Who is most at risk of dying if infected with SARS-CoV-2? A mortality risk factor analysis using machine learning of COVID-19 patients over time in a large Mexican population.

Lauren D. Liao,[1*] Alan E. Hubbard,[1] Juan Pablo Gutiérrez,[2] Arturo Juárez-Flores,[2] Kendall Kikkawa,[3] Ronit Gupta,[1] Yana Yarmolich,[1] Iván de Jesús Ascencio-Montiel,[4] Stefano M. Bertozzi[1,5,6]

Affiliations:

[1] University of California, Berkeley, School of Public Health, Berkeley, CA, USA

[2] Center for Policy, Population & Health Research, School of Medicine, Universidad Nacional Autónoma de México, Mexico City, Mexico

[3] Micron Technology, Boise, ID, USA

[4] Instituto Mexicano del Seguro Social, CDMX, México

[5] University of Washington School of Public Health, Seattle, WA, USA

[6] Instituto Nacional de Salud Pública, Cuernavaca, MOR, México

***Correspondence to:** ldliao@berkeley.edu (LDL)

Word count: 2884/3000

Key messages

What is already known on this topic

Studies for Mexico and other countries have suggested that pre-existing conditions such as renal disease, diabetes, hypertension, and obesity are strongly associated with COVID-19 mortality. While age and the presence of pre-existing conditions have been shown to predict mortality, other studies have typically used less powerful statistical approaches, have had smaller sample sizes, and have not been able to describe changes over time.

What this study adds

This study examines mortality risk in a very large population (> 60 M); it uses powerful ensemble machine learning methods that outperform regression analyses; and it demonstrates marked changes over time in the degree to which different risk factors predict mortality.

How this study might affect research, practice or policy

Because we show an important improvement in predictive performance over traditional regression analyses, and the ability to update estimates as the pandemic evolves, we argue that these methods should be much more widely used to inform national programming in Mexico and elsewhere. Programs that assume that predictive models don't change over time as variants emerge and as pre-existing immunity evolves due to vaccination and prior infection will not accurately predict mortality risk.

Abstract

Background: COVID-19 would kill fewer people if health programs can predict who is at higher risk of mortality because resources can be targeted to protect those people from infection. We predict mortality in a very large population in Mexico with machine learning using demographic variables and pre-existing conditions.

Methods: We conducted a population-based cohort study with over 1.4 million laboratory-confirmed COVID-19 patients using the Mexican social security database. Analysis is performed on data from March 2020 to November 2021 and over three phases: (1) from March to October in 2020, (2) from November 2020 to March 2021, and (3) from April to November 2021. We predict mortality using an ensemble machine learning method, *super learner*, and independently estimate the adjusted mortality relative risk of each pre-existing condition using targeted maximum likelihood estimation.

Results: Super learner fit has a high predictive performance (C-statistic: 0.907), where age is the most predictive factor for mortality. After adjusting for demographic factors, renal disease, hypertension, diabetes, and obesity are the most impactful pre-existing conditions. Phase analysis shows that the adjusted mortality risk decreased over time while relative risk increased for each pre-existing condition.

Conclusions: While age is the most important predictor of mortality, younger individuals with hypertension, diabetes and obesity are at comparable mortality risk as individuals who are 20 years older without any of the three conditions. Our model can be continuously updated to identify individuals who should most be protected against infection as the pandemic evolves.

Keywords: mortality; death; COVID-19; biostatistics; international health

Introduction

The probability that someone infected with SARS-CoV-2 dies has varied enormously over time, among countries, and among population groups within countries. Interest in understanding who is at a higher risk of death has grown as this heterogeneity became more apparent. Identifying people at higher risk of severe disease and death will help health systems better respond and focus prevention resources on protecting them. We examine Mexico, a country with a very high reported case-fatality rate (4.7%) among those who have laboratory-confirmed coronavirus disease 2019 (COVID-19) as of September 23, 2022 [1].

Previous analyses in Mexico have found diabetes, obesity, hypertension, immunosuppression, and renal disease to be significant risk factors along with age and sex. Multiple authors have identified obesity and diabetes as important risk factors for mortality [2–5]. Escobedo de-la Peña et al. also found a strong association with hypertension, which is consistent with results from Giannouchos et al. [5,6]. Late-stage chronic kidney disease, although less prevalent, has also consistently been identified as a COVID-19 mortality risk factor . Older/ male patients tend to have higher mortality risks than younger/ female patients [3,5,6]. In a previous analysis, we found interactions between those comorbidities, suggesting a synergic effect when having more than one of diabetes, hypertension, and obesity (larger odds ratio when reporting the 3 conditions vs. one or two) [7]. We also found that the odds ratio increased by age group with those over age 80 having 30-fold the risk of those 20 to 29 [7]. One important consideration is that the prevalence of diabetes and hypertension is positively associated with age, so it has not been clear how this interaction is related to mortality risk. A more adaptive analysis performed by Martínez-Martínez et al. developed a prediction model for severity of COVID-19, defined by hospitalization and/or mortality. They examined the relationship of 14 variables with

hospitalization and mortality using interaction terms and splines to account for non-linear relationships [8].

