Predicting COVID-19 Mortality in Mexico from Comorbidity and Demographic Data

By Sarina Kopf and Martin Pollack

Abstract

Predictive models of mortality due to COVID-19 have clinical importance since they can guide life-saving resource allocation decisions; however, most of the models created so far have either used outdated data or not considered the effects of interactions between different factors. To address this need, we utilized logistic regression to create a model that considered interaction terms to predict COVID-19 mortality, fitting the model using a recent dataset with information on COVID-19 cases in Mexico. We find that the explanatory variables sex, age, whether or not the person self identifies as indegenous or smokes, and whether or not the patient has medical conditions including diabetes, chronic obstructive pulmonary disease (COPD), athsma, hyptertension, immunosuppression, obesity, chronic kidney failure, cardiovascular illness, or other illnesses were all significant predictors of COVID-19 mortality. Several significant interactions between comorbidities and sex, comorbidities and age, and between different comorbidities were also identified. Our final model had good predictive power and allowed us to identify how a person's risk of dying from COVID-19 in Mexico changes with various factors.

Introduction

Since its emergence up until December 22, 2020, the novel SARS-CoV-2 coronavirus, the virus responsible for the disease COVID-19, has infected a total of 77.5 million people and killed over 1.7 million people worldwide (Johns Hopkins Coronavirus Resource Center). This highly infectious disease has affected people unequally: men, older people and those with pre-existing conditions are more likely to experience a severe course of the disease as well as die from it (Moula et al. 2020). Specifically, meta-analysis of human studies shows that conditions such as obesity, cardiovascular disease, diabetes, chronic obstructive pulmonary disorder, and chronic kidney disease, but not smoking, increase mortality due to COVID-19 (Flaherty et al. 2020).

Understanding the risk presented by these comorbidities has clinical significance; for example, it can assist physicians in prescribing treatment and behavioral changes (e.g. social distancing) in a patient-specific context. Further, it can guide life-saving resource-allocation decisions such as who should be given priority for ICU beds or the newly-approved vaccines. However, much of the research on the effect of comorbidities on mortality due to COVID-19 is outdated (based on data published before July 2020) and follows small cohorts of people, restricting the ability to generalize these findings. These analyses are further limited since they fail to consider the interactions between comorbidities and demographic factors and the corresponding impact on the risk of death from COVID-19. We aimed to address these limitations in previous research by developing a predictive model with interaction terms based on a recent dataset from Mexico in order to evaluate a patient's risk of dying from COVID-19.

Methods:

For our analysis, we utilized the publicly available dataset compiled by the General Bureau of Epidemiology and released by the Mexican Ministry of Health which contains demographic and comorbidity data for over 2 million Mexican citizens between the ages of 18 and 85 who have been "associated with COVID-19" (Bases de datos COVID-19, 2020). At the time of writing, this dataset included information about all Mexicans who have been suspected of having, tested for, or confirmed to have COVID-19 from the onset of the pandemic until November 30, 2020.

Since we were interested in mortality due to COVID-19 given a SARS-CoV-2 infection, we filtered our dataset to only contain cases who were confirmed by a laboratory to have SARS-CoV-2. Next, we selected only the columns corresponding to our response variable, whether or not a patient died, and the variables with demographic or commorbidity information such as sex, a person's exact age, whether or not the person self identifies as indegenous or smokes, and whether or not the patient has medical conditions including diabetes, chronic obstructive pulmonary disease (COPD), athsma, hyptertension, immunosuppression, obesity, chronic kidney failure, cardiovascular illness, or other illnesses. Finally, we removed all rows containing at least one missing value (about 30,000 out of ~1.03 million of the remaining observations). After carefully examining these observations with missing values, we found that deleting them did not significantly alter the distributions of our variables of interest. Therefore, we concluded that the elimination of these patients would likely not introduce substantial bias.

After cleaning our data, our final dataset contained 992,463 observations, of which 93,893 had died (approximately 9.5%). We split this dataset in half to obtain our training and testing datasets, maintaining approximately a 9.5% death rate in each. Using our training data, we attempted to fit the best logistic regression model to predict COVID-19 mortality from a patient's demographic and comorbidity information using manual, backwards stepwise regression. In total we created three models and compared them with metrics such as residual deviance, area under the ROC curve (AUC), and accuracies. Once we selected our best model, we evaluated its performance using the testing dataset and these same metrics.

