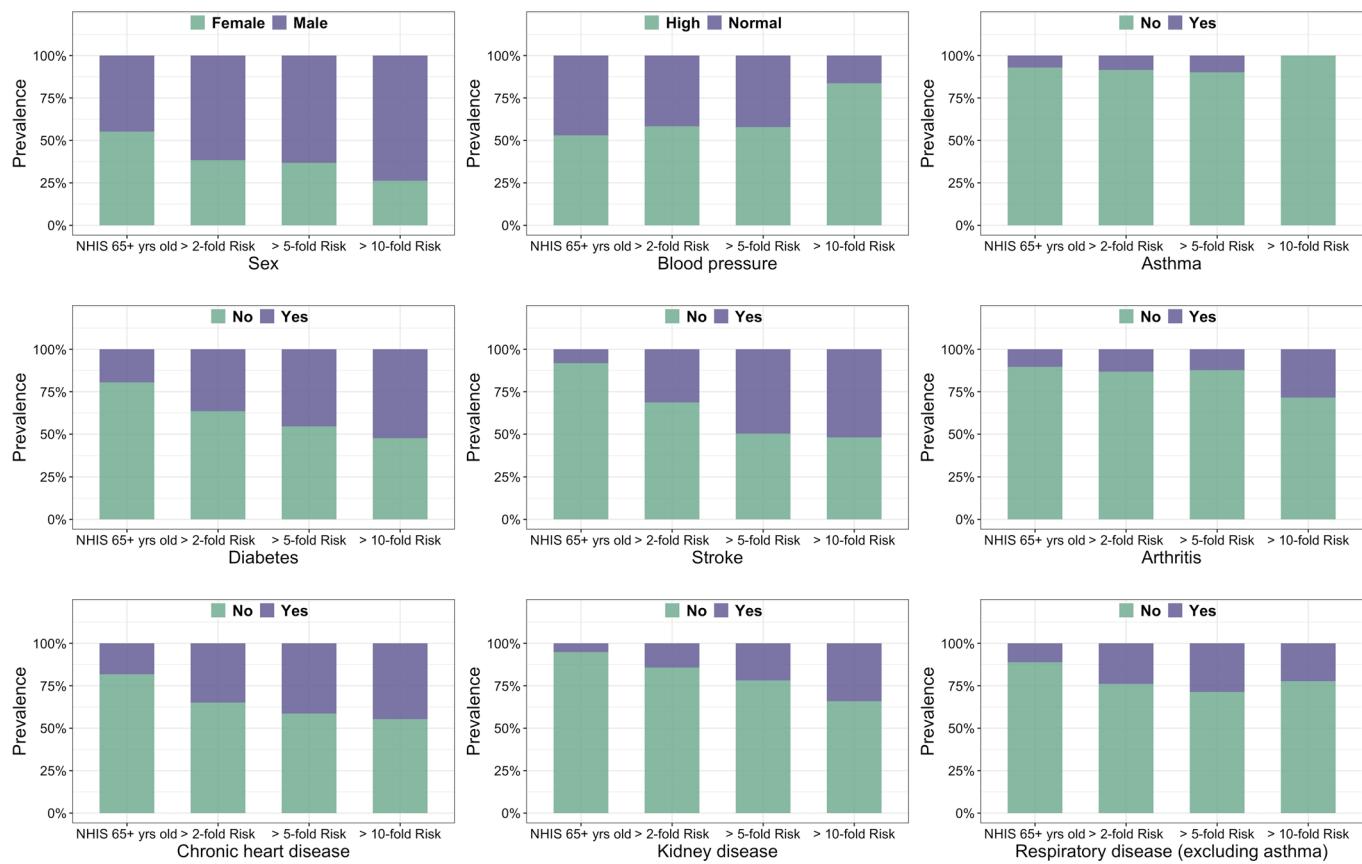
**Extended Data Fig. 3 | Distribution of risk factors in the general NHIS population and among individuals in different risk groups.** The risk score is calculated based on age, sex, race/ethnicity, body mass index (BMI), smoking status and 12 different health conditions. Social Deprivation Index (SDI) was not available in NHIS and was excluded in this analysis. The risk thresholds are defined with respect to the average risk of the NHIS population. NHIS: National Health Interview Survey.



