Received: 23 March 2021 | Accepted: 3 August 2022

DOI: 10.1111/rssa.12947

ORIGINAL ARTICLE

COVID-19 clinical footprint to infer about mortality

Carlos E. Rodríguez | **Ramsés H. Mena**

Department of Probability and Statistics at the Research Institute for Applied Mathematics and Systems, Universidad Nacional Autónoma de México, México city, México

Correspondence
Carlos E. Rodríguez, Department of Probability and Statistics at the Research Institute for Applied Mathematics and Systems, Universidad Nacional Autónoma de México, México City, México.
Email: carloserwin@sigma.iimas.unam.mx

Funding information
PAPIIT-UNAM, Grant/Award Numbers: IA103220, IG100221, IV100220

Abstract

Information on 4.1 million patients identified as COVID-19 positive in Mexico is used to understand the relationship between comorbidities, symptoms, hospitalisations and deaths due to the COVID-19 disease. Using the presence or absence of these variables a clinical footprint for each patient is created. The risk, expected mortality and the prediction of death outcomes, among other relevant quantities, are obtained and analysed by means of a multivariate Bernoulli distribution. The proposal considers all possible footprint combinations resulting in a robust model suitable for Bayesian inference. The analysis is carried out considering the information on the monthly COVID-19 cases, from March 2020 to the first days of January 2022. This allows one to appreciate the evolution of the mortality risk over time and the effect the strategies of the health authorities have had on it. Supporting information for this article, containing code and the dataset used for the analysis, is available online.

KEYWORDS

Bayesian modelling, correlation, COVID-19 footprint, mortality risk, prediction

1 | INTRODUCTION

On 31 December 2019, the World Health Organization (WHO) received a troubling report from Chinese health officials (WHO, 2021). A mystery pneumonia had sickened dozens of people in

Wuhan, the capital of Hubei Province in China. A virus, that we know now as SARS-CoV-2, had been transmitted from an unknown animal host to humans and since then it has turned up lives worldwide with an unprecedented speed.

As of 8 January 2022, the WHO database has confirmed 298,915,721 COVID-19 cases globally with 5,469,303 reported deaths from 237 countries. The most affected countries are the United States with 826,022 deaths out of 57,535,858 confirmed cases; Brazil with 619,513 deaths out of 22,351,104 cases; India with 483,178 deaths out of 35,226,386 confirmed cases; the Russian Federation with 314,604 deaths out of 10,618,035 confirmed cases and Mexico with 300,303 deaths out of 4,113,789 confirmed cases (WHO, 2021).

1.1 | Clinical characteristics of COVID-19

The incubation period of COVID-19, defined as the time between exposure to the virus and symptom onset, is on average 5 to 6 days, but it can be as long as 14 days (Lauer et al., 2020). The symptoms of COVID-19 range from those that might not be noticeable to severe life-threatening illness. Some infected people have no symptoms, known as asymptomatic or pre-symptomatic carriers (Arons et al., 2020).

According to the WHO most infected people will develop mild to moderate illness and recover without hospitalisation. The WHO divides COVID-19 symptoms in three groups: *most common symptoms*; fever, dry cough and tiredness, *less common symptoms*; aches and pains, sore throat, diarrhoea, conjunctivitis, headache, loss of taste or smell, a rash on skin, or discolouration of fingers or toes, and *serious symptoms*; breathing difficulty or shortness of breath, chest pain or pressure, loss of speech or movement (WHO, 2021).

It is important to mention that individual symptoms appear to have poor diagnostic properties. Indeed, based on currently available data, neither absence nor presence of any symptoms are accurate enough to rule in or rule out the disease (Struyf et al., 2020). Thus, the gold standard for COVID-19 diagnosis is the laboratory technique known as Reverse-Transcription Polymerase Chain Reaction test; however, there are other alternatives (see Oliveira et al., 2020).

1.2 | Comorbidities and their effects in COVID-19 patients

People of any age who have underlying medical conditions, such as hypertension and diabetes, have shown worse prognosis (Sanyaolu et al., 2020). Diabetic patients have increased morbidity and mortality rates which have been linked to more hospitalisation and intensive care unit admissions (Singh et al., 2020). People with chronic obstructive pulmonary disease (COPD) or any respiratory illnesses are also at higher risk for severe illness from COVID-19 (Zhao et al., 2020).

1.3 | Impact of sex and age on COVID-19 outcomes

From the first reports from China a sex imbalance with regard to the fatality rate of COVID-19 patients has been detected. Case fatality rates reported in China, Italy, Spain, France, Germany and Switzerland support the view that a consistent biological phenomenon is operating, accounting for a higher case fatality in men. Such observation is independent of country-specific demographics and testing strategies (Gebhard et al., 2020).

Age has also been identified as a variable with high impact over the mortality rate of COVID-19 cases. All age groups appear to have significantly higher mortality compared with the immediately younger age group (Bonanad et al., 2020).

1.4 | Objectives and methods

The purpose of this research is to use the available data from the COVID-19 pandemic in Mexico to gain insight into the COVID-19 disease. The first particular objective is to understand the relationship between comorbidities, symptoms, hospitalisation and mortality. The second is to identify differences by sex and age group. The third last objective is the mortality risk prediction of a patient identified as COVID-19 positive, given commodities, symptoms, age and sex. It is important to observe that almost all relevant variables in this study are binary and indicate the presence or absence of a symptom or a comorbidity, or whether a patient has died, has been hospitalised or not.

A straightforward first strategy to analyse the impact of comorbidities and symptoms over death outcomes and hospitalisations, under a binary data setting, is to use contingency tables. Thus, denoting generically by Y_j's the presence or absence of these clinical features, basic probabilities such as $P(Y_j = 1)$ and $P(Y_i = 1|Y_j = 1)$ can be easily computed via the observed frequencies. This can be extended to account for higher order interactions of the form $P(Y_i|Y_j, Y_r)$. The idea to compute these probabilities efficiently is to concatenate the observed combinations y_jy_r and then to obtain the correspondent frequencies. The same ideas could be applied to obtain probabilities such as $P(Y_i|Y_{j_1}, Y_{j_2}, \dots, Y_{j_k})$. Moving to a modelling framework, a second vanilla strategy is to use logistic regression models, and estimate death outcome probabilities using all available variables. The challenge is to select a small subset of variables that describe the death outcomes effectively. However, it is not possible to obtain all the summaries of interest, as the joint distribution over all variables is sometimes needed and not robustly available for such a regression method. A third and possibly more general approach is to use a multivariate binary distribution.

To achieve the aforementioned objectives via the generation of relevant information summaries such as those obtained via contingency tables, while maintaining the predictive capabilities of the logistic regression, we use the multivariate Bernoulli distribution (see Dai et al., 2013). In this case all the relevant quantities are obtained in closed form and their implementation is straightforward. This model is a generalisation of the well-known Bernoulli distribution that takes the value 1 (success) with probability p and the value 0 (failure) with probability $1 - p$. In the multivariate case each observation is a vector of k successes/failures. The multivariate Bernoulli distribution assigns positive probability to each of the $\mathbb{S} := 2^k$ possible combination of successes/failures. It is important to note that this model can estimate not only the main effects and the interactions between pairs of variables, but is also capable of modelling all higher-order interactions. For us the information of each patient will be encoded by a k-dimensional vector of ones and zeros: 1 (presence) or 0 (absence) of each comorbidity, symptom, hospitalisation and death. Such data composition will be referred to as the COVID-19 footprint for each patient. With the multivariate Bernoulli distribution and the footprint for each patient is relatively simple to use the Bayesian machinery to obtain meaningful inference. Furthermore, as we will see, it is possible to obtain inferences about quantities such as the mortality risk given certain comorbidity, sex and age group. Indeed, this will be done for all patients identified as COVID-19 cases.

2 | COVID-19 PANDEMIC IN MEXICO

In Mexico, the first cases of SARS-CoV-2 were detected on 29 February 2020, and by 8 January 2022, there were 4,113,789 confirmed cases and 300,303 deaths. With this Mexico has become one of the countries with the highest death toll due to this disease.

2.1 | Diabetes, obesity and hypertension in Mexico

The WHO has said people with underlying medical problems like high blood pressure, heart and lung problems, diabetes, or cancer, are among those most vulnerable to severe cases of the new coronavirus disease, along with the elderly (WHO, 2021).

Over the past 30 years, Mexico has become one of the countries in the world most heavily affected by the global epidemic of obesity. It is now the second country worldwide with obesity prevalence. Between 2006, 2012 and 2018, overweight or obesity prevalence increased from 69.5% to 71.3% and then to 75.2%, respectively, in population of 20 years and over, while the rate of obesity alone rose from 30% in 2006 to 32.4% in 2012 and then to 36.1% in 2018. Also, Mexico is now one of the countries with the highest child obesity rates in the world with one in three children being overweight or obese. Diabetes, the chronic disease most directly linked with obesity, is spreading rapidly and in Mexico in 2018 affected 10.3% of the adult population—aged over 20 years—, while in 2012 it affected 9.2%.

High blood pressure, or hypertension, has less noticeable symptoms, but if untreated, it increases the risk of serious problems such as heart attacks and strokes. In Mexico, the prevalence of hypertension in 2012 was of 30.2% and in 2018 of 32.7%. It is worth mentioning that these estimates are not directly comparable as there was a methodological change in the National Health and Nutrition Survey[1] (NHNS) of 2018. Specifically, digital baumanometers were introduced; being less susceptible to measurement error, they provide interviewers with better estimates of blood pressure values, see Campos et al. (2019). All these information have been obtained via the NHNS, which is a survey conducted every 6 years by the National Institute of Statistics and Geography, an autonomous agency of the Mexican Government dedicated to coordinate the National System of Statistical and Geographical Information of Mexico.

2.2 | Data and variables

The source of information for this work is the database of the National Epidemiological Surveillance System for monitoring possible cases of COVID-19 in Mexico (SINAVE/SISVER for its acronym in Spanish), coordinated by the Secretariat of Health. The SINAVE/SISVER platform considers cases that are suspected of COVID-19. People who have had flu-like symptoms, or that believe to have been infected with the SARS-CoV-2 virus, are entitled to attend any public or private health service in Mexico, and if after an initial examination is suspected to suffer from the COVID-19 disease, are registered on this database.

By 8 January 2022, this database had information for a total of 12,693,821 suspected cases and 115 variables. We work with the information of 4,064,259 confirmed COVID-19 cases. For

[1] https://www.inegi.org.mx/programas/ensanut/2018/

our analysis we consider relevant information of comorbidities, symptoms, sex, age and disease outcomes, encoded in 35 dichotomous variables. So for each patient we use a vector generically denoted by

$$\boldsymbol{y} = (y_1, y_2, \dots, y_{35}), \tag{1}$$

that we call the COVID-19 footprint.