The pattern of age, sex, and comorbidities being associated with higher mortality risk is not specific to Mexico, and the global literature on such associations is extensive. Researchers have identified old age, diabetes, obesity, chronic renal failure, and congestive heart failure to be strongly associated with severe infection amongst both sexes in the Spanish population [9]. Researchers in Brazil showed that older age, male, kidney disease, obesity and/ or diabetes are strong predictors of mortality amongst other comorbidities such as chronic liver disease, immunosuppression, and cardiovascular disease [10,11]. Another study used United Kingdom Biobank data and showed that pre-existing dementia, diabetes, chronic obstructive pulmonary disease (COPD), pneumonia, and depression were positively associated with risk of hospitalization and death [12]. An analysis from France found age, diabetes, hypertension, obesity, cancer, and kidney and lung transplants to be associated with risk of COVID-19-related hospitalization and mortality, among others [13]. A Canadian study reported dementia, chronic kidney disease, cardiovascular disease, diabetes, COPD, severe mental illness, organ transplant, hypertension, and cancer to be significant predictors of mortality [14]. Our goal in this study is not only to predict mortality using demographic factors and comorbidities, but to show how those predictions change over time in this rapidly evolving pandemic.

Although mortality risk estimation and risk factor identification have been examined in prior studies, we are concerned about the statistical validity and interpretation of the standard methods. A commonly used prediction tool, logistic regression, assumes a linear relationship of predictors against the log odds of mortality risk, but this logit-linear assumption will lead inevitably to biased estimates of risk (either under- or over-predict the risk) for subsets of the

population. We instead used flexible, data-adaptive methods that can capture non-linearities in the dose-response, such as potential nonlinear interactions between the predictors (e.g., the potential interaction of age and diabetes on predicting death) [15,16]. The better the model fits the study population; the more likely estimates are closer to the true joint relationship of mortality and risk factors.

We included pre-existing conditions, demographic variables, the Mexican state where the patient was treated, and the month that the patient initiated care to fit our prediction algorithm. We conducted the analysis using an ensemble machine learning algorithm, super learner, to form optimal combination of predictions from multiple machine learning methods [15,16]. We also estimated the comparative importance of variables for mortality risk prediction (holding all other variables constant) by nonparametrically estimating quantities inspired by causal parameters (parameters that compare so-called counterfactual distributions, in our case, causal relative risks). The statistical goal is to estimate and provide robust inference for impact estimates of the predictors without the arbitrary modeling assumptions that characterize the great majority of prior work [17].

Methods

Study population and design

The study population is drawn from the Mexican Social Security Institute (IMSS), a vertically integrated insurance and health system that provides coverage for over 60 million private sector employees and their families, including their parents, children and spouse. IMSS also provided care as part of the COVID-19 response for some non-beneficiaries, who are also included in the dataset.

The data were recorded from March 1st, 2020, to November 3rd, 2021 in a platform called SINOLAVE. They reflect the entire population of 4,482,292 patients who were registered as receiving care for suspected COVID-19 at an IMSS facility. The dataset and the data entry process have been described previously [18]. The demographic variables include age, sex, insured by IMSS, and indigenous status. The data contains pre-existing conditions reported by the patient or the family at presentation: asthma, cardiovascular disease, chronic liver disease, chronic obstructive pulmonary disease, diabetes, hemolytic anemia, human immunodeficiency virus, hypertension, immunosuppression, neurological disease, obesity, cancer, renal disease and tuberculosis, as well as whether the patient currently smokes. Patients were asked at presentation about their pre-existing health conditions; these were not ascertained with reference to the patient's medical record, even for those patients insured by the IMSS. The data also includes the Mexican state in which the patient received care, COVID-19 test results (from both polymerase chain reaction (PCR) tests and antigen tests), the month that the patient initiated care, and mortality. In addition, we extracted a different dataset from the National Council of Science and Technology to determine the dominant circulating variant in each month [19]. A short summary can be found in **Table 1 (Supplemental Table S1)**. We define COVID-19 positive as a positive PCR or antigen test.

From the full data set, we generated an analytic sample (n = 1,423,720) **(Supplemental Figure S1)**. We exclude those under the age of 20 years, those without any positive COVID-19 test result from either the PCR or antigen tests, and those with unknown pre-existing conditions. We also create a phase variable that corresponds to changes in the epidemic curve into three: phase 1 is from March 1st, 2020, to October 31st, 2020, phase 2 is from November 1st, 2020, to

March 31st, 2021, and phase 3 is from April 1st, 2021, to November 3rd, 2021 as previously described [18].