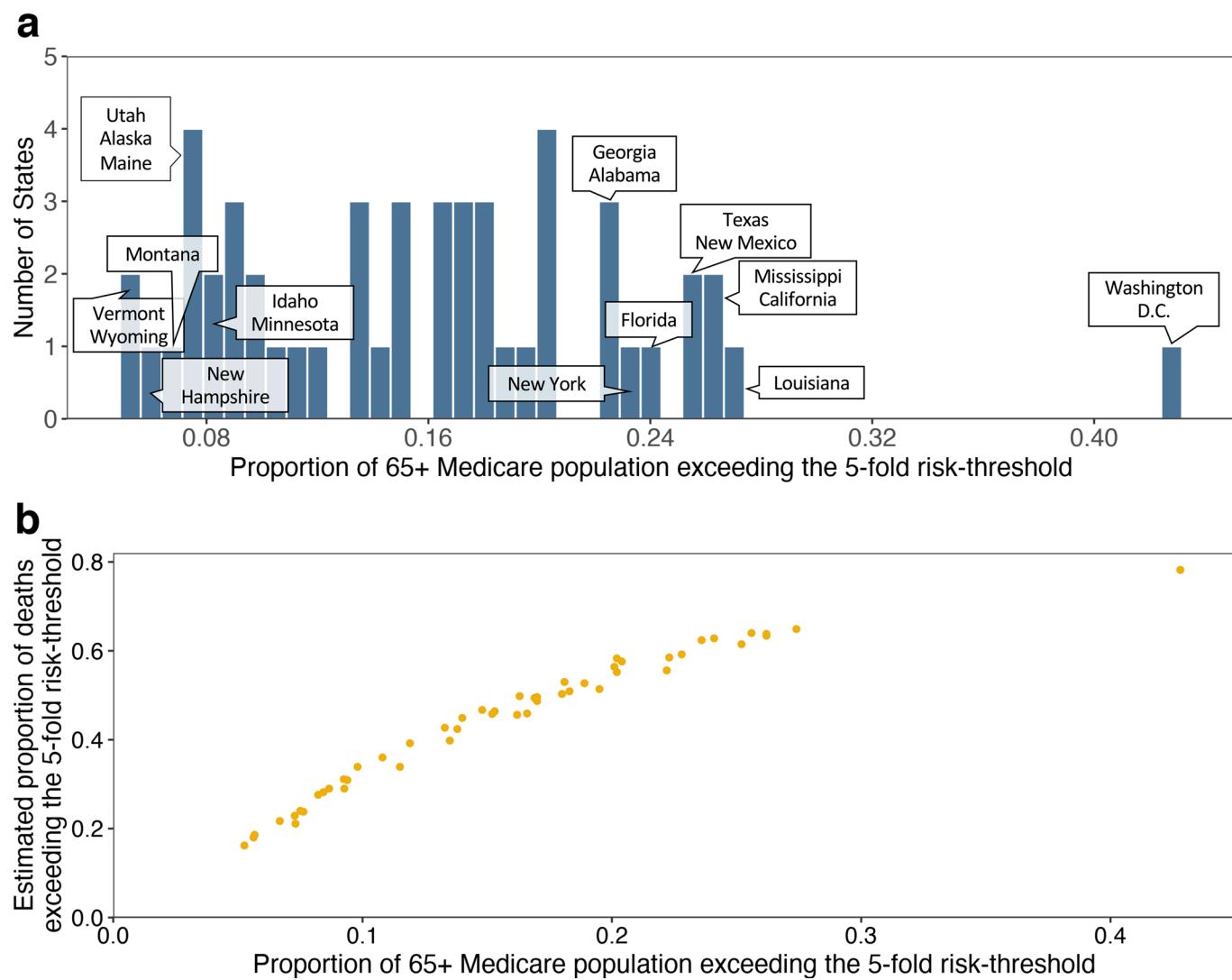
**Extended Data Fig. 4 | Distribution of risk in the general NHIS population and among individuals in different risk groups (continued).** The risk scores were calculated based on age, sex, race/ethnicity, body mass index (BMI), smoking status and 12 different health conditions. Social Deprivation Index (SDI) was not available in NHIS and was excluded in this analysis. The risk thresholds are defined with respect to the average risk of the NHIS population. NHIS: National Health Interview Survey.