Results:

Using our training dataset, we created a logistic regression model to predict COVID-19 death using only our thirteen explanatory variables, all of which appeared significant. Building upon this no-interaction model, we created a model with all explanatory variables and all two-way interaction terms. A drop-in-deviance test comparing this all-interaction model to the one with no interactions suggested that at least one interaction was important and should be included in the model ($p < 0.001$). Although adding all interactions resulted in noticeable improvements in various metrics (Table 1), our model became much more complicated by adding all 78 interaction terms, most of which did not seem significant.

As a result, we created a reduced model with our training dataset containing only the most important interactions. We chose which terms to include according to their p-values in the all-interactions model. Specifically, to keep our Type I error rate manageable in multiple comparisons, we employed the Bonferroni method to determine our significance threshold. Based on this threshold, we kept 19 of the 78 interaction terms to create our reduced-interaction model (Appendix 1). Although Bonferonni methods are sometimes too restrictive, resulting in the elimination of important variables, this reduced-interaction model captures most of the improvements in residual deviance and AUC resulting from the addition of interaction terms (Table 1).

Despite the fact that the model with some interactions had the lowest overall accuracy, it was best at predicting the outcome for those who actually died. This is preferable for saving lives since we want to correctly predict death to assure proper care, even if this means an increased rate of incorrectly predicting death and giving overly-cautious treatment. Since this reduced-interaction model balanced predictive power and simplicity most effectively, we chose to use it as our best overall model. Evaluating this model with our testing data, we find that although the overall accuracy has decreased, the AUC and accuracy for those who died is higher than any model using the training dataset (Table 1), suggesting our model is not overfit.

Model	Residual Deviance	AUC	Overall Accuracy	Accuracy for those who actually died
No interactions (train)	236,259	0.8539	74.77%	82.46%
Some interactions (train)	233,269	0.8559	74.73%	82.86%
All interactions (train)	232,978	0.8562	75.41%	82.01%
Some interactions (test)	N/A	0.8567	73.52%	84.13%

Table 1: Evaluation metrics for various logistic regression models.

Our final reduced-interaction model highlights that sex and age have significant interactions with various diseases—of the 19 interaction terms in the model, 15 included sex or age (Appendix 1). It also reveals how different factors affect a person's risk of dying due to COVID-19. For example, a person's chance of dying tends to increase slowly with age until they are about 50 years old and then very quickly thereafter. The odds of a man dying due to COVID-19, assuming no comorbidities are present, are higher than those odds for a woman (between 2.8 and 1.7 times higher for younger versus older individuals) (Appendix 1). Further, chronic kidney disease tends to affect younger individuals more than older ones (Figure 1A), whereas obesity drastically increases one's risk of dying from COVID-19 for all ages (Figure 1B and 1C). Notably, cardiovascular disease slightly increases risk for young people but decreases it for old people, and smoking increases the risk of dying for men over 60 but decreases risk for young men and women of all ages (Figure 1C and D). See Appendix 5 for the effects of other comorbidities on COVID-19 mortality.