Statistical analysis

Mortality risk prediction using super learner (SL)

We predict mortality risks with SL [15,16], using predictors: pre-existing conditions, demographic variables, the Mexican state where the patient was treated, and the month that the patient initiated care. SL combines a set of user-supplied machine learning algorithms, which includes both simple, parametric fits and flexible algorithms, to create an optimally-weighted combination. This optimal fit is found by creating a combination of algorithms that minimize the cross-validated risk (in our case, the negative log-likelihood). SL has the property that asymptotically it will perform at least as well as the best fitting algorithm in the library [15,16]. Thus, it is important to include a diverse and large set of learners as candidates to ensure the model can fit complex patterns if warranted, but also, simpler, parametric models if simpler fits are sufficient. The following learners were included in the SL library: Bayesian additive regression trees [21], Bayesian generalized linear model [22], elastic net regression [23], empirical mean, generalized additive model [24], least absolute shrinkage and selection operator regression [25], logistic regression, multivariate adaptive regression splines [26], random forest [27], ridge regression [28], and extreme gradient boosting algorithms [29]. We estimate the prediction performance, via the AUC, and derive a 95% confidence interval for the estimated AUC [30]. We compare the SL fit using all predictors listed above to a logistic regression with only age entered as a linear term. We compute the AUC for the resulting SL/logistic regression fits with 3-fold cross validation on the 80%, both on the same data used to estimate SL/logistic

regression models (training AUC), as well as a more realistic assessment by using the test set – the left-out 20% of the available data (testing AUC).

To interpret the final prediction model generated by the SL fit, we use the permutation-based variable importance measure to identify variables that influence the SL model's prediction [27]. This is performed by permuting the predictor variables one at a time (keeping the other variables fixed) and measuring the magnitude of the decline on the predictive performance (as measured by the change in the average negative log-likelihood). This provides a list of variables ranked by the relative importance to prediction fit but does not provide information on the variable impact on mortality, which led us to another measure of relative risk (RR) using targeted maximum likelihood estimation (TMLE).

Pre-existing condition relative risk estimate through targeted maximum likelihood estimation

For pre-existing conditions, we estimated a different variable importance measure that is not focused on prediction accuracy but on estimating potential impacts of pre-existing conditions on mortality risk. The impact is estimated by the RR of adjusted means (adjusted for baseline confounders) for the population if everyone had the specific pre-existing condition of interest (the numerator) versus the same population where no one has the specific pre-existing condition (the denominator). To estimate RRs, we used cross-validated targeted minimum-loss-based estimation (cross-validated TMLE). TMLE is a semiparametric, substitution estimator that has shown to be asymptotically efficient (unlike the inverse probability of treatment-weighting estimators [31]). It also has some robustness advantages over other semiparametric efficient approaches, such as augmented inverse probability weighting. TMLE estimates parameters that,

under certain assumptions, can be interpreted as potential causal impacts of these factors on mortality, in our case, in the form of a causal relative risk. Our ensemble machine learning is optimized for prediction, but it does not directly provide measures of individual variable importance. We conducted follow-up procedure (TMLE) to generate interpretable estimates of variable impact with robust standard errors [32, 33].

Results

Descriptive results show the age distribution of laboratory-confirmed patients across the three different epidemic phases **(Supplemental Figure S2**). Phases 1 and 2 have similar distributions, and there are more young people (under 30) in phase 3. The six most prevalent pre-existing conditions are hypertension, obesity, diabetes, smoking, asthma, and renal disease **(Supplemental Figure S3**). The prevalence of all pre-existing conditions decreased over the three phases, and prevalence of hypertension, obesity, and diabetes were drastically reduced in phase 3.

Super learner (SL) prediction

SL fit has high prediction accuracy on the testing set (AUC: 0.907 (95% CI: (0.905-0.908)). SL leverages XGBoost models (**Supplemental Table S2)** and significantly outperforms the simple logistic regression model (testing AUC: 0.874 (95% CI: (0.872-0.876)) (**Table 2**). The logistic regression model overpredicts mortality risks for those roughly above age 75 compared to the SL prediction (**Fig. 1**). Permuted variable importance shows, while holding other variables constant, age is consistently the most important for SL prediction in average mortality risk (**Supplemental Figure S4 and Table S3)**. Having multiple comorbidities can dramatically increase risk for those individuals (**Fig. 2**).

Relative risks of pre-existing conditions

To assess the impact of each pre-existing condition, we estimate their respective relative risks (RRs) of mortality, adjusting for demographic variables. We report the estimated RRs in **Table 3**, ordered by impact (most to least) (**Supplemental Figure S5**). The RRs compare the expected risk if all patients have the pre-existing condition (with) versus if all patients do not have the condition (without). The highest impact pre-existing condition is renal disease (RR: 3.783, 95% CI: (3.705, 3.862)); diabetes, obesity, and hypertension also have high impact individually (RR: 1.432-1.847). Minimal differences between the risk estimates are shown for smoking and asthma (RR: 1.049 and 1.037, respectively).

The phase analyses indicate pre-existing conditions are especially important in phase 3. Phase 1 and 2 are very similar in terms of both risk prediction and adjusted mortality risk estimates. However, in phase 3, age is less important in prediction (**Supplemental Table S3)** and RRs drastically increase for every comorbidity. The adjusted risks show the decrease for each pre-existing condition in phase 3 (**Supplemental Table S4)**.

Discussion

Our analysis of (>1.4 million) laboratory-confirmed COVID-19 patients demonstrates that age is by far the most important predictor of average mortality. For those patients with renal disease, diabetes, hypertension, or obesity, having the comorbidity further increases their risk of mortality. A patient with diabetes, hypertension, and obesity is roughly comparable to a patient 20 years older with none of the conditions, based on the predicted mortality (**Fig. 2**). Thus, having a comorbidity increases risk of mortality and should be considered at any age. The reason that comorbidities add little to the predictive power at younger ages is that hypertension and

249 diabetes are age-related and the reported onset is often for those over 30, so the pre-existing
250 conditions are far less prevalent.