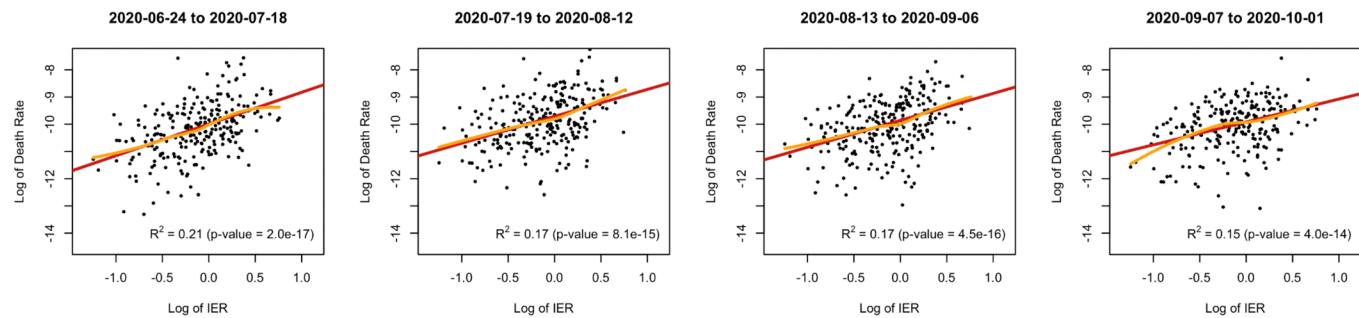
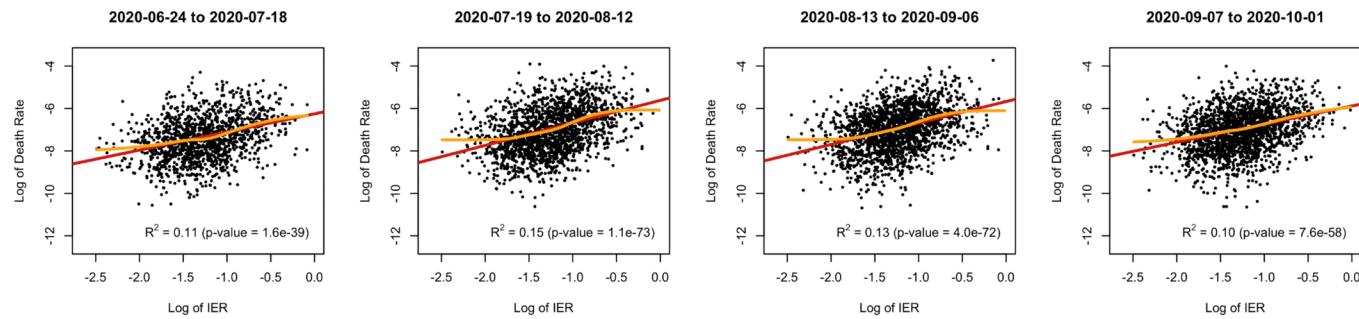
**Extended Data Fig. 5 | Distribution of risk factors in the 65 years and older NHIS population and among individuals in different risk groups.** The risk scores were calculated based on age, sex, race/ethnicity, body mass index (BMI), smoking status and 12 different health conditions. Social Deprivation Index (SDI) was not available in NHIS and was excluded in this analysis. The risk thresholds are defined with respect to the average risk of the NHIS population. NHIS: National Health Interview Survey.


**Extended Data Fig. 6 | Distribution of risk factors in the 65 years and older NHIS population and among individuals in different risk groups (continued).**

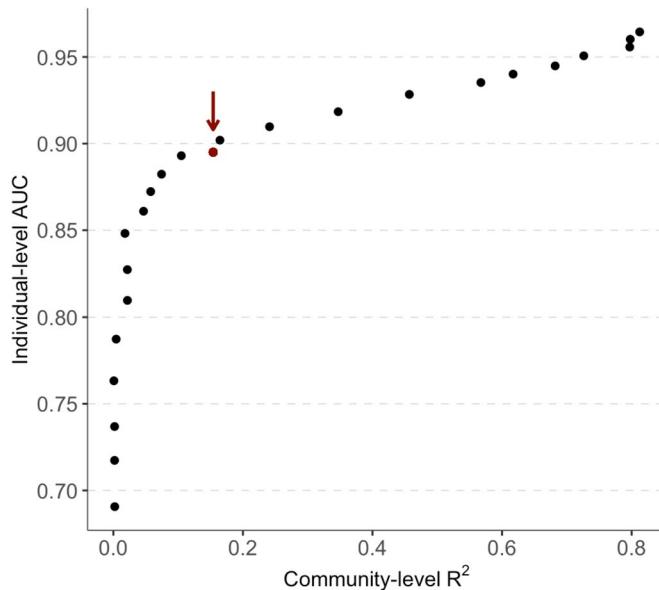
The risk scores were calculated based on age, sex, race/ethnicity, body mass index (BMI), smoking status and 12 different health conditions. Social Deprivation Index (SDI) was not available in NHIS and was excluded in this analysis. The risk thresholds are defined with respect to the average risk of the NHIS population. NHIS: National Health Interview Survey.