Variables in (1) are ordered as follows: 1-diabetes, 2-COPD, 3-asthma, 4-immunosuppression, 5-hypertension, 6-heart disease, 7-obesity, 8-chronic kidney failure, 9-smoking, 10-fever, 11-cough, 12-ears pain, 13-breathing difficulty (or shortness of breath), 14-irritability, 15-diarrhoea, 16-chest pain, 17-chills, 18-headache, 19-muscle pain, 20-joint pain, 21-attack general state, 22-nasal discharge, 23-increased respiratory frequency and depth, 24-vomiting, 25-abdominal pain, 26-conjunctivitis, 27-black colour lack of oxygen, 28-sudden onset of symptoms, 29-death, 30-hospitalisation, 31-sex and finally, the four age groups are given by 32- $[0, 20)$, 33- $[20, 40)$, 34- $[40, 60)$ and 35- $[60,)$. Notice that the first nine variables are the comorbidities, while the following 19 variables are the symptoms. For these variables each patient has only two possible values presence 1 or absence 0. Variables 29-death and 30-hospitalisation are the disease outcomes. In these cases $y_{29} = 1$ indicates the patient has died and $y_{30} = 1$ the patient has been hospitalised. The variable with information about the sex of the patient, $y_{31} = 1$ indicates a male patient. Finally, for the age group variables, value 1 indicates membership to the corresponding group.

Clinical knowledge about the COVID-19 disease has evolved during the pandemic and several countries are implementing vaccination programmes. Fortunately, these factors have been influencing hospitalisation and mortality rates of the disease across time. To have a glance of their impact in Mexico we have included the onset date of symptoms in the analysis. We use this variable to select COVID-19 positive patients by month to perform all the analysis as well as to have an idea of the evolution through time.

2.3 | COVID-19 case

Health authorities in Mexico classify a patient as a COVID-19 case if he/she falls in one of the following three categories:

- Confirmed by epidemiological clinical association. This instance applies when the case had contact with a COVID-19 case, and the latter is registered in SISVER platform. The actual case was not tested or the test was invalid.
- Confirmed by an assessment committee. This category only applies to deaths when the case was not tested or a test was taken, but it was invalid.
- Confirmed by SARS-CoV-2 test. The case has a laboratory test or antigenic test and is positive for SARS-CoV-2, regardless of whether the case has a clinical epidemiological association.

2.4 | Open access data

The Secretariat of Health via the General Directorate of Epidemiology (Spanish: Dirección General de Epidemiología) updates daily the data base with the suspected cases and 48 variables, these data can be accessed using the following link https://www.gob.mx/salud/documentos/datos-abiertos-152127. Additional variables are available upon request via the platform http://covid-19.iimas.unam.mx/, as described in Loza et al. (2020).

2.5 | Missing data

The complete data base has 12,693,821 records. Focusing only in COVID-19 cases, there are 4,113,789 patients. At this instance, there are no missing values in the variables with information about deaths, hospitalisations, sex and age. However, if the 19 symptoms and nine comorbidities are included, there are 0.27% of cells, in the $4{,}113{,}789 \times 28$ database, with missing values. The percentage of missing values over the comorbidities ranges from 0.22% in the variable obesity to 0.25% in the variable diabetes, while for the symptoms ranges from 0.22% in the variable breathing difficulty to 0.46% in sudden onset of symptoms. Assuming a missing completely at random mechanism, we followed a case deletion strategy to obtain the 4,064,259 registers that we work in this analysis.

2.6 | Uncertainty and bias

As an additional source of uncertainty we have that the information about the symptoms and comorbidities is mainly self declared by patients and the Mexican health authorities know there must be many patients not fully aware if they suffer from certain comorbidities.

The number of SARS-CoV-2 positive tests has turned into an important indicator, it has been used to decide whether or not nations or regions around the world can open their economies. This is assessed using what has been called the percent positive which is simply the percentage of all coronavirus tests performed that are actually positive. The WHO recommended that the percent positive remain below 5% for at least 2 weeks before governments consider reopening after a lock down period (WHO, 2021). In Mexico, the percent positive has attained 30% and there even have been periods where it almost reached 50%. These very high percentages are easy to explain, the Mexican health authorities operate with limited resources, which forbids a widespread testing. Thus mainly patients with COVID-19 symptoms have been tested. Hence, it is clear that there is bias in the data base of 12,693,821 suspected cases, and is far from a random sample.

All this said, the 4,064,259 positive cases are expected to provide us with important information for inference.

2.7 | Onset date of symptoms

We use the date on which the patient's symptoms began to perform a monthly analysis. Information for a 23-month period from March 2020 to the first days of January 2022 is then used to track the evolution of the COVID-19 pandemic in Mexico. The information of the 4,064,259 COVID-19 positive cases is divided into monthly cases using the onset date of symptoms.

As the number of tests varies over the 23-month period, the infection mortality rate changes, due to improved hospital care, treatments, etc., and other factors also change over time. As it will become evident, the monthly analysis will reduce the bias and thus more reliable inferences will

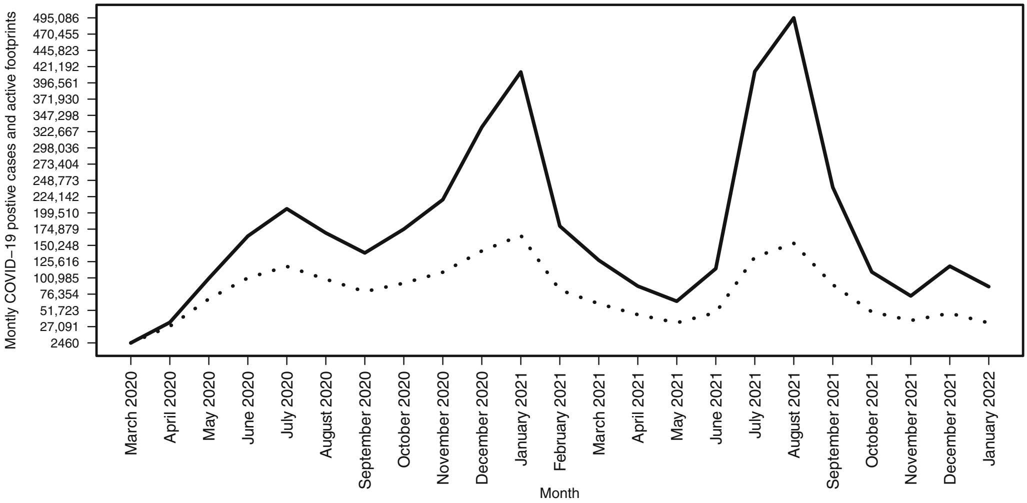


FIGURE 1 Monthly COVID-19 positive cases (continuous line) and active footprints (dotted line) in Mexico using as a reference the onset date of symptoms

be at hand. In Figure 1, the number monthly COVID-19 cases in Mexico, using as a reference the onset date of symptoms, is displayed.

It is easy to identify the so far three waves of the pandemic in Mexico: the first one attaining its peak in July 2020, the second during January 2021 and the last wave (at the time of writing this manuscript) in August 2021. The number of monthly COVID-19 cases ranges from 2,462 in March 2020 when the pandemic started in Mexico to 495,079 cases in August 2021. On average there have been 176,707 monthly cases while the median is of 138,812 positive cases.

3 | THE MODEL

Let $\boldsymbol{Y} = (Y_1, Y_2, \dots, Y_k)$ be a k-dimensional random vector of possibly correlated Bernoulli random variables and $\boldsymbol{y} = (y_1, y_2, \dots, y_k)$ a realisation of it. The multivariate Bernoulli distribution can be described via its mass probability function

$$\begin{aligned}
p(\boldsymbol{y}|\boldsymbol{w}) = {} & w(0,0,0,\dots,0)^{\left\{\prod_{j=1}^{k}(1-y_j)\right\}} \\
& \times w(1,0,0,\dots,0)^{\left\{y_1 \prod_{j\neq 1}(1-y_j)\right\}} \\
& \times w(0,1,0,\dots,0)^{\left\{y_2 \prod_{j\neq 2}(1-y_j)\right\}} \\
& \times \dots \qquad \dots \qquad \dots \\
& \times w(1,1,0,\dots,0)^{\left\{y_1 y_2 \prod_{j\neq 1,2}(1-y_j)\right\}} \\
& \times \dots \qquad \dots \qquad \dots \\
& \times w(1,1,1,\dots,1)^{\left\{\prod_{j=1}^{k} y_j\right\}}.
\end{aligned}$$

where

$$\boldsymbol{w} = (w(0,0,0,\dots,0), w(1,0,0,\dots,0), w(0,1,0,\dots,0), \dots, w(1,1,1,\dots,1)), \tag{2}$$

is the vector of probabilities associated to each possible outcome, and there are $\mathbb{S} := 2^k$ possible outcomes. Observe the probability vector (2) sums up to 1.

Taking $\boldsymbol{\rho} = \left\{ \{\rho_{l,j}\}_{l=1}^{\mathbb{S}} \right\}_{j=1}^{k}$ as the matrix where the rows are the $\mathbb{S}$ possible outcomes of the random vector $\boldsymbol{Y}$ and $\rho_{l,\cdot}$ as the k-dimensional vector of the l possible outcome. It is possible to write the above mass probability function in shorter forms, that is

$$
\begin{aligned}
p(\boldsymbol{y}|\boldsymbol{w}) &= \sum_{l=1}^{\mathbb{S}} w(\rho_{l,\cdot})\ \gamma_l(\boldsymbol{y}), && (3)\\
&= \prod_{l=1}^{\mathbb{S}} w(\rho_{l,\cdot})^{\gamma_l(\boldsymbol{y})}, && (4)
\end{aligned}
$$

where $\gamma_l(\boldsymbol{y}) = \mathbb{1}(\rho_{l,.} = \boldsymbol{y})$. Thus $\gamma_l(\boldsymbol{y})$ indicates whether $\boldsymbol{y}$ is the lth outcome or not of the $\mathbb{S}$ possible outcomes. Notice that expression (4) defines a function to map the outcome $\boldsymbol{y}$ to its corresponding success probability.

3.1 | Properties

A multivariate Bernoulli model will be assumed to describe the footprint of each patient that has been identified as COVID-19 positive, see (1). Thus, in our case $\mathbb{S} = 2^{35}$, namely all possible footprints. However, in the analysis section below we focus our attention on describing marginal relationships; for example, the impact of sex, age group and each comorbidity or symptom on death and hospitalisation outcomes. Also, under a prediction setting the aim is to find the smallest subset of variables with good predictive power. This clearly implies a dimension reduction of our multivariate Bernoulli random variable, for example, $s < k$, which consequently reduces the cardinality of the support $\mathbb{S}$, say to $\mathbb{S}'$. Considering these and further scenarios, it is important to know that marginal distributions of any order of a multivariate Bernoulli distribution are still in the same family.