251 Our prediction results using machine learning methods predict better than previous
252 studies, and we demonstrated the feasibility and robustness of using machine learning methods
253 targeted for prediction and variable impact. SL model prediction has an AUC of 0.907, which is
254 higher than any previous Mexican study (AUCs from 0.634 to 0.824) [8,34]. Although age has
255 been well reported by previous studies as important [5,34,35], our analysis is more robust
256 because we do not assume a pre-specified functional relationship between the explanatory
257 variables and the predicted variable, and thereby avoid any arbitrary groupings into age
258 categories. Moreover, since those above age 60 have a higher prevalence of comorbidities,
259 relying on simple logistic regression models can greatly overpredict the average mortality risk
260 for the elder patients. Our study applies TMLE to estimate the adjusted mortality risk ratios for
261 each comorbidity to provide more robust impact estimates that respect time ordering and account
262 for background variables.

263 We find consistent results of comorbidities compared to previous studies, and present
264 phase analyses highlighting the changes in relative risks over time. Previous results from logistic
265 regressions indicated odds ratios of 1.458-2.48 for renal disease, 1.237-1.74 for diabetes, 1.173-
266 1.47 for obesity, 1.194-1.315 for hypertension, 0.852-1.02 for smoking, and 0.74-1.420 for
267 asthma [34–36]. Although our analysis is generally consistent with previous findings, our RR
268 estimations have less uncertainty. Renal disease has the greatest impact on mortality, followed
269 by diabetes, hypertension, and obesity; smoking and asthma have negligible impact on mortality
270 risk.

This phase-specific analysis produced a seemingly paradoxical finding. The impact of comorbidities on predicted mortality decreased with time (primarily between the second and third wave), but the RR on mortality dramatically increased for the same conditions **(Supplemental Table S4 and Figure S5)**. The apparent explanation is that mortality risk for people without the comorbidities fell faster than for people with them, increasing the relative risk. The decrease in mortality risk is multifactorial and includes a decrease in susceptibility over time (due to prior infection and vaccination), improved treatment, enhanced healthcare response and opportunity to be admitted to a hospital or ICU, and less virulent viral subtypes. This implies that as herd immunity increases, medical resources should focus even more on protecting vulnerable people at older age and those with comorbidities since they are even more likely to experience severe outcomes compared to those who are younger and/or healthier.

Readers should be cautious about extrapolating our findings to other populations. Although our sample is large and includes patients from all parts of Mexico, most of the patients were IMSS beneficiaries. In order to access IMSS health services, patients require: a) be a formal-sector worker or retired, b) be a direct dependent of such an employee, c) be a bachelor or postgraduate student in a public institution, d) voluntarily enroll by paying a fee. Thus, the IMSS population skews toward the upper half of the income distribution. Populations without similar access to health services may have different results. It is also important to consider the potential impact of data quality. Pre-existing conditions were self-reported and likely also inconsistently recorded, perhaps in systematic ways that could have biased the results. For example, if people with severe diabetes were more likely to report diabetes as a pre-existing condition, we may overestimate the impact of diabetes on mortality.

It is also important to consider what predictive variables are included in this model. We sought to predict risk for an individual in the population using their characteristics prior to infection. In other words, what is this person's risk of death from COVID-19 if they were to be infected? The answer to this question best informs the question of who should be prioritized for protection against infection or for early therapeutic interventions following infection. It does not attempt to predict the likely mortality of a patient who presents to the health services with COVID-19 because information about that patient's severity of their COVID-19-related symptoms will represent important additional predictors of their mortality risk.

Abbreviations

AUC, area under the receiver operating characteristic curve; CI, confidence interval; COPD, chronic obstructive pulmonary disease; COVID-19, coronavirus disease of 2019; IMSS, Mexican Social Security Institute; PCR, polymerase chain reaction; RR, relative risk; SL, super learner; TMLE, targeted maximum likelihood estimation; XGBoost, extreme gradient boosting

Declarations:

Acknowledgements

We thank the staff of C3.ai DTI for their technical support and our colleagues at University of California, Berkeley, the Mexican National Autonomous University, and the Mexican Social Security Institute (IMSS) for all of the administrative and technical support that has allowed this collaboration to flourish.

Authors' contributions

LDL and AEH contributed to the study design and methodology. AJ, YY, and IA contributed to data acquisition. LDL, YY, and KK contributed to data cleaning. LDL led the data analysis and visualization. LDL, AEH, JPG, and SMB interpreted the results. LDL drafted the manuscript with support from RG on literature search. AEH and SMB significantly contributed to the revision of the manuscript. All authors participated in review and edited the manuscript; all authors have read and approved the final manuscript. All authors had full access to all the data in the study and accepted responsibility to submit for publication. All authors take responsibility for the integrity of the data and the accuracy of the data analysis.