**Extended Data Fig. 7 | Projections for high-risk for the 65 years and older Medicare population across N=51 US states.** **a**, Histogram of the proportion of population exceeding the 5-fold risk threshold across states. **b**, Scatter plot of the proportion of population exceeding the fivefold risk threshold against the proportion of deaths among the population that are expected to occur within the  $\geq$ fivefold risk group. Results for additional risk thresholds and the corresponding 95% CIs are provided in Supplementary Table 5.

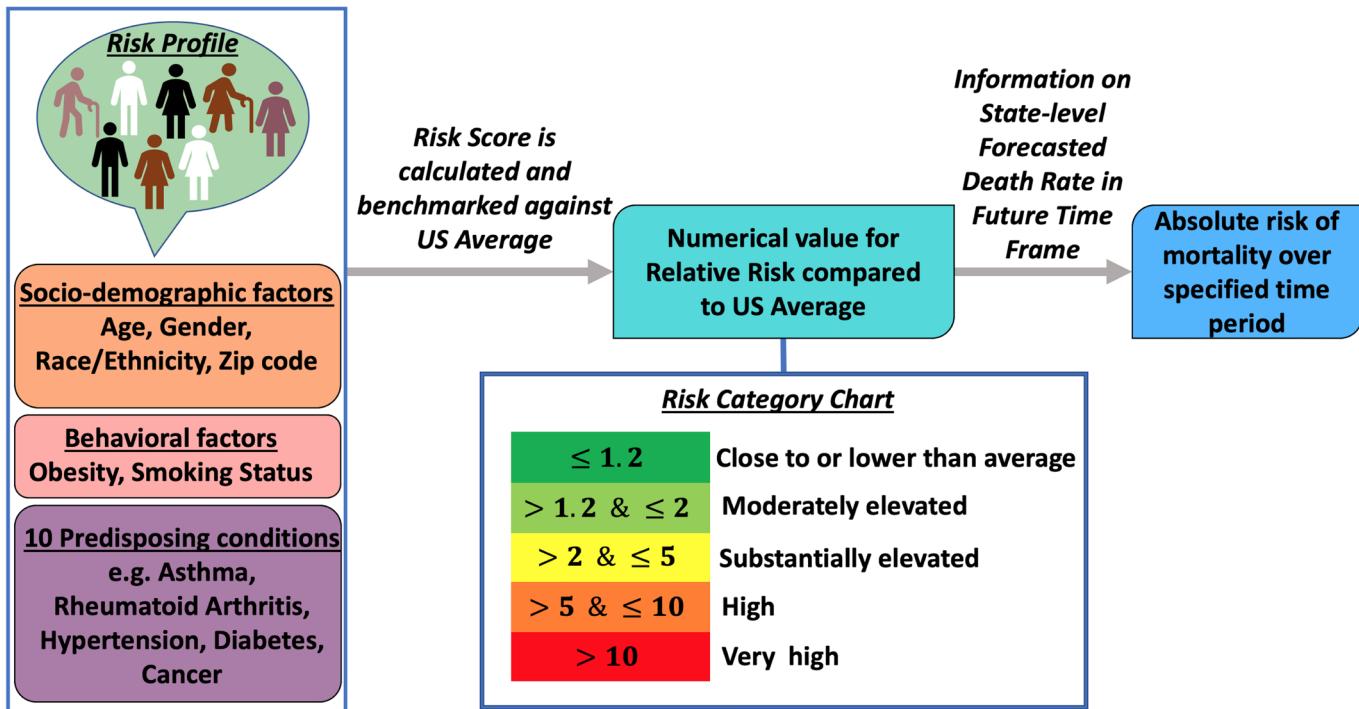
**a. 18+ year old population in 259 US counties representing 477 cities****b. 65+ year old Medicare population in 2999 US counties**

**Extended Data Fig. 8 | Scatter plots of the log of death rate against log(IER).** **a**, IER derived for the 18 years and older population is plotted against observed death rates on the log scale across N=259 US counties representing the 477 cities. **b**, IER derived for the 65 years and older Medicare population is plotted against observed death rates across N=2,999 US counties. The orange curve is fitted by loess regression. IER: Index of Excess Risk.