To ease notation, we will denote by $\boldsymbol{Y}'$ the random variable with reduced dimension, $s < k$, and correspondingly its values, $\boldsymbol{y}'$, parameters $\boldsymbol{w}'$ and support values $\rho'_{\cdot,\cdot}$.

Thus,

$$
\begin{aligned}
p(\boldsymbol{y}'|\boldsymbol{w}') &= \sum_{y_{s+1}=0}^{1} \cdots \sum_{y_k=0}^{1} p(\boldsymbol{y}|\boldsymbol{w}),\\
&= \prod_{r=1}^{\mathbb{S}'} w'(\rho'_{r,.})^{\gamma_l(\boldsymbol{y}')}, && (5)
\end{aligned}
$$

is a multivariate Bernoulli distribution where

$$
w'(\rho'_{r,.}) = \sum_{l=1}^{\mathbb{S}} \left[\mathbb{1}(\rho_{l,1:s} = \rho'_{r,.})\ w(\rho_{l,.}) \right],\ \text{for } r = 1, \ldots, \mathbb{S}',
$$

with $\rho_{l,1:s}$ as the first s observations of the l outcome of the random variable $\boldsymbol{Y}$, and $\rho'_{r,.}$ as the s dimensional vector of the r possible outcome of the marginalised random variable $\boldsymbol{Y}'$.

As a straightforward consequence, for $j \in \{1, \dots, k\}$, we have

$$p(Y_j = y_j|\boldsymbol{w}) = \theta_j^{y_j}(1-\theta_j)^{1-y_j}, \tag{6}$$

where

$$\theta_j = \sum_{l=1}^{\mathbb{S}} \mathbb{1}(\rho_{l,j} = 1)\, w(\rho_{l,\cdot}). \tag{7}$$

Thus each Y_j follows a Bernoulli distribution as expected, and to obtain the probability of success we need to add up the probabilities of all the footprints where $Y_j = 1$.

Finally, in this brief outline, the covariance between any two random variables Y_j and Y_s of the random vector $\boldsymbol{Y}$, is given by

$$\begin{aligned}\mathrm{Cov}(Y_j, Y_s) &= p(Y_j = 1, Y_s = 1|\boldsymbol{w}) - \theta_j\theta_s,\\ &= \theta_{j,s} - \theta_j\theta_s,\end{aligned}$$

where

$$\theta_{j,s} = \sum_{l=1}^{\mathbb{S}} \mathbb{1}(\rho_{l,j} = 1)\mathbb{1}(\rho_{l,s} = 1)\, w(\rho_{l,\cdot}). \tag{8}$$

Then, the correlation is given by

$$\mathrm{Corr}(Y_j, Y_s) = \frac{\theta_{j,s} - \theta_j\theta_s}{\sqrt{\theta_j(1-\theta_j)}\sqrt{\theta_s(1-\theta_s)}}, \tag{9}$$

where the numerator can take negative or positive values. Indeed, $\mathrm{Corr}(Y_j, Y_s)$ can potentially take any values in $[-1, 1]$.

3.2 | Bayesian inference

Let $\boldsymbol{Y}^{(n)} := (\boldsymbol{Y}_1, \boldsymbol{Y}_2, \dots, \boldsymbol{Y}_n)$ be a random sample of size n from the multivariate Bernoulli distribution (4). Then, the likelihood is given by

$$\begin{aligned}\prod_{i=1}^{n} p(\boldsymbol{y}_i|\boldsymbol{w}) &= \prod_{i=1}^{n}\left[\prod_{l=1}^{\mathbb{S}} w(\rho_{l,\cdot})^{\gamma_l(\boldsymbol{y}_i)}\right],\\ &= \prod_{l=1}^{\mathbb{S}} w(\rho_{l,\cdot})^{r_l},\end{aligned}$$

where

$$r_l = \sum_{i=1}^{n}\gamma_l(\boldsymbol{y}_i) = \sum_{i=1}^{n}\mathbb{1}(\rho_{l,\cdot} = \boldsymbol{y}_i), \text{ for } l = 1, \dots, \mathbb{S}. \tag{10}$$

Hence, r_l counts the number of occurrences of the lth possible footprint in the data.

If a Dirichlet distribution is chosen as a prior distribution over $\boldsymbol{w}$, that is,

$$p(\boldsymbol{w}|\delta_1, \dots, \delta_{\mathbb{S}}) = \mathrm{Dir}(\boldsymbol{w}|\delta_1, \dots, \delta_{\mathbb{S}}), \\ \propto \prod_{l=1}^{\mathbb{S}} w(\rho_{l,\cdot})^{\delta_l - 1}, \tag{11}$$

Thus δ_l can be thought as the prior pseudocount assigned to the lth footprint. Now is trivial to obtain the posterior distribution, which is given by

$$p\left(\boldsymbol{w}|\boldsymbol{y}^{(n)}\right) \propto p(\boldsymbol{w}|\delta_1, \dots, \delta_{\mathbb{S}}) \prod_{i=1}^{n} p\left(\boldsymbol{y}_i|\boldsymbol{w}\right), \\ \propto \mathrm{Dir}(\boldsymbol{w}|\alpha_1, \dots, \alpha_{\mathbb{S}}). \tag{12}$$

where $\alpha_l = r_l + \delta_l$, for $l = 1, \dots, \mathbb{S}$. In other words, we have a conjugate model and α_l is an update to the pseudocount of footprint lth once the data have been observed.

3.3 | Posterior probabilities

Due to the marginalisation properties of the multivariate Bernoulli and also of the Dirichlet distribution, it is easy to obtain, for example, the posterior probability of death and hospitalisation, or that a patient suffers certain comorbidities within the COVID-19 positive cases. This, is done by computing

$$p\left(Y_j = 1|\boldsymbol{y}^{(n)}\right) \overset{d}{=} \left\{\Theta_j|\boldsymbol{y}^{(n)}\right\} \sim \mathrm{Beta}(\theta_j|\eta_j, \alpha_0 - \eta_j), \tag{13}$$

where

$$\alpha_0 = \sum_{l=1}^{\mathbb{S}} \alpha_l, \quad \text{and} \quad \eta_j = \sum_{l=1}^{\mathbb{S}} \mathbb{1}(\rho_{l,j} = 1)\alpha_l,$$

with Θ_j is the random variable corresponding to statistic (7). If only a point estimate is needed, the posterior expectation is easily obtained as

$$\mathbb{E}\left(\Theta_j|\boldsymbol{y}^{(n)}\right) = \frac{\eta_j}{\alpha_0}. \tag{14}$$

Bivariate joint and conditional probabilities between any two dichotomous variables are straightforward as well, that is,

$$p\left(Y_j = 1, Y_s = 1|\boldsymbol{y}^{(n)}\right) \overset{d}{=} \left\{\Theta_{j,s}|\boldsymbol{y}^{(n)}\right\} \sim \mathrm{Beta}(\theta_{j,s}|\eta_{j,s}, \alpha_0 - \eta_{j,s}), \tag{15}$$

where, analogously, $\Theta_{j,s}$ denotes the random variable corresponding to statistic (8), and

$$\eta_{j,s} = \sum_{l=1}^{\mathbb{S}} \mathbb{1}(\rho_{l,j} = 1)\mathbb{1}(\rho_{l,s} = 1)\alpha_l.$$

In Appendix A it is shown that

$$p\left(Y_s = 1 | Y_j = 1, \boldsymbol{y}^{(n)}\right) \overset{d}{=} \left\{ \frac{\Theta_{j,s}}{\Theta_j} | \boldsymbol{y}^{(n)} \right\} \sim \text{Beta}\left(\frac{\theta_{j,s}}{\theta_j} | \eta_{j,s}, \eta_j - \eta_{j,s} \right). \tag{16}$$

and then

$$\mathbb{E}\left(\frac{\Theta_{j,s}}{\Theta_j} | \boldsymbol{y}^{(n)} \right) = \frac{\eta_{j,s}}{\eta_j}. \tag{17}$$

These results are easily extended to

$$p\left(Y_{j_1} = 1, \dots, Y_{j_r} = 1, Y_{s_1} = 1, \dots, Y_{s_u} = 1 | \boldsymbol{y}^{(n)}\right),$$

and $p\left(Y_{j_1} = 1, \dots, Y_{j_r} = 1 | Y_{s_1} = 1, \dots, Y_{s_u} = 1, \boldsymbol{y}^{(n)}\right)$. Obtaining beta distributions as well.

4 | COVID-19 FOOTPRINT AND MORTALITY

In this section our purpose is to study the relationship between comorbidities, symptoms, sex and age, with death and hospitalisation outcomes for all those patients that resulted COVID-19 positive in Mexico. We will assume that the monthly case information is a random sample from a multivariate Bernoulli distribution. In the 23 months from March 2020 to the first days of January 2022, the monthly number of COVID-19 cases ranged from a minimum of 2,462 cases to a maximum 495,079 cases, see Figure 1. By considering the order of any relevant variables, described in the paragraph after expression (1), all the posterior distributions described in the previous section can be obtained. We will repeat our analysis monthly, so given a footprint sample, $\boldsymbol{y}^{(n)}$, in a particular month, for any COVID-19 positive patient the posterior distribution of, for example, a death outcome, given the patient suffers from diabetes, is obtained computing $p(Y_{29} = 1 | Y_1 = 1, \boldsymbol{y}^{(n)})$. Similarly, if interest lies in the posterior probability of hospitalisation given the patient is a 37-year-old male that suffers from obesity we compute $p(Y_{30} = 1 | Y_7 = 1, Y_{31} = 1, Y_{33} = 1, \boldsymbol{y}^{(n)})$.

4.1 | Computational shortcut

Our theoretical sample space is of $\mathbb{S} := 2^{35}$ possible footprints[2]. The key to handle this huge support is simply to identify the active footprints in (10), that is, the unique different footprints that have been observed in the data. Denoting by $\beta_{l,\cdot}$ for $l = 1, \dots, m$, the active footprints, we can order the rows of the matrix with all the possible footprints such that in the first m rows we have the active footprints, thus $\rho_{l,\cdot} = \beta_{l,\cdot}$ for $l = 1, \dots, m$. With this we can write (10) as

$$r_l = \sum_{i=1}^{n} \mathbb{1}(\rho_{l,\cdot} = \boldsymbol{y}_i) = \sum_{i=1}^{n} \mathbb{1}(\beta_{l,\cdot} = \boldsymbol{y}_i), \text{ for } l = 1 \dots, m, \tag{18}$$

[2] It is worth noting that for the four variables of age group there are only four possible results, namely the same person cannot belong to two different age groups at the same time. Hence, the true sample space of our model is reduced to 2^{32} outcomes.

and we known that $r_l = 0$, for $l > m$. Clearly, $\alpha_l = r_l + \delta_l$, for $l = 1 \ldots, m$, and $\alpha_l = \delta_l$ for $l = m + 1, \ldots, \mathbb{S}$.