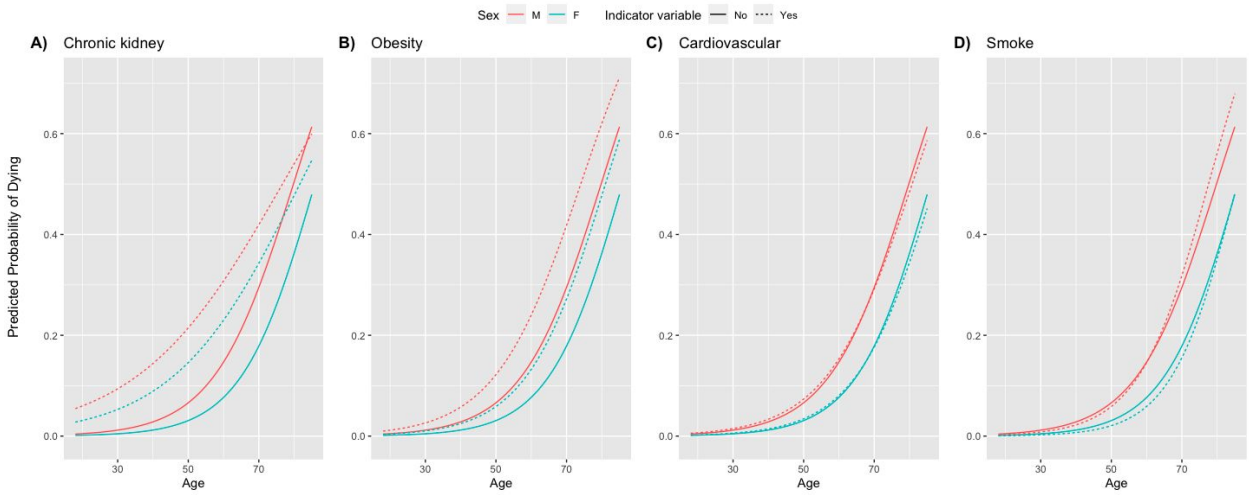


Figure 1: Selected graphs showing predicted probability of dying according to our reduced-interaction model given sex, age, and one other variable, assuming no other comorbidities.

Discussion:

Since our reduced-interaction model demonstrated improvements in metrics such as accuracy and AUC over both the no- and all-interactions models, we successfully incorporated interaction terms into a predictive model for COVID-19 mortality in Mexico. Although our goal was to create a predictive model, not to test the significance of individual variables, and despite the difficulty in interpreting coefficients due to interaction terms, overall trends in mortality by comorbidity and demographic factors are mostly reflected in the literature (Flaherty et al. 2020). These trends all have a plausible biological basis, rendering our model more generalizable (Hajifathalian et al. 2020). One exception was asthma, which our model suggests slightly decreases risk of dying for all ages despite the fact that other studies suggest the opposite (Abrams et al. 2020). Further, although cardiovascular disease increases the risk of mortality in young people, it does so only slightly, and it actually decreases risk for elderly people (Figure 1B, C). This change in risk is much smaller than that caused by obesity, which is surprising given that cardiovascular disease is much deadlier than obesity in a healthy population (Figure 1) (CDC Heart Disease). We hypothesize that this disparity is due to the high incidence of obesity in Mexico (~17%) compared to all other explanatory variables. Perhaps the high absolute number of deaths of patients with obesity artificially inflated the disease's effect on COVID-19 mortality.

However, these results must be taken in context since our model is limited in several aspects. First, many explanatory variables are subjective and their precise definitions are unclear, such as the criteria for being considered a smoker. The classification system of cause of death is further unclear and as such, we do not know whether people in our dataset died from COVID-19 or from unrelated causes. Both factors restrict the generalizability of our model and could be the source of a confounding variable if classification methods varied by region. Our data analysis is further limited by the fact that we did not create a time-to-death horizon and thus some observations in our dataset are right-censored, potentially introducing bias into our model. Finally, our model fails to account for the false-positive and false-negative rates of laboratory COVID-19 tests. Future models using this dataset would benefit from defining an event horizon and accounting for the accuracy of COVID-19 tests.

Our addition of interaction terms, which has not been discussed in the current literature, may improve predictive capacity and therefore policy decisions; however, the model's limitations must be addressed before it is deployed in a clinical setting in Mexico to assist in resource-allocation decisions.

References:

Abrams EM, Geert W'J, Yang CL. Asthma and COVID-19. (2020). CMAJ. 192(20):E551. https://doi-org.idm.oclc.org/10.1503/cmaj.200617.

Bases de datos COVID-19. (2020). Retrieved December 1, 2020, from https://www.gob.mx/salud/documentos/datos-abiertos-152127.

CDC Heart Disease. (2020). Retrieved December 21, 2020, from https://www.cdc.gov/nchs/fastats/heart-disease.htm

Flaherty, H. (2020). COVID-19 in adult patients with pre-existing chronic cardiac, respiratory and metabolic disease: a critical literature review with clinical recommendations. Tropical Diseases, Travel Medicine and Vaccines, 6(1), 1–16. https://doi.org/10.1186/s40794-020-00118-y

Hajifathalian, K., Sharaiha, R. Z., Kumar, S., Krisko, T., Skaf, D., Ang, B., . . . Fortune, B. E. (2020). Development and external validation of a prediction risk model for short-term mortality among hospitalized U.S. COVID-19 patients: A proposal for the COVID-AID risk tool. PLoS One, 15(9) doi:http://dx.doi.org.idm.oclc.org/10.1371/journal.pone.0239536