Funding

This research effort was funded by the C3.ai Digital Transformation Institute. The C3.ai DTI was established by C3.ai, Microsoft, the University of California, Berkeley (UC Berkeley), the University of Illinois at Urbana-Champaign (UIUC), Carnegie Mellon University, University of Chicago, MIT, and Princeton University. It is being funded in cash and in kind by C3.ai, Microsoft Azure, and the Lawrence Berkeley National Laboratory. The funders had no role in access to data, design of the research, or analyses conducted. They have not seen or contributed to the manuscript in any way.

In addition, LDL received funding from the National Science Foundation (DGE 2146752). AEH received funding from a global development grant (OPP1165144) from the Bill & Melinda Gates Foundation to the University of California, Berkeley, CA, USA.

Availability of data and materials

The study was conducted using confidential patient records subject to strict access controls and we are therefore unable to share the data that were used for this study.

Ethics approval and consent to participate

This data-only study was approved on November 4th, 2020, by the Scientific Research National Committee (Social Security Mexican Institute) with R-2020-785-165. The University of California, Berkeley Institutional Review Board (IRB) determined that the project was exempt from IRB approval.

Consent for publication

Not applicable.

Competing interests

All authors declare no competing interests.

References

1. Hopkins. University of Medicine. Coronavirus Resource Center. *Data Stream* 2020.

2. Singer M. Deadly Companions: COVID-19 and Diabetes in Mexico. *Med Anthropol* 2020;**39**:660–5.

3. Bello-Chavolla OY, Bahena-López JP, Antonio-Villa NE, *et al.* Predicting Mortality Due to SARS-CoV-2: A Mechanistic Score Relating Obesity and Diabetes to COVID-19 Outcomes in Mexico. *J Clin Endocrinol Metab* 2020;**105**. doi:10.1210/clinem/dgaa346

4. Noyola DE, Hermosillo-Arredondo N, Ramírez-Juárez C, *et al.* Association between obesity and diabetes prevalence and COVID-19 mortality in Mexico: an ecological study. *J Infect Dev Ctries* 2021;**15**:1396–403.

5. Giannouchos TV, Sussman RA, Mier Odriozola JM, *et al.* Characteristics and risk factors for COVID-19 diagnosis and adverse outcomes in Mexico: an analysis of 89,756 laboratory–confirmed COVID-19 cases. bioRxiv. 2020. doi:10.1101/2020.06.04.20122481

6. Peña JE la, Rascón-Pacheco RA, Ascencio-Montiel I de J, *et al.* Hypertension, Diabetes and Obesity, Major Risk Factors for Death in Patients with COVID-19 in Mexico. *Arch Med Res* 2021;**52**:443–9.

7. Gutierrez JP, Bertozzi SM. Non-communicable diseases and inequalities increase risk of death among COVID-19 patients in Mexico. *PLoS One* 2020;**15**:e0240394.

8. Martínez-Martínez MU, Alpízar-Rodríguez D, Flores-Ramírez R, *et al.* An Analysis COVID-19 in Mexico: a Prediction of Severity. *J Gen Intern Med* 2022;**37**:624–31.

9. Gimeno-Miguel A, Bliek-Bueno K, Poblador-Plou B, *et al.* Chronic diseases associated with increased likelihood of hospitalization and mortality in 68,913 COVID-19 confirmed cases in Spain: A population-based cohort study. *PLoS One* 2021;**16**:e0259822.

10. Soares R de CM, Mattos LR, Raposo LM. Risk Factors for Hospitalization and Mortality due to COVID-19 in Espírito Santo State, Brazil. *Am J Trop Med Hyg* 2020;**103**:1184–90.

11. Wollenstein-Betech S, Silva AAB, Fleck JL, *et al.* Physiological and socioeconomic characteristics predict COVID-19 mortality and resource utilization in Brazil. *PLoS One* 2020;**15**:e0240346.

12. Atkins JL, Masoli JAH, Delgado J, *et al.* Preexisting Comorbidities Predicting COVID-19 and Mortality in the UK Biobank Community Cohort. *J Gerontol A Biol Sci Med Sci* 2020;**75**:2224–30.

13. Semenzato, Botton, Drouin, *et al.* Chronic diseases, health conditions and risk of COVID-19-related hospitalization and in-hospital mortality during the first wave of the epidemic in France: a …. *The Lancet Regional* https://www.sciencedirect.com/science/article/pii/S2666776221001356

14. Ge E, Li Y, Wu S, *et al.* Association of pre-existing comorbidities with mortality and disease severity among 167,500 individuals with COVID-19 in Canada: A population-based cohort study. *PLoS One* 2021;**16**:e0258154.

15. van der Laan MJ, Polley EC, Hubbard AE. Super learner. *Stat Appl Genet Mol Biol* 2007;**6**:Article25.

16. Polley EC, van der Laan MJ. Super Learner In Prediction. Published Online First: 2010.https://biostats.bepress.com/ucbbiostat/paper266/?TB_iframe=true&width=370.8&height=658.8 (accessed 26 Jul 2022).

17. Petersen ML, van der Laan MJ. Causal models and learning from data: integrating causal modeling and statistical estimation. *Epidemiology* 2014;**25**:418–26.