Notice that the number of occurrences for every outcome in the matrix $\boldsymbol{\beta}$ can be calculated efficiently. First, the footprint of each patient is concatenated into a single variable. Second, obtain the number of counts for each category in this new variable. This, gives us the active footprints with their respective number of occurrences. Third, the concatenated variable is separated into dichotomous variables. In particular, with the programming language R (R Core Team, 2021), using the function `unite`, from the library `tidyr` (see Wickham, 2021) and the functions `table` and `merge` from the base library this is done efficiently. From the 4,064,259 cases, we have observed $m = 941{,}943$ active footprints. However, remember that to track the evolution of the COVID-19 pandemic we perform a monthly analysis. In Figure 1 the monthly number of active footprints are displayed (dotted line). These are clearly correlated to the number of monthly observed cases (continuous line).

4.2 | Model specification

We will assume $\delta_l = \delta$, for $l = 1, \ldots, \mathbb{S}$, in the Dirichlet prior distribution for the weights in (11). Thus giving a symmetric prior distribution over all possible footprints. Under the active footprint setting, see (18), the update for the pseudo-counts is given by $\alpha_l = r_l + \delta$, for $l = 1 \ldots, m$, and $\alpha_l = \delta$, for $l = m + 1, \ldots, \mathbb{S}$.

The study of the symmetric Dirichlet prior distribution dates back to Good (1965), where it was used as a prior distribution over the probabilities in the Multinomial model. Good observed that when setting $\delta = 0$, the posterior expectation corresponds to the MLE estimator. Interest lied in the cases $\delta = 1/2, 1$ and $1/H$, with H the number of categories, and also placing a prior distribution over δ. In the context of finite mixture models, the symmetric Dirichlet distribution is often used as a prior distribution over the mixture weights. Under this latter setting, usually latent allocation variables indicating to which component each observation belongs are introduced. These latent allocations follow a Multinomial distribution with the mixture weights as the vector of probabilities. See Frühwirth-Schnatter et al. (2019) for a recent review of mixture models. Thus, we are again in Good's framework. For finite mixture-based clustering, Ishwaran and Zarepour (2020) observed that if $\delta = 1$ is used, then the tendency of the model is to allocate observations in many different components. Instead if $\delta = \frac{\nu}{H}$ is chosen, where now H is the number of mixture components, the allocation of observations favours a few dominant clusters. Finally, they recommend $\nu = 1$, namely $\delta = \frac{1}{H}$, as a reasonable minimally informative default. More recent work in mixture models, for an unknown number of components, first sets an unusually high value for H and then proposes a gamma prior distribution over δ, where the mean of the gamma distribution is fixed at a value close to zero. These are called sparse finite mixture models (see Frühwirth-Schnatter and Malsiner-Walli, 2019).

In our case, with monthly sample sizes ranging from 2,462 to 495,079 patients, observed active footprints ranged between 2,350 and 165,356 out of all the $\mathbb{S} = 2^{35}$ possibilities; as depicted in Figure 1. Considering the mixture modelling ideas outlined above, it is reasonable to assume $\delta = \nu/\mathbb{S}$. To set ν, our benchmark will be the posterior distributions (13) and (15). First, it is easy to see that

$$\alpha_0 = \sum_{l=1}^{m} r_l + \delta(2^{\mathbb{S}}) = n + \nu.$$

Second, computing

$$\begin{aligned}
\eta_j &= \sum_{l=1}^{m} \mathbb{1}(\beta_{l,j} = 1) r_l + \delta \sum_{l=1}^{\mathbb{S}} \mathbb{1}(\rho_{l,j} = 1),\\
&= \sum_{l=1}^{m} \mathbb{1}(\beta_{l,j} = 1) \sum_{i=1}^{n} \mathbb{1}(\beta_{l,\cdot} = \boldsymbol{y}_i) + (\delta)(2^{35-1}),\\
&= \sum_{i=1}^{n} \mathbb{1}(y_{i,j} = 1) + (\delta)(2^{35-1}),\\
&= \sum_{i=1}^{n} \mathbb{1}(y_{i,j} = 1) + \nu/2,
\end{aligned}$$

and

$$\begin{aligned}
\eta_{j,s} &= \sum_{l=1}^{\mathbb{S}} \mathbb{1}(\rho_{l,j} = 1)\mathbb{1}(\rho_{l,s} = 1)(r_l + \delta),\\
&= \sum_{i=1}^{n} \mathbb{1}(y_{i,j} = 1)\mathbb{1}(y_{i,s} = 1) + \delta \sum_{l=1}^{\mathbb{S}} \mathbb{1}(\rho_{l,j} = 1)\mathbb{1}\left(\rho_{l,s}^{35} = 1\right),\\
&= \sum_{i=1}^{n} \mathbb{1}(y_{i,j} = 1)\mathbb{1}(y_{i,s} = 1) + (\delta)(2^{35-2}),\\
&= \sum_{i=1}^{n} \mathbb{1}(y_{i,j} = 1)\mathbb{1}(y_{i,s} = 1) + \nu/4,
\end{aligned}$$

we obtain

$$\mathbb{E}\left(\Theta_j | \boldsymbol{y}^{(n)}\right) = \frac{\sum_{i=1}^{n} \mathbb{1}(y_{i,j} = 1) + \nu/2}{n + \nu}, \tag{19}$$

and

$$\mathbb{E}\left(\frac{\Theta_{j,s}}{\Theta_j} | \boldsymbol{y}^{(n)}\right) = \frac{\sum_{i=1}^{n} \mathbb{1}(y_{i,j} = 1)\mathbb{1}(y_{i,s} = 1) + \nu/4}{\sum_{i=1}^{n} \mathbb{1}(y_{i,j} = 1) + \nu/2}. \tag{20}$$

Hence, if $\nu = 1$, that is $\delta = \frac{1}{\mathbb{S}}$, our inference would be exclusively based on the information available. However, to give weight to the unobserved footprints we have set $\nu = 100$, and our inference will be heavily based on the information available but giving a small weight to the unobserved footprints.

5 | PREDICTION

If a new patient has been diagnosed as COVID-19 positive, a natural question that arises is that of quantifying his/her death probability given a realisation of his/her 'clinical footprint', $\boldsymbol{y}^*$, namely where all but death outcomes are recorded. Remember that in the variable ordering described in the introduction, Y_{29} corresponds to the death outcome. Then, to obtain the predictive probability of death, we compute

$$p\big(Y_{29}=1|\boldsymbol{Y}=\boldsymbol{y}^{*},\ \boldsymbol{y}^{(n)}\big)=\frac{p\big(\boldsymbol{Y}=\boldsymbol{y}^{*,1}|\boldsymbol{y}^{(n)}\big)}{\sum_{r=0}^{1}p\big(\boldsymbol{Y}=\boldsymbol{y}^{*,r}|\boldsymbol{y}^{(n)}\big)},$$

where

$$\begin{aligned}\boldsymbol{y}^{*,0}&=\big(y_1^{*},\ \dots\ ,y_{28}^{*},0,y_{30}^{*},\ \dots\ ,y_{35}^{*}\big),\\ \boldsymbol{y}^{*,1}&=\big(y_1^{*},\ \dots\ ,y_{28}^{*},1,y_{30}^{*},\ \dots\ ,y_{35}^{*}\big).\end{aligned}$$

As before, $\boldsymbol{Y}$ denotes the 35-dimensional random vector.

On the other hand, if one is looking for the full predictive footprint, that is, not only his/her death outcome, this can be computed as

$$\begin{aligned}p\big(\boldsymbol{Y}=\boldsymbol{y}^{*}|\boldsymbol{y}^{(n)}\big)&=\int p(\boldsymbol{Y}=\boldsymbol{y}^{*}|\boldsymbol{w})p\big(\boldsymbol{w}|\boldsymbol{y}^{(n)}\big)\mathrm{d}\boldsymbol{W},\\ &\propto\int w(\boldsymbol{y}^{*})\prod_{l=1}^{\mathbb{S}}w(\rho_{l,.})^{\alpha_l-1}\ \mathrm{d}\boldsymbol{W},\\ &=\int\prod_{l=1}^{\mathbb{S}}w(\rho_{l,.})^{\alpha_l+\mathbb{1}(\rho_{l,.}=\boldsymbol{y}^{*})-1}\ \mathrm{d}\boldsymbol{W},\\ &\propto\frac{\alpha_{l*}+1}{\alpha_0},\end{aligned}\tag{21}$$

where $\boldsymbol{W}$ refers to the $\mathbb{S}-1$-dimensional simplex, and with $l*$ denoting the row of the matrix consisting of the $\mathbb{S}$ possible outcomes such that $\mathbb{1}(\rho_{l*,.}=\boldsymbol{y}^{*})=1$.

Then,

$$p(Y_{29}=1|\boldsymbol{Y}=\boldsymbol{y}^{*},\boldsymbol{y}^{(n)})=\frac{\alpha_{l_1*}+1}{\alpha_{l_1*}+\alpha_{l_2*}+2},\tag{22}$$

where l_1* is the row of the matrix with the $\mathbb{S}$ possible outcomes such that $\mathbb{1}(\rho_{l_1*,.}=\boldsymbol{y}^{*,1})=1$, and l_2* is such that $\mathbb{1}(\rho_{l_2*,.}=\boldsymbol{y}^{*,0})=1$.

Going back to expression (21) and considering the active footprint setting, expression (18). We can obtain α_{l*} in an easier manner, namely

$$\alpha_{l*}=\begin{cases}r_{l*}+\delta, & \text{if } \boldsymbol{y}^{*} \text{ is in the active footprints and } \mathbb{1}(\beta_{l*,.}=\boldsymbol{y}^{*})=1.\\ \delta, & \text{otherwise.}\end{cases}$$

For predicting death or hospitalisation outcomes, the purpose will be to find the subset of variables that are most closely related to these outcomes. With a model with a few variables we are ignoring irrelevant variables that decrease the precision and increase the complexity of the model. Thus, we will apply the exact same ideas as before with the obvious modifications.

6 | ANALYSIS

Focusing the analysis on the monthly positive cases, here we compute mortality risk given a particular footprint configuration. Also the expected mortality, correlation between variables among other posterior summaries are obtained. The section is closed with a prediction of death outcomes.

6.1 | Mortality risk

To identify the comorbidities and symptoms that lead to a greater mortality risk, we need to obtain the posterior expectations $\mathbb{E}\left(\frac{\Theta_{j,29}}{\Theta_j}|\boldsymbol{y}^{(n)}\right)$, for $j = 1, \dots, 28$ by month. See expression (20). In Figure 2, we display the time series of the twenty variables with higher expected mortality risk.