Johns Hopkins Coronavirus Resource Center. (2020). Retrieved December 22, 2020, from https://coronavirus.jhu.edu/

Moula, A. I., Micali, L. R., Matteucci, F., Lucà, F., Rao, C. M., Parise, O., . . . Gelsomino, S. (2020). Quantification of death risk in relation to sex, pre-existing cardiovascular diseases and risk factors in COVID-19 patients: Let's take stock and see where we are. Journal of Clinical Medicine, 9(9) doi:http://dx.doi.org.idm.oclc.org/10.3390/jcm9092685

Appendix:

Terms in reduced-interaction model	Estimate	Std. Error	Z value	p-value
(Intercept)	-7.0920	0.0389	-182.2500	<0.0001
SEX	-1.1650	0.0526	-22.1627	<0.0001
AGE	0.0889	0.0007	135.3933	<0.0001
INDIGENOUS	0.7780	0.2026	3.8404	0.0001
DIABETES	1.9793	0.0629	31.4480	<0.0001
COPD	1.4475	0.1762	8.2146	<0.0001
ASTHMA	-0.1416	0.0379	-3.7390	0.0002
IMMUNOCOMPROMISED	1.9386	0.1684	11.5116	<0.0001
HYPERTENSION	1.2532	0.0623	20.1190	<0.0001
OTHER DISEASE	1.5681	0.1221	12.8455	<0.0001
CARDIOVASCULAR	0.4345	0.1513	2.8717	0.0040
OBESITY	1.0104	0.0585	17.2659	<0.0001
CHRONIC KIDNEY	3.3668	0.1078	31.2184	<0.0001
SMOKE	-0.7201	0.0955	-7.5380	<0.0001
SEX:AGE	0.0073	0.0009	8.3160	<0.0001
SEX:DIABETES	0.1885	0.0245	7.6873	<0.0001
SEX:HYPERTENSION	0.1121	0.0245	4.5790	<0.0001
SEX:CHRONIC KIDNEY	0.3296	0.0487	6.7652	<0.0001
SEX:SMOKE	-0.2775	0.0526	-5.2777	<0.0001
AGE:INDIGENOUS	-0.0110	0.0032	-3.4063	0.0007
AGE:DIABETES	-0.0240	0.0010	-24.2203	<0.0001
AGE:COPD	-0.0168	0.0026	-6.5127	<0.0001
AGE:IMMUNOCOMPROMISED	-0.0238	0.0027	-8.6815	<0.0001
AGE:HYPERTENSION	-0.0165	0.0010	-16.9824	<0.0001
AGE:OTHER DISEASE	-0.0138	0.0020	-7.0683	<0.0001
AGE:CARDIOVASCULAR	-0.0064	0.0023	-2.8331	0.0046
AGE:OBESITY	-0.0067	0.0010	-6.5083	<0.0001
AGE:CHRONIC KIDNEY	-0.0403	0.0018	-22.9004	<0.0001
AGE:SMOKE	0.0119	0.0016	7.6354	<0.0001
DIABETES:OBESITY	-0.2621	0.0275	-9.5490	<0.0001
IMMUNOCOMPROMISED:CHRONIC KIDNEY	-0.9236	0.0982	-9.4078	<0.0001
HYPERTENSION:OBESITY	-0.1626	0.0277	-5.8593	<0.0001
OTHER DISEASE:OBESITY	-0.4321	0.0612	-7.0645	<0.0001

Appendix 1: Coefficients for all terms in the reduced-interaction model along with their standard errors and Z and p-values.

Model with no interactions	**Actually Survived**	**Actually Died**
Predicted Survived	332261	8243
Predicted Died	116977	38751