18. Juárez-Flores A, Ascencio-Montiel IJ, Gutiérrez JP, *et al.* COVID-19 in the Mexican Social Security Institute (IMSS) population. Prevalent symptoms. bioRxiv. 2022. doi:10.1101/2022.04.12.22273734

19. Vigilancia de variantes del virus SARS-CoV-2. Vigilancia de variantes del virus SARS-CoV-2. https://salud.conacyt.mx/coronavirus/variantes/. (accessed 29 Jul 2022).

20. Hubbard AE, Kennedy CJ, van der Laan MJ. Data-Adaptive Target Parameters. In: van der Laan MJ, Rose S, eds. *Targeted Learning in Data Science: Causal Inference for Complex Longitudinal Studies*. Cham: : Springer International Publishing 2018. 125–42.

21. Chipman HA, George EI, McCulloch RE. BART: Bayesian additive regression trees. *aoas* 2010;**4**:266–98.

22. Dey DK, Ghosh SK, Mallick BK. *Generalized linear models: A Bayesian perspective*. CRC Press 2000.

23. Hans C. Elastic Net Regression Modeling With the Orthant Normal Prior. *J Am Stat Assoc* 2011;**106**:1383–93.

24. Liu. Generalized additive model. *Rep Univ Jyvaskyla Dep Math Stat* Published Online First: 2008.http://people.vcu.edu/~dbandyop/BIOS625/GAM.pdf

25. Tibshirani R. Regression Shrinkage and Selection via the Lasso. *J R Stat Soc Series B Stat Methodol* 1996;**58**:267–88.

26. Friedman JH. Multivariate Adaptive Regression Splines. *aos* 1991;**19**:1–67.

27. Breiman L. Random Forests. *Mach Learn* 2001;**45**:5–32.

28. Hoerl AE, Kennard RW. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* 1970;**12**:55–67.

29. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: : Association for Computing Machinery 2016. 785–94.

30. LeDell E, Petersen M, van der Laan M. Computationally efficient confidence intervals for cross-validated area under the ROC curve estimates. *Electron J Stat* 2015;**9**:1583–607.

31. Rosenbaum PR. Model-Based Direct Adjustment. *J Am Stat Assoc* 1987;**82**:387–94.

32. van der Laan MJ, Rubin D. Targeted Maximum Likelihood Learning. *Int J Biostat* 2006;**2**. doi:10.2202/1557-4679.1043

33. Porter KE, Gruber S, van der Laan MJ, *et al*. The relative performance of targeted maximum likelihood estimators. *Int J Biostat* 2011;**7**. doi:10.2202/1557-4679.1308

34. Wollenstein-Betech S, Cassandras CG, Paschalidis IC. Personalized predictive models for symptomatic COVID-19 patients using basic preconditions: Hospitalizations, mortality, and the need for an ICU or ventilator. *Int J Med Inform* 2020;**142**:104258.

35. Parra-Bracamonte GM, Lopez-Villalobos N, Parra-Bracamonte FE. Clinical characteristics and risk factors for mortality of patients with COVID-19 in a large data set from Mexico. *Ann Epidemiol* 2020;**52**:93–8.e2.

36. Hernández-Galdamez DR, González-Block MÁ, Romo-Dueñas DK, *et al.* Increased Risk of Hospitalization and Death in Patients with COVID-19 and Pre-existing Noncommunicable Diseases and Modifiable Risk Factors in Mexico. *Arch Med Res* 2020;**51**:683–9.

Table 1. Summary table of baseline variables and pre-existing conditions

	All time (2020/03-2021/11)	**Phase 1** (2020/03-2020/10)	**Phase 2** (2020/11-2021/03)	**Phase 3** (2021/04-2021/11)
Sample size	1,423,720	303,278	425,698	694,744
Demographic variables				
Age in years (mean (SD))	42.15 (15.70)	46.41 (16.04)	44.89 (16.27)	38.61 (14.34)
Sex = male (%)	729,782 (51.3)	158,248 (52.2)	218,165 (51.2)	353,369 (50.9)
Insured by IMSS = yes (%)	1,358,440 (95.4)	288,588 (95.2)	402,754 (94.6)	667,098 (96.0)
Indigenous = yes (%)	7,381 (0.5)	2,200 (0.7)	1,628 (0.4)	3,553 (0.5)
Pre-existing conditions				
Hypertension = yes (%)	228,901 (16.1)	72,615 (23.9)	83,735 (19.7)	72,551 (10.4)
Diabetes = yes (%)	169,869 (11.9)	55,551 (18.3)	61,120 (14.4)	53,198 (7.7)
Obesity = yes (%)	181,736 (12.8)	55,965 (18.5)	60,217 (14.1)	65,554 (9.4)
Smoking = yes (%)	87,161 (6.1)	21,253 (7.0)	28,346 (6.7)	37,562 (5.4)
Asthma = yes (%)	25,297 (1.8)	7,951 (2.6)	7,765 (1.8)	9,581 (1.4)
Renal Disease Diagnosis = yes (%)	24,099 (1.7)	8,912 (2.9)	8,555 (2.0)	6,632 (1.0)
Outcome				
Death = yes (%)	149,805 (10.5)	53,530 (17.7)	62,517 (14.7)	33,758 (4.9)