First, it is important to note that the coloured labels in Figure 2 follow a decreasing order, considering the average expected mortality risk over the 23 months. Thus, for example, chronic kidney failure, hospitalisation, COPD, bluish colouration due to lack of oxygen, etc., are the comorbidities or symptoms that lead to higher mortality risk. Second, in these labels we have included a "-c" for comorbidity and a "-s" for symptom to differentiate each condition. The symptoms in the group with higher mortality risk are all related to lungs becoming inflamed, while the comorbidities range from chronic kidney failure, COPD and heart disease.

Furthermore, considering the strength of the pandemic waves, see Figure 1, it is also possible to observe the three waves through Figure 2. However, even during the third wave (the highest) there is a decreasing trend of the mortality risk.

It is straightforward to generate this information desegregated by sex and age group. In this case we need to calculate,

$$\mathbb{E}\left[P\left(Y_{29} = 1|Y_j = 1, Y_{31} = s, (Y_{32}, Y_{33}, Y_{34}, Y_{35}) = g, \boldsymbol{y}^{(n)}\right)\right], \text{ for } j = 1 \dots, 28, \tag{23}$$

where $s \in \{0, 1\}$ and $g \in \{(1, 0, 0, 0), (0, 1, 0, 0), (0, 0, 1, 0), (0, 0, 0, 1)\}$. This is displayed in Figure 3, where we only show the ten comorbidities or symptoms with higher mortality risk across the 23 months by sex and age group.

As described in the Section 4.2, we assume $\nu = 10$ $(\delta = 10/\mathbb{S})$ for the analysis. However, it is important to mention that for the age groups $[0, 20)$ and $[20, 40)$ inference changes drastically if we take, for example, $\nu = 1$ $(\delta = 1/\mathbb{S})$ for the symmetric Dirichlet prior distribution. This happens

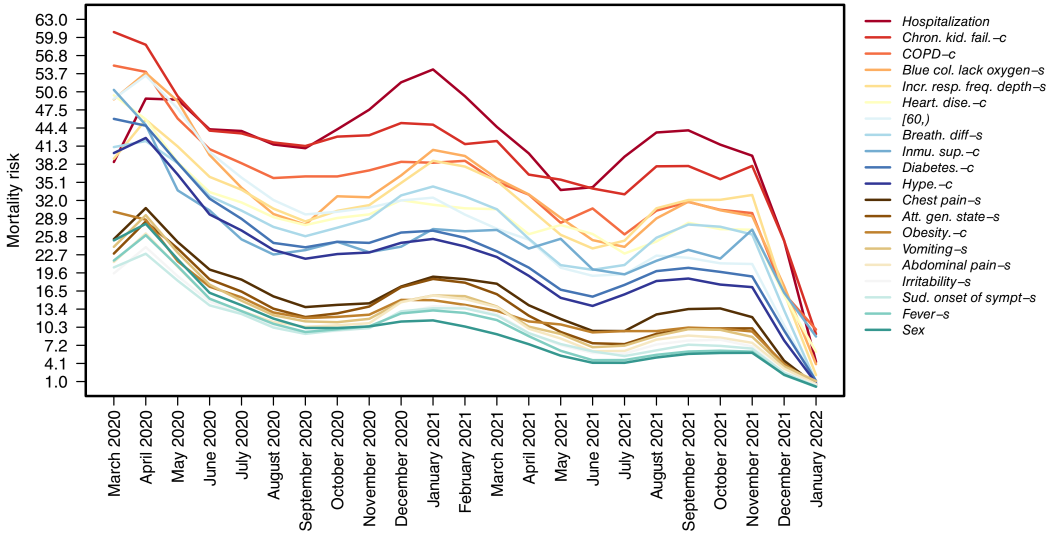


FIGURE 2 Monthly expected mortality risk for the 20 variables with highest mortality risk, that is, posterior expectation $\mathbb{E}\left(\frac{\Theta_{j,29}}{\Theta_j}|\boldsymbol{y}^{(n)}\right)$ for $j = 1, \dots, 28$. We are only displaying the 20 variables with the higher risk. The coloured labels follow a decreasing order, considering the average expected mortality risk over the 23 months. [Colour figure can be viewed at wileyonlinelibrary.com]

because the number of monthly COVID-19 cases that fall into these groups intersected with most of the comorbidities is very small. Thus, inference for these two groups exhibits great variability. Instead, inference for the age groups [40, 60) and 60 and more, is more reliable as there are large sample sizes. This can be also seen in Figure 3.

Considering the monthly estimates for each age group the trend of the mortality risk for age groups [0, 20) and [20, 40) changes drastically. However, it clearly decreases on age groups [40, 60)

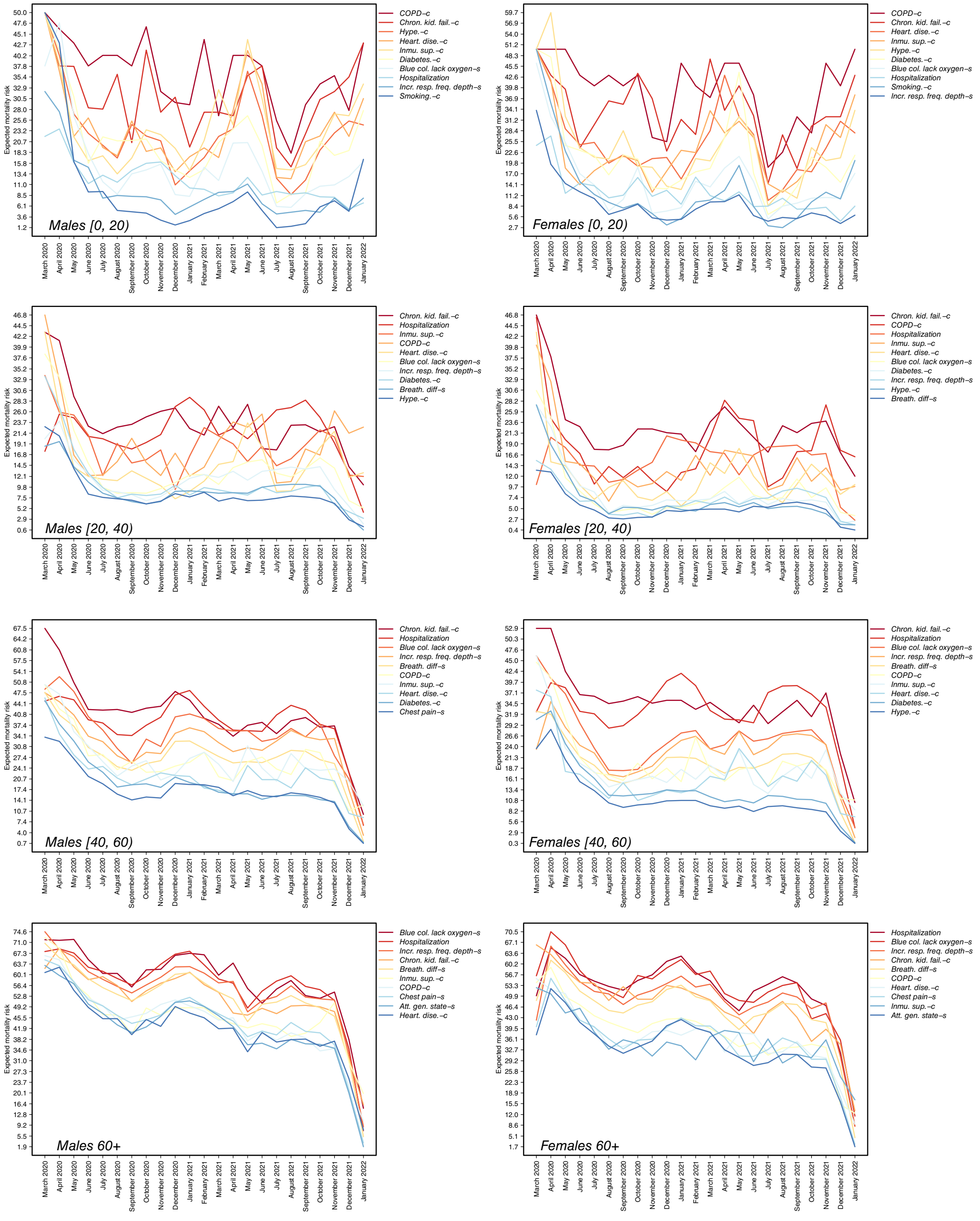


FIGURE 3 Expected mortality risk given comorbidities/symptoms, sex and age. The coloured labels for each graphic follow a decreasing order, considering the average expected mortality risk over the 23 months. [Colour figure can be viewed at wileyonlinelibrary.com]

and 60 and more. This is for all given comorbidities or symptoms. Sex is also a key variable, and it can be observed from its clear impact on ages [40, 60) and 60 and more. Female patients always show lower mortality risks, which is easily observed in the range of the graphics. Other interesting findings from Figure 3, and not considering hospitalisations, are the following:

- Cases between 40 and 60 years old with chronic kidney failure (comorbidity) are associated with the highest mortality risk for males and females, then these associate with a bluish colouration due to lack of oxygen (symptom) and an increased respiratory rate and depth (symptom), in second and third places, respectively.
- For cases above 60 years old, bluish colouration due to lack of oxygen (symptom) is associated with the highest mortality risk for males and females, then for males and females follow increased respiratory rate and depth (symptom), and the third is chronic kidney failure (comorbidity).

6.2 | COVID-19 mortality

Let us assume that $\nu = 100{,}000$ people in Mexico were diagnosed as COVID-19 case. To appreciate which comorbidities and/or symptoms have greater impact over COVID-19 mortality, from these ν cases, we can approximate the number of people who suffer from comorbidity (or symptom) j, and this is given by $\nu P(Y_j = 1|\boldsymbol{y}^{(n)})$. Then to approximate the mortality within those with comorbidity j would be $\nu P(Y_j = 1|\boldsymbol{y}^{(n)})P(Y_{29} = 1|Y_j = 1, \boldsymbol{y}^{(n)}) = \nu P(Y_{29} = 1, Y_j = 1|\boldsymbol{y}^{(n)})$. Instead of working with with $\nu P(Y_{29} = 1, Y_j = 1|\boldsymbol{y}^{(n)})$ we obtain its expectation and this is shown in Figure 4.