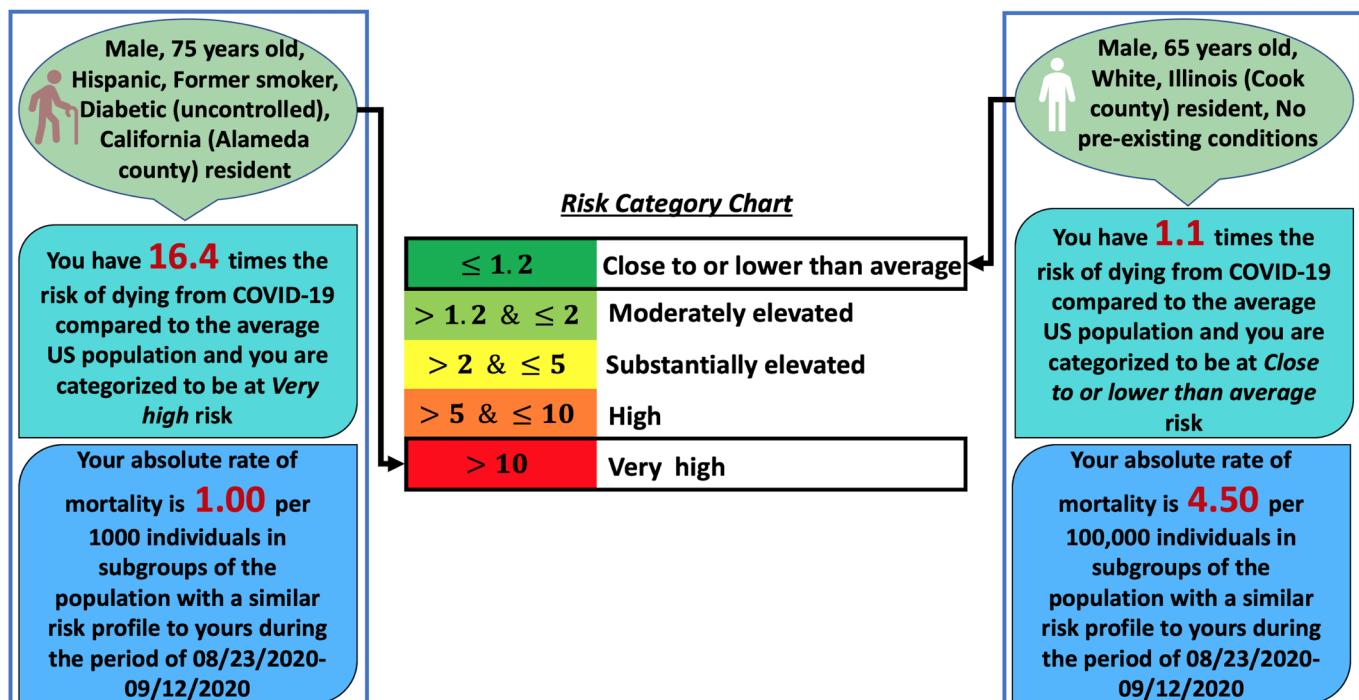


**Extended Data Fig. 9 | Relationship between  $R^2$  for predicting group-level risk and AUC for individual-level prediction observed on simulated data sets.** We simulated individual-level outcome and risk-factor data for approximately 4.1 million individuals based on the risk model, and randomly divided the population into 100 risk groups. We varied the coefficient of the risk-score in the model to achieve different strengths of association between the risk-score and the outcome. The dark red colored plot corresponds to a value of  $R^2=0.154$ , which is the average value of performance of the model predicting death rates across 259 counties representing the 477 studied cities over two-week windows between 7 June 2020 and 1 October 2020, and  $AUC=0.895$ .