IMSS: Mexican Institute of Social Security; SD: standard deviation

Table 2. Prediction results

	All time (2020/03-2021/11) AUC (95% CI)	Phase 1 (2020/03-2020/10) AUC (95%CI)	Phase 2 (2020/11-2021/03) AUC (95%CI)	Phase 3 (2021/04-2021/11) AUC (95%CI)
Super learner fit	Training: 0.916 (0.915-0.917) Testing: 0.907 (0.905-0.908)	Training: 0.887 (0.885-0.888) Testing: 0.873 (0.870-0.876)	Training: 0.904 (0.903-0.906) Testing: 0.895 (0.892-0.897)	Training: 0.914 (0.913-0.916) Testing: 0.906 (0.902-0.909)
Age only logistic regression fit	Training: 0.874 (0.873-0.875) Testing: 0.874 (0.872-0.876)	Training: 0.845 (0.843-0.846) Testing: 0.846 (0.842-0.850)	Training: 0.868 (0.866-0.870) Testing: 0.871 (0.868-0.874)	Training: 0.867 (0.865-0.869) Testing: 0.871 (0.866-0.875)

AUC: area under the receiver operating characteristic curve; CI: confidence interval

Table 3. Targeted maximum likelihood estimation relative risk results for each pre-existing condition

	All time (2020/03-2021/11) Relative Risk (95% CI)	Phase 1 (2020/03-2020/10) Relative Risk (95%CI)	Phase 2 (2020/11-2021/03) Relative Risk (95%CI)	Phase 3 (2021/04-2021/11) Relative Risk (95%CI)
Renal disease	3.783 (3.705, 3.862)	2.588 (2.521, 2.657)	2.994 (2.910, 3.080)	6.638 (6.361, 6.927)
Diabetes	1.847 (1.820, 1.875)	1.536 (1.508, 1.566)	1.594 (1.564, 1.625)	2.508 (2.423, 2.596)
Hypertension	1.745 (1.721, 1.770)	1.427 (1.402, 1.452)	1.500 (1.474, 1.527)	2.356 (2.279, 2.436)
Obesity	1.432 (1.417, 1.447)	1.269 (1.249, 1.288)	1.259 (1.239, 1.279)	1.794 (1.750, 1.840)
Smoking	1.049 (1.030, 1.068)	1.001 (0.975, 1.028)	0.992 (0.966, 1.018)	1.158 (1.107, 1.210)
Asthma	1.037 (1.002, 1.073)	0.941 (0.895, 0.989)	0.942 (0.892, 0.995)	1.223 (1.134, 1.319)

CI: confidence interval

Fig. 1. Mortality risk prediction comparing age only logistic regression and super learner

GAM: generalized additive model
The smoothed true mortality risk curve is generated using a GAM with integrated smoothness estimation fitted with cubic splines.

Fig. 2. Super learner predicted mortality risk averaged by specific age in two subgroups: those having all obesity, diabetes, and hypertension pre-existing conditions versus those without

Supplemental Material

SupplementalMaterial.pdf:

Table S1. TableS1 – [Complete table of baseline variables and pre-existing conditions]

Table S2. TableS2 – [Weighted combination of the super learner fit]

Table S3. TableS3 – [Top 5 ranked most important variables for prediction]

Table S4. TableS4 – [Targeted maximum likelihood estimation adjusted mortality risk, with or without the pre-existing condition]

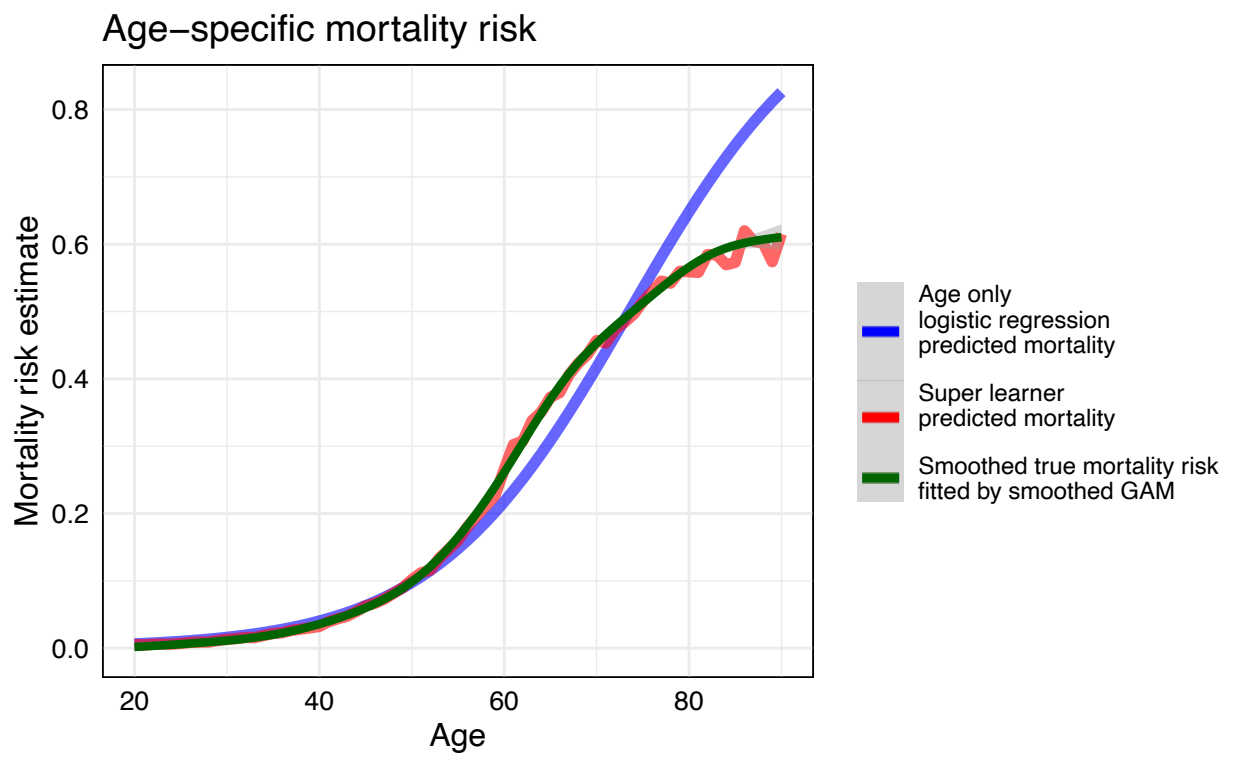
Figure S1. FigS1 – [Flowchart for analytic sample development]