The expected mortality is decreasing on all cases. Hospitalisation is clearly associated with the highest expected mortality, however it is neither a comorbidity nor a symptom. Following hospitalisation, there are five symptoms, and the one that leads to the a greater mortality is breathing difficulty. The comorbidities that have greater impact over the mortality are hypertension and

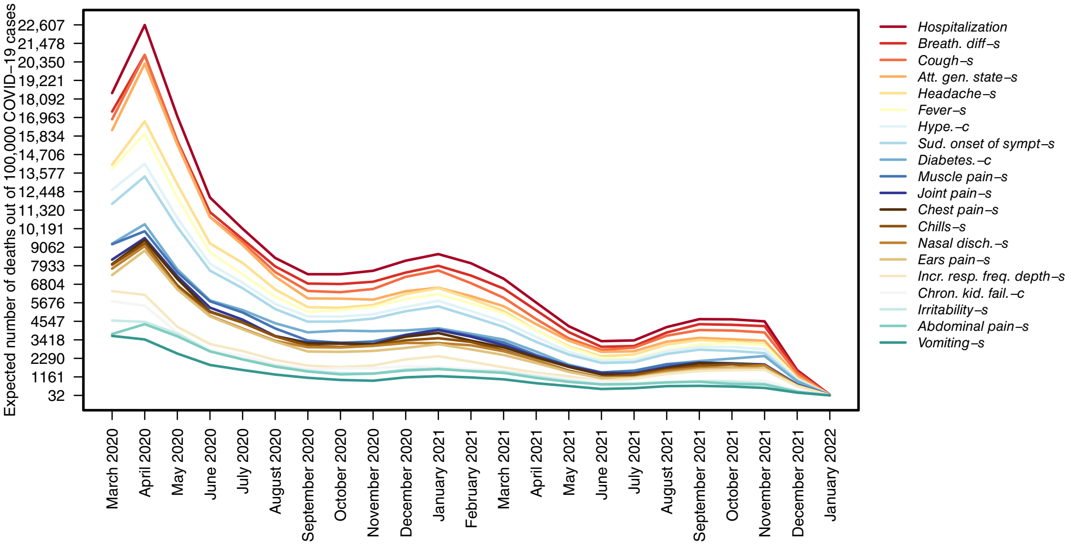


FIGURE 4 Expected mortality related to each comorbidity or symptom out of 100,000 cases. The coloured labels follow a decreasing order, considering the average expected mortality over the 23 months. [Colour figure can be viewed at wileyonlinelibrary.com]

diabetes. This is of great concern to health authorities due to the enormous problem of diabetes and hypertension in Mexico. Now we can do the complete exercise to estimate the mortality related to each comorbidity (or symptom) by sex and age group. In this case we compute,

$$\nu\, \mathbb{E}\big(P\big(Y_{29} = 1, Y_j = 1, Y_{31} = s, (Y_{32}, Y_{33}, Y_{34}, Y_{35}) = g|\boldsymbol{y}^{(n)}\big)\big),$$

where all the indexes are as in expression (23). This is displayed in Figure 5, where we only show the 10 comorbidities or symptoms with higher mortality across the 23 months by sex and age group.

Figure 3 shows that the mortality risk exhibited great variability for younger age groups; however the expected mortality out of 100,000 cases shows a steep decrease trend for both male and female patients and in every age group. This is given for all comorbidities or symptoms. Again, as in the case of the mortality risk, female patients show lower mortality than males. Indeed, the mortality of male patients is twice (or more) than that of female patients.

Finally, it is important to observe that symptoms largely determine mortality, while comorbidities start to gain relevance for older age groups. However, it would be important to understand how the presence of certain comorbidities exacerbates some symptoms, in particular breathing difficulty. This is done in the correlation section.

6.3 | Impact of sex and age of the patients over symptoms

In this section patterns of symptoms determined by sex and age group are identified. This is done only for the COVID-19 cases in August 2021, which is the highest registered number of cases among the three waves of the pandemic that Mexico has suffered to date, see Figure 1. Thus, Figure 6 shows the expected probability of each symptom given sex and age group and a heat map along with a hierarchical clustering is also included.

We have seen that shortness of breath (or breathing difficulty) is the symptom leading to the highest mortality, see Figure 4. The hierarchical clustering indicates the symptoms diarrhoea, irritability, chest pain and breathing difficulty often appear together, and increases with age and have greater impact over male patients. Cough, fever and headache are the most common symptoms, but the probability of having the first two increases marginally with age. Also, the expected probability of the symptoms such as chest pain and increased respiratory frequency and depth increases with age, which is for both male and female patients. On the other hand, the probability of having nasal discharge decreases with age.

6.4 | Correlation between comorbidities, symptoms, hospitalisation and deaths

The multivariate Bernoulli model give us the possibility to obtain the correlation between the comorbidities, deaths and hospitalisations. This is done using Equation (9), but here we need to approximate the posterior mean

$$\mathbb{E}\left(\mathrm{Corr}(Y_j, Y_s)|\boldsymbol{y}^{(n)}\right) = \mathbb{E}\left(\frac{\Theta_{j,s} - \Theta_j\Theta_s}{\sqrt{\Theta_j(1-\Theta_j)}\sqrt{\Theta_s(1-\Theta_s)}}|\boldsymbol{y}^{(n)}\right),$$

via classic Monte Carlo, and the idea is described in Appendix B.

With the matrix of pairwise correlations a hierarchical clustering is used to obtain groups of variables closely associated to each other. In Figure 7 the correlation plot and the groups are displayed. This is done only for the COVID-19 cases of August 2021, however the correlation matrices for the 23 months can be obtained using the code given as Data S1.

The correlation plot indicates that there are groups of variables closely related to each other. The group containing the outcomes hospitalisations and deaths includes the comorbidities COPD,

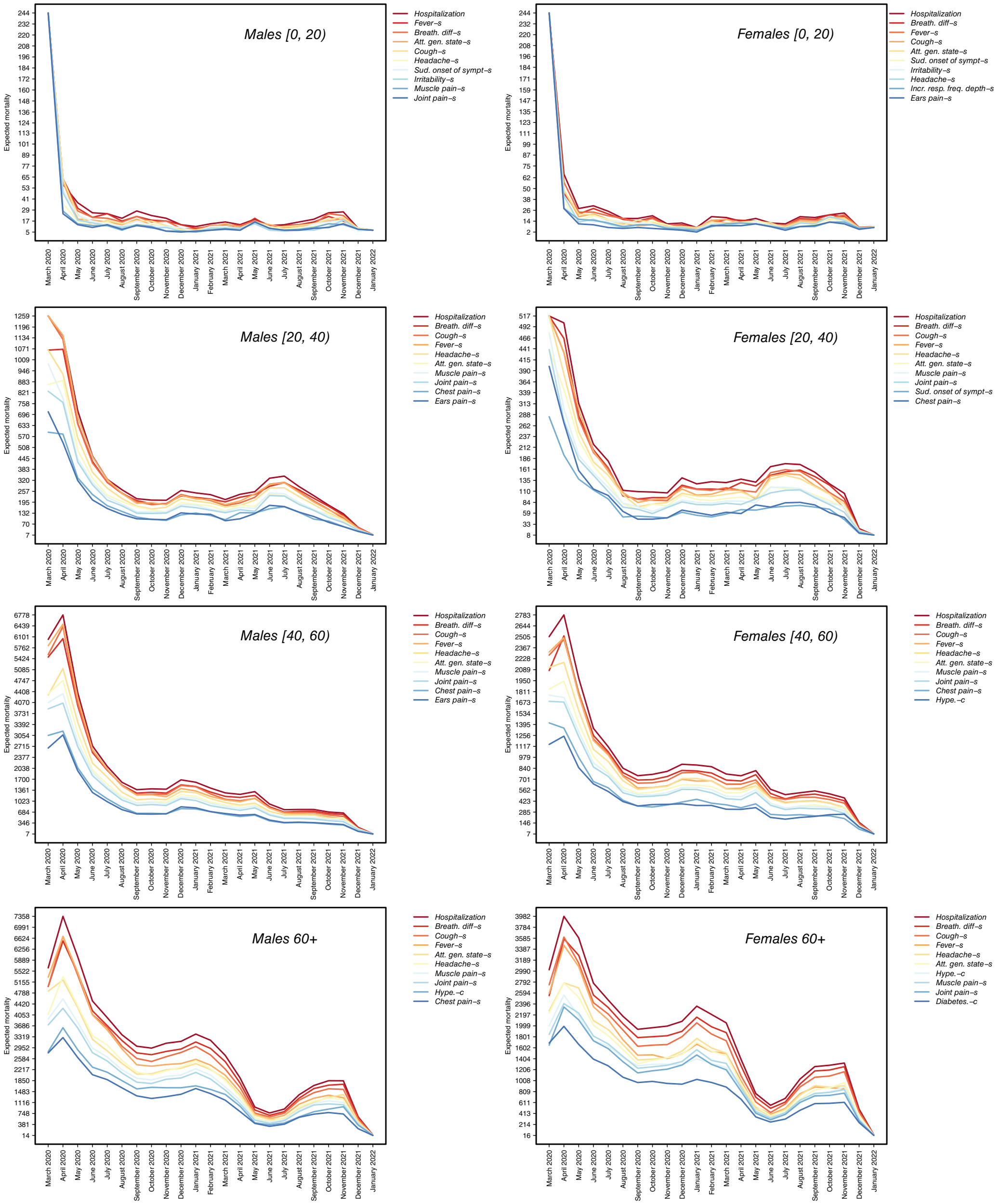
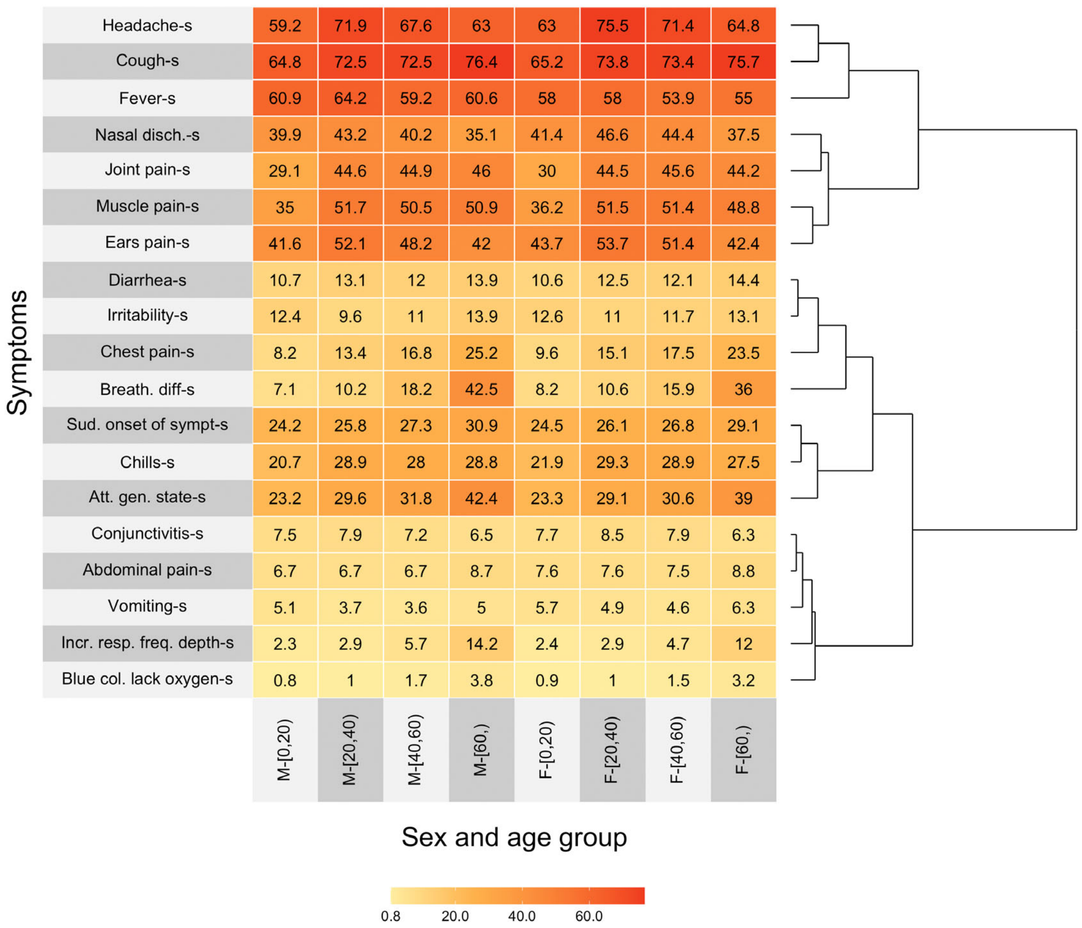