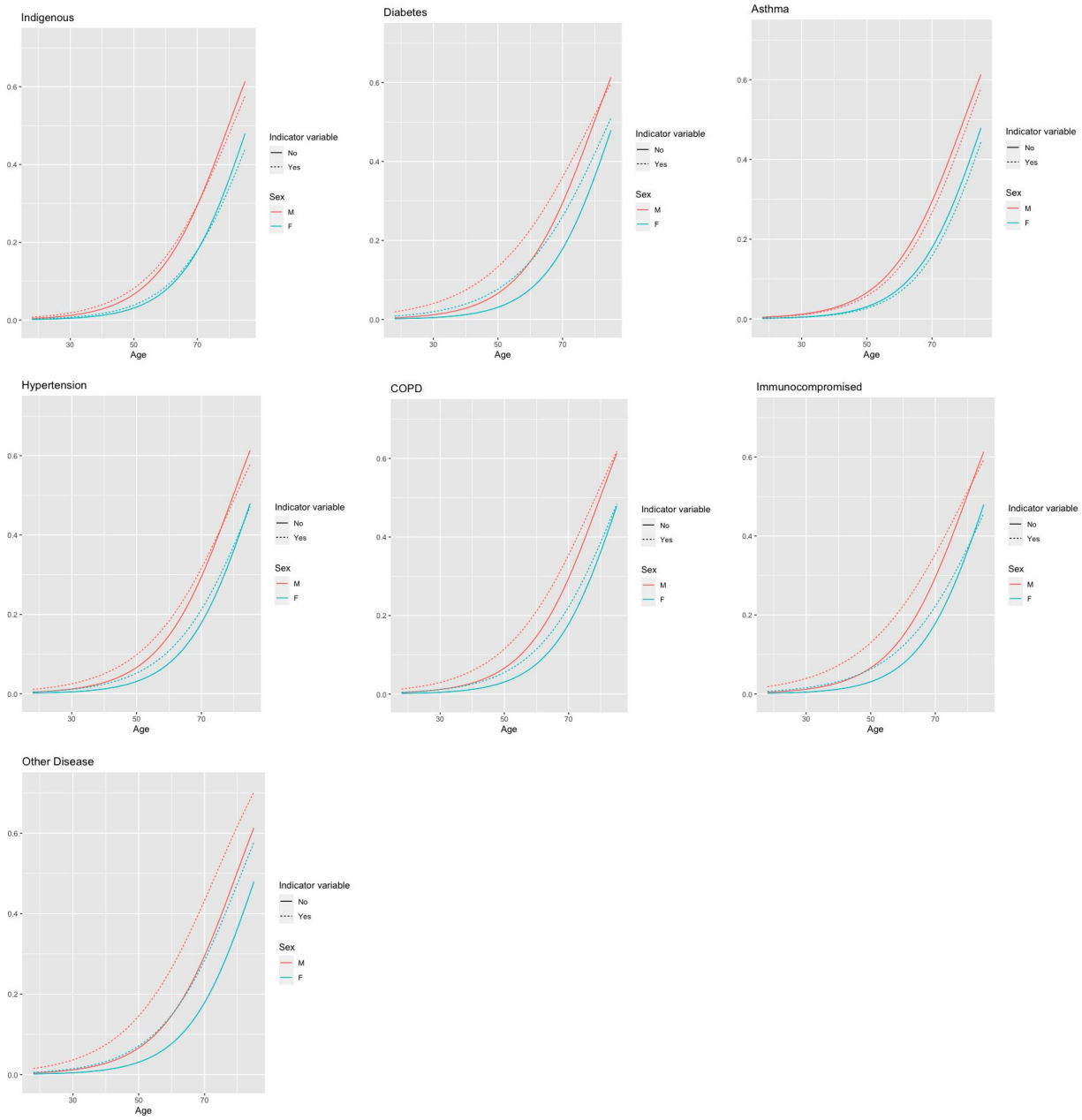
Appendix 2: Confusion matrix for the logistic regression model with no interaction terms on the training data using a predicted probability threshold of 0.083, obtained from the ROC, to classify predictions.

Model with some interactions	**Actually Survived**	**Actually Died**
Predicted Survived	331890	8056
Predicted Died	117348	38938

Appendix 3: Confusion matrix for the logistic regression model with certain interaction terms on the training data using a predicted probability threshold of 0.091, obtained from the ROC, to classify predictions.

Model with all interactions	**Actually Survived**	**Actually Died**
Predicted Survived	335664	8452
Predicted Died	113574	38542

Appendix 4: Confusion matrix for the logistic regression model with all interaction terms on the training data using a predicted probability threshold of 0.095, obtained from the ROC, to classify predictions.



Appendix 5: Predicted probability of dying according to our reduced-interaction model for different sexes and ages and depending on whether a person is indigenous or has diabetes, has asthma, has hypertension, has COPD, is immunocompromised, or has another disease.