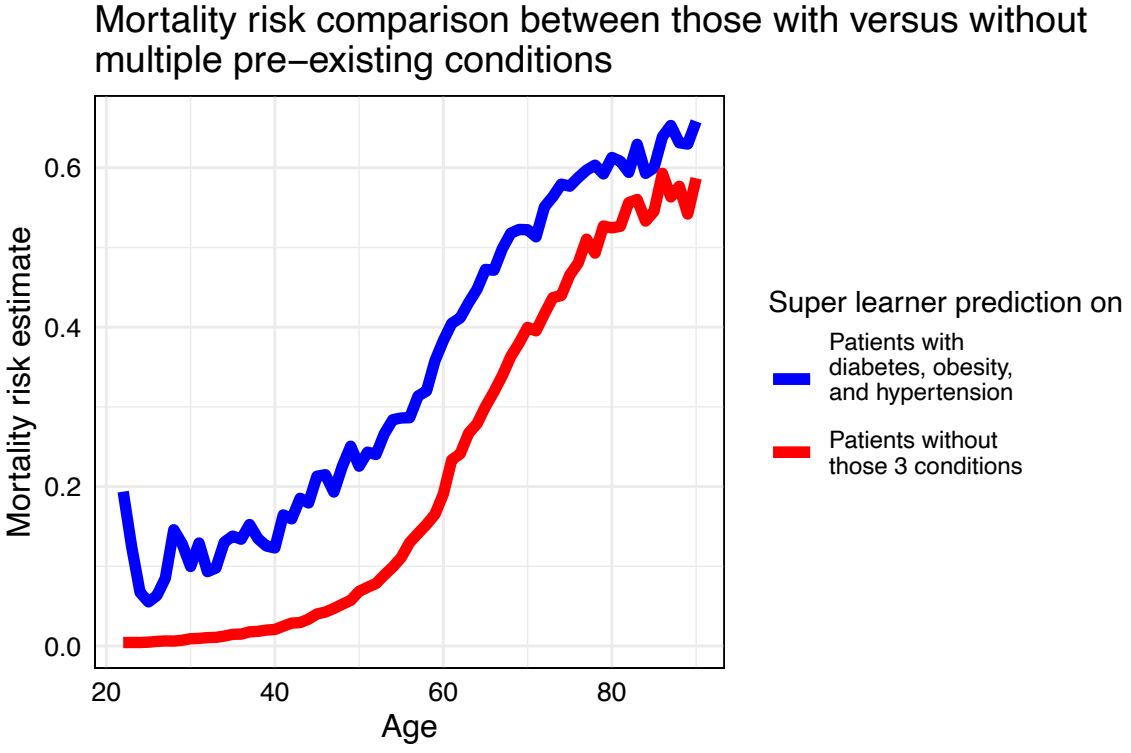
Figure S2. FigS2 – [Age distribution for laboratory-confirmed COVID-19 patients]

Figure S3. FigS3 – [Prevalence of pre-existing conditions prevalence over time]

Figure S4. FigS4 – [Prediction variable importance predicted using the super learner fit]

Figure S5. FigS5 – [Relative risk for each pre-existing condition associated with mortality]


Age–specific mortality risk
Mortality risk estimate
0.0
0.2
0.4
0.6
0.8
20
40
60
80
Age
Age only
logistic regression
predicted mortality
Super learner
predicted mortality
Smoothed true mortality risk
fitted by smoothed GAM


Mortality risk comparison between those with versus without multiple pre−existing conditions
Mortality risk estimate
0.0
0.2
0.4
0.6
20
40
60
80
Age
Super learner prediction on
Patients with diabetes, obesity, and hypertension
Patients without those 3 conditions