FIGURE 5 Expected mortality related to each comorbidity (or symptom) by sex and age group out of 100,000 cases. The coloured labels of each graphic follow a decreasing order, considering the average expected mortality over the 23 months. [Colour figure can be viewed at wileyonlinelibrary.com]



	M-[0,20)	M-[20,40)	M-[40,60)	M-[60,)	F-[0,20)	F-[20,40)	F-[40,60)	F-[60,)
Headache-s	59.2	71.9	67.6	63	63	75.5	71.4	64.8
Cough-s	64.8	72.5	72.5	76.4	65.2	73.8	73.4	75.7
Fever-s	60.9	64.2	59.2	60.6	58	58	53.9	55
Nasal disch.-s	39.9	43.2	40.2	35.1	41.4	46.6	44.4	37.5
Joint pain-s	29.1	44.6	44.9	46	30	44.5	45.6	44.2
Muscle pain-s	35	51.7	50.5	50.9	36.2	51.5	51.4	48.8
Ears pain-s	41.6	52.1	48.2	42	43.7	53.7	51.4	42.4
Diarrhea-s	10.7	13.1	12	13.9	10.6	12.5	12.1	14.4
Irritability-s	12.4	9.6	11	13.9	12.6	11	11.7	13.1
Chest pain-s	8.2	13.4	16.8	25.2	9.6	15.1	17.5	23.5
Breath. diff-s	7.1	10.2	18.2	42.5	8.2	10.6	15.9	36
Sud. onset of sympt-s	24.2	25.8	27.3	30.9	24.5	26.1	26.8	29.1
Chills-s	20.7	28.9	28	28.8	21.9	29.3	28.9	27.5
Att. gen. state-s	23.2	29.6	31.8	42.4	23.3	29.1	30.6	39
Conjunctivitis-s	7.5	7.9	7.2	6.5	7.7	8.5	7.9	6.3
Abdominal pain-s	6.7	6.7	6.7	8.7	7.6	7.6	7.5	8.8
Vomiting-s	5.1	3.7	3.6	5	5.7	4.9	4.6	6.3
Incr. resp. freq. depth-s	2.3	2.9	5.7	14.2	2.4	2.9	4.7	12
Blue col. lack oxygen-s	0.8	1	1.7	3.8	0.9	1	1.5	3.2

FIGURE 6 Expected probability of suffering each one of 19 symptoms by sex and age group, for all those were diagnosed as COVID-19 positive case during August 2021. A heat map along with a hierarchical clustering is also included. [Colour figure can be viewed at wileyonlinelibrary.com]

heart disease, diabetes and hypertension. The only symptom that is included in this group is breathing difficulty. Finally, the age group of all those older than 60 years is also included. Thus, the presence of any of these comorbidities or being an elderly exacerbates significantly the symptom that leads to higher mortality. Instead the symptoms ears pain, nasal discharge and belonging to the age groups $[0, 20)$ or $[20, 40)$ have a negative correlation to breathing difficulty.

6.5 | Prediction of death outcomes

In this section, our purpose is to use the footprints from August 2021 to predict the deaths in September 2021 and then to use September 2021 footprints to predict the mortality in October 2021. We compare our prediction with the one obtained via the classic logistic regression model where instead of using the age group variables, the actual age of each patient is used as a continuous variable.

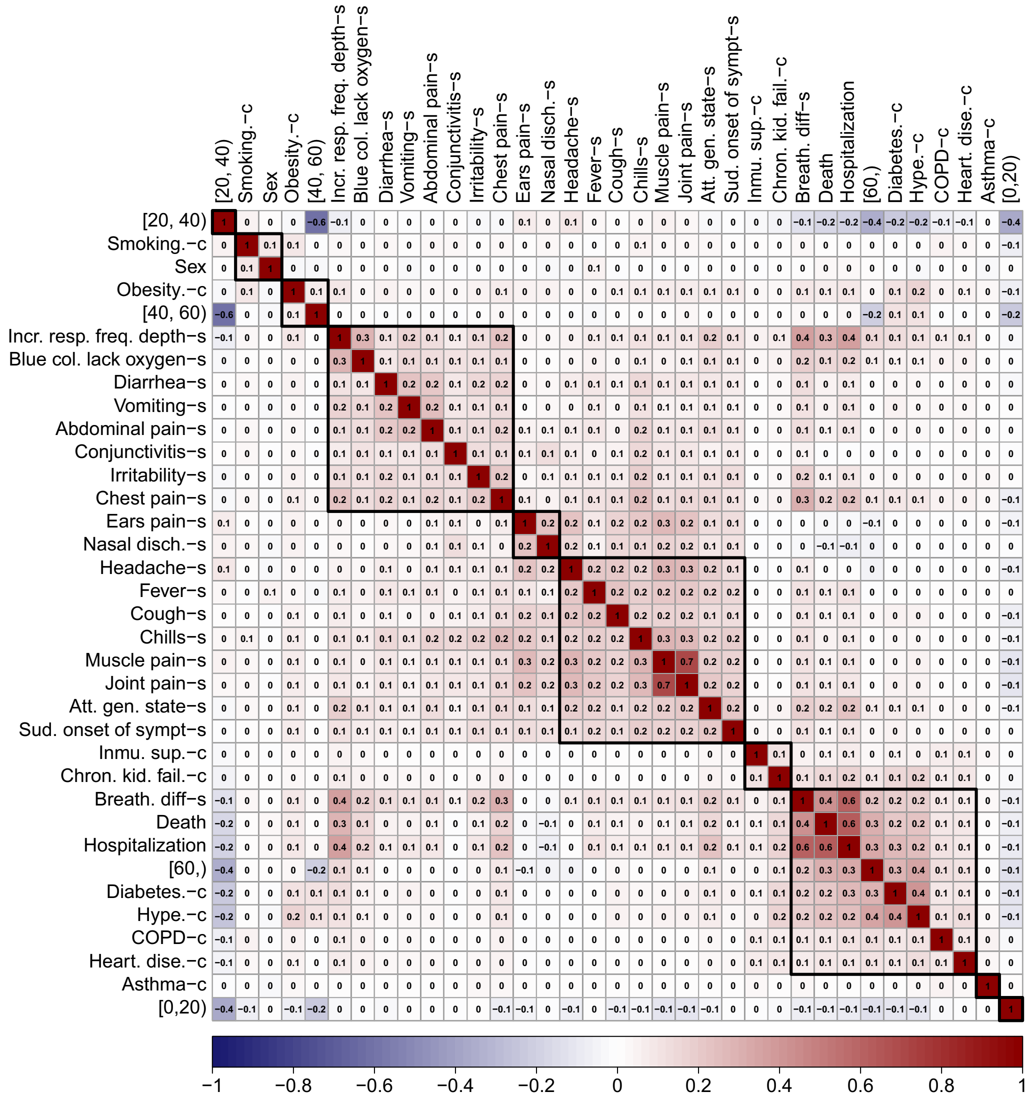


FIGURE 7 Correlation plot of comorbidities, symptoms, sex, age groups, hospitalisation and death. This plot is drawn from all those who were diagnosed as COVID-19 positive case during August 2021. [Colour figure can be viewed at wileyonlinelibrary.com]

To predict COVID-19 death outcomes using the ideas outlined in Section 5, only seven variables are used: COPD, heart disease, diabetes, hypertension, breathing difficulty, sex and the age group [60,). These variables have been included as the correlation plot in Figure 7 indicates that these are the variables with higher correlation with the death outcomes during August 2021. Hospitalisation is not included because for a newly detected COVID-19 case hospitalisation and death outcomes would be unknown. The correlation plot of September 2021 shows that the same variables are correlated with death outcomes as in August 2021. The same model is fitted via logistic regression, and as mentioned before, the actual age of each patient is used as a continuous variable instead of the age group [60,). It could seem that relationships of higher order are not been considered, however we ran this parsimonious model against the complete model with 34 variables and obtained similar results (not shown).

6.5.1 | Optimal cut-point

To determine the cut-point c and make a decision via the estimated mortality risk (obtained via the multivariate Bernoulli and logistic regression) we maximise the proportion of death outputs that are correctly identified (true positive rate or TPR) plus the proportion of alive outcomes that are correctly identified (true negative rate or TNR). Thus, the optimal cut-point maximises the quantity TPR + TNR. The library of R `cutpointr` (see Thiele, 2020) is used to obtained the optimal cut-point. This optimisation exercise is performed using the information from August 2021 to predict the deaths of the same month. The optimal cut-point for the multivariate Bernoulli and logistic models were $\hat{c} = 0.043$ and $\hat{c} = 0.035$ respectively. With these optimal cut-points we use the footprints from August 2021 to predict the deaths of September 2021. The resulting confusion matrices are displayed in Tables 1 and 2.

Via the multivariate Bernoulli we obtain (TPR, TNR)= (85.3%, 93.3%), while using the logistic regression (TPR, TNR)= (86.5%, 93.3%) is achieved. An alternative measure of comparison is the area under the ROC curve (AUC). For the multivariate Bernoulli, it was found to be equal to 0.942, while for the logistic regression was equal to 0.95. In general terms classifiers with larger AUC are considered better to discriminate between groups, hence according to this criterion there is virtually no difference between the discrimination ability of both models. Repeating the exercise but now predicting the death outcomes of October 2021 using the information of September 2021 the confusion matrices shown in Tables 3 and 4 are obtained. Results are very similar to those described above.