## a Schema of Risk Calculator



## b Illustration using two specific profiles



**Extended Data Fig. 10 | Risk calculator workflow.** **a**, General schema of the risk calculator which inputs information on socio-demographic, behavioural, and predisposing conditions of an individual to estimate their relative risk compared to the average risk of the US adult population (aged 18 years and older). Based on the projected death rate in the state where the individual resides in, the tool evaluates the individual's absolute risk of death due to COVID-19 during future time frame. **b**, Output from the risk calculator for two hypothetical profiles.

## Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

### Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- The exact sample size ( $n$ ) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided  
*Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g.  $F$ ,  $t$ ,  $r$ ) with confidence intervals, effect sizes, degrees of freedom and  $P$  value noted  
*Give P values as exact values whenever suitable.*
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's  $d$ , Pearson's  $r$ ), indicating how they were calculated

*Our web collection on [statistics for biologists](#) contains articles on many of the points above.*

### Software and code

Policy information about [availability of computer code](#)

Data collection

The R codes for data management in this article can be accessed at <https://github.com/nchatterjeelab/COVID19Risk>. Versions used were R v3.6.0 and v3.5.2. Interactive maps for viewing the city-, county-, state- and national-level risk projections in the US and the web-based tool for the individualized risk calculator are available at <http://covid19risktools.com/>.

Data analysis

The R codes for data analyses in this article can be accessed at <https://github.com/nchatterjeelab/COVID19Risk>. Versions used were R v3.6.0 and v3.5.2. Interactive maps for viewing the city-, county-, state- and national-level risk projections in the US and the web-based tool for the individualized risk calculator are available at <http://covid19risktools.com/>.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

### Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All data used in the manuscript are publicly available and can be accessed at <https://github.com/nchatterjeelab/COVID19Risk/tree/master/data>.

# Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- Life sciences       Behavioural & social sciences       Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://nature.com/documents/nr-reporting-summary-flat.pdf)

## Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

### Sample size

We built a risk model using information from two sources: (1) The UK OpenSAFELY Study and (2) state level COVID-19 deaths reported by age and sex across 51 states. The UK OpenSAFELY study is a cohort study involving a total of 17,278,392 individuals including 10,926 reported deaths from COVID-19. The CDC datasets involved entire populations of 51 states and included a total of 99,866 deaths observed prior to June 7.

We performed risk projections at population levels using the following sets of data or summary-data available from different sources National Health Interview Survey (N=22,109) for national and state level projections.

Estimate of risk factor prevalence available from Behavioral Risk Factor Surveillance (BRFSS) for projecting risks across 477 cities (Total sample size varied from 39,510 to 754,950 with mean response rate 44.9%).

Estimates of risk factor prevalence for Medicare 65+ year old population across 3,113 counties (total sample: the whole medicare population).

We performed model validation using city-level death rates and county-level death rates during the period of June 7 - October 1. The total number of deaths included in the analysis were 54,444 and 94,420, respectively.

Our analysis is descriptive in nature and is not designed for hypothesis testing and thus no power calculations were performed to determine sample size. We used information from some of the largest publicly available datasets involving millions of individuals, and tens of thousands of COVID-19 deaths. These have resulted in precise confidence intervals around risk estimates reported.

### Data exclusions

Our risk estimates were designed for the US adult population (>18 year old). Thus, we excluded individuals with age <18 from various datasets. The exclusion criterion was pre-determined.

### Replication

All our results are fully reproducible using the datasets and codes made available through GitHub.

### Randomization

None. The project involves development and validation of a risk-prediction model using various existing datasets based on observational cohort studies. Because we are not performing hypothesis testing for an effect of an individual exposure/treatment, randomization is not applicable here.

### Blinding

Not applicable. There was no randomized group allocations in the datasets.

## Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

### Materials & experimental systems

- |                                     |  |
|-------------------------------------|--|
| n/a                                 | Involved in the study                                  |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Antibodies                    |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Eukaryotic cell lines         |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Palaeontology and archaeology |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Animals and other organisms   |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Human research participants   |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Clinical data                 |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Dual use research of concern  |

### Methods

- |                                     |   |
|-------------------------------------|---|
| n/a                                 | Involved in the study                           |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> ChIP-seq               |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Flow cytometry         |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> MRI-based neuroimaging |