TABLE 1 Multivariate Bernoulli: prediction of death outcomes related to the COVID-19 disease of September 2021 via the footprints of August 2021. Confusion matrix (left) and column percentages

		Actual				Actual	
	Outcomes	Alive	Death		Outcomes	Alive	Death
Predicted	Alive	191,822	720	Predicted	Alive	85.3%	6.7%
	Death	33,084	10,018		Death	14.7%	93.3%

TABLE 2 Logistic regression: prediction of death outcomes related to the COVID-19 disease of September 2021 via the footprints of August 2021. Confusion matrix (left) and column percentages

		Actual				Actual	
	Outcomes	Alive	Death		Outcomes	Alive	Death
Predicted	Alive	194,541	718	Predicted	Alive	86.5%	6.7%
	Death	30,365	10,020		Death	13.5%	93.3%

TABLE 3 Multivariate Bernoulli: prediction of death outcomes related to the COVID-19 disease of October 2021 via the footprints of September 2021. Confusion matrix (left) and column percentages

		Actual				Actual	
	Outcomes	Alive	Death		Outcomes	Alive	Death
Predicted	Alive	88,474	314	Predicted	Alive	86%	7.5%
	Death	14,429	3874		Death	14%	92.5%

TABLE 4 Logistic regression: prediction of death outcomes related to the COVID-19 disease of October 2021 via the footprints of September 2021. Confusion matrix (left) and column percentages

		Actual				Actual	
	Outcomes	**Alive**	**Death**		**Outcomes**	**Alive**	**Death**
Predicted	Alive	90,686	385	Predicted	Alive	88.1%	9.2%
	Death	12,217	3803		Death	11.9%	90.8%

Thus, classification-wise the results are comparable, and it is worth noting that there is no difference in performance between a model that uses only the the dichotomous variable $[60,)$ with a model that instead uses the actual age of the patients as a continuous variable.

7 | DISCUSSION

We have used a multivariate Bernoulli distribution to analyse monthly information about the COVID-19 pandemic in Mexico. In particular we have used the data of patients identified as COVID-19 positive and determined, among other interesting findings, that: (1) the mortality risk on age groups $[40, 60)$ and $[60,)$ exhibits a decreasing trend; (2) the expected mortality trend decreases on all age groups; (3) the symptom breathing difficulty is the one associated with the highest mortality and it is exacerbated by the comorbidities COPD, heart disease, diabetes, hypertension and also in patients of 60 years old and above; (4) the symptom breathing difficulty has a negative correlation with the symptoms ears pain and nasal discharge. Belonging to the age groups $[0, 20)$ or $[20, 40)$ also shows a negative correlation to breathing difficulty; (5) the symptoms diarrhoea, irritability, chest pain and breathing difficulty often appear together; (6) using only four comorbidities (COPD, heart disease, diabetes and hypertension), one symptom (breathing difficulty), the sex of the patient and information about whether or not the patient belongs to the group of 60 and older, it is possible to predict the mortality of the COVID-19 disease with the following accuracy: 93.3% of death outputs have been correctly classified and 85.3% of alive outputs have been correctly identified. Finally, when comparing with a logistic regression, it can be seen that with our proposal one could obtain similar predictive performance with fewer variables which are closely correlated to the death outcomes.

The Mexican government's vaccination strategy began in December 2020. Health personnel were the first to be vaccinated. Population aged 60 years and over began to be vaccinated during February 2021. This explains the steep slope during March (Figures 3 and 5). Indeed we believe the decreasing trends are due to the increasing knowledge about the COVID-19 disease, which clearly has translated into improved treatments as well as vaccines.

Finally, the multivariate Bernoulli has 2^k parameters, that is one parameter value per each possible footprint. Indeed, much in the same spirit as for the univariate Bernoulli distribution, the multivariate version constitutes a dense model in its support. Namely all possible combinations are considered. The finite mixture representation (3) suggests the possibility of using a discrete random probability measure as a potential model, namely a Bayesian nonparametric distribution. Such an approach, for example, via the Dirichlet processes, would work under the assumption of an infinite number of footprints and a diffuse baseline distribution to ease its implementation. While potentially possible, some adaptations and interpretations would be needed, so we prefer to keep it simple here.

ACKNOWLEDGEMENTS

The authors are grateful for the detailed comments and suggestions from two reviewers and the associate editor which helped to improve the quality of the manuscript. This work was supported by UNAM-PAPIIT projects IA103220, IV100220 and IG100221.

DATA AVAILABILITY STATEMENT

The data that supports the findings of this study are available in Data S1.

ORCID

Carlos E. Rodríguez https://orcid.org/0000-0002-4796-0784
Ramsés H. Mena https://orcid.org/0000-0001-9608-8059

REFERENCES

Arons, M.M., Hatfield, K.M., Reddy, S.C., Kimball, A., James, A., Jacobs, J.R. et al. (2020) Presymptomatic SARS-CoV-2 infections and transmission in a skilled nursing facility. *New England Journal of Medicine*, 382(22), 2081–2090.

Bonanad, C., García-Blas, S., Tarazona-Santabalbina, F., Sanchis, J., Bertomeu-González, V., Fácila, L. et al. (2020) The effect of age on mortality in patients with COVID-19: a meta-analysis with 611,583 subjects. *Journal of the American Medical Directors Association*, 21(7), 915–918.

Campos, I., Hernández, L., Flores, A., Gómez, E. & Barquera, S. (2019) Prevalencia, diagnóstico y control de hipertensión arterial en adultos mexicanos en condición de vulnerabilidad. Resultados de la ENSANUT 100k. *Salud Pública de México*, 61(6), 888–897.

Dai, B., Ding, S. & Wahba, G. (2013) Multivariate Bernoulli distribution. *Bernoulli*, 19(4), 1465–1483.

Frühwirth-Schnatter, S., Celeux, G. & Robert, C.P. (Eds.). (2019) *Handbook of mixture analysis. Handbooks of Modern Statistical Methods*. Boca Raton: Chapman & Hall/CRC Press.

Frühwirth-Schnatter, S. & Malsiner-Walli, G. (2019) From here to infinity: sparse finite versus Dirichlet process mixtures in model-based clustering. *Advances in Data Analysis and Classification*, 13, 33–64.

Gebhard, C., Regitz-Zagrosek, V., Neuhauser, H.K., R.M. & L.K.S. (2020) Impact of sex and gender on COVID-19 outcomes in Europe. *Biology of Sex Differences*, 11(1), 11–29.

Good, I.J. (1965) *The estimation of probabilities: an essay on modern Bayesian methods*. MIT: *Press*.

Ishwaran, H. & Zarepour, M. (2020) Dirichlet prior sieves in finite normal mixtures. *Statistica Sinica*, 12(3), 941–963.

Lauer, S.A., Grantz, K.H., Bi, Q., Jones, F.K., Zheng, Q., Meredith, H.R. et al. (2020) The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: estimation and application. *Annals of Internal Medicine*, 172(9), 577–582.

Loza, A., Pineda, L., Dyer, D., Benítez, H., Ciria, R., Cruz, R. et al. (2020) Sistema de información nacional depurado sobre la evolución de la pandemia COVID-19. *Biotecnología en Movimiento*, 5(23), 8–11.

Oliveira, B.A., Oliveira, L.C.d., Sabino, E.C. & Okay, T.S. (2020) SARS-CoV-2 and the COVID-19 disease: a mini review on diagnostic methods. *Revista do Instituto de Medicina Tropical de Sao Paulo*, 62, 1–8.

R Core Team (2021) *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Available from: https://www.R-project.org/

Sanyaolu, A., Okorie, C., Marinkovic, A., Patidar, R., Younis, K., Desai, P. et al. (2020) Comorbidity and its impact on patients with COVID-19. *SN Comprehensive Clinical Medicine*, 2(8), 1069–1076.

Singh, A.K., Gupta, R., Ghosh, A. & Misra, A. (2020) Diabetes in COVID-19: prevalence, pathophysiology, prognosis and practical considerations. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, 14(4), 303–310.

Struyf, T., Deeks, J.J., Dinnes, J., Takwoingi, Y., Davenport, C., Leeflang, M.M.G. et al. (2020) Signs and symptoms to determine if a patient presenting in primary care or hospital outpatient settings has COVID-19 disease. *Cochrane Database of Systematic Reviews*, 7(7), 1–342.

Thiele, C. (2020) *cutpointr: determine and evaluate optimal cutpoints in binary classification tasks. R package version 1.0.32*. Available from: https://CRAN.R-project.org/package=cutpointr

WHO. (2021) *Coronavirus disease (COVID-19) pandemic*. Available from: https://covid19.who.int/ [Accessed 15th November 2021].

Wickham, H. (2021) *tidyr: tidy messy data*. R package version 1.1.3. Available from: https://CRAN.R-project.org/package=tidyr

Zhao, Q., Meng, M., Kumar, R., Wu, Y., Huang, J., Lian, N. et al. (2020) The impact of COPD and smoking history on the severity of COVID-19: a systemic review and meta-analysis. *Journal of Medical Virology*, 92(10), 1915–1921.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Rodríguez, C.E. & Mena, R.H. (2022) COVID-19 clinical footprint to infer about mortality. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 185(Suppl. 2), S547–S572. Available from: https://doi.org/10.1111/rssa.12947

APPENDIX A. DISTRIBUTIONAL RESULTS

Let

$$\theta_{j,-s} = \sum_{l=1}^{\mathbb{S}} \mathbb{1}(\rho_{l,j} = 1)\mathbb{1}(\rho_{l,s} \neq 1)\, w(\rho_{l,\cdot}),$$

$$\eta_{j,-s} = \sum_{l=1}^{\mathbb{S}} \left(\mathbb{1}(\rho_{l,j} = 1)\mathbb{1}(\rho_{l,s} \neq 1)\right)(r_l + \delta_l),$$

with equivalent definitions for $\theta_{s,-j}$ and $\eta_{s,-j}$. First, it is easy to see that

$$\begin{aligned}\theta_{j,s} + \theta_{j,-s} &= \sum_{l=1}^{\mathbb{S}} (\mathbb{1}(\rho_{l,j} = 1)\mathbb{1}(\rho_{l,s} = 1) + \mathbb{1}(\rho_{l,j} = 1)\mathbb{1}(\rho_{l,s} \neq 1))\, w(\rho_{l,\cdot}),\\ &= \sum_{l=1}^{\mathbb{S}} \left(\mathbb{1}(\rho_{l,s} = 1) + \mathbb{1}(\rho_{l,s} \neq 1)\right)\mathbb{1}(\rho_{l,j} = 1) w(\rho_{l,\cdot}),\\ &= \sum_{l=1}^{\mathbb{S}} \mathbb{1}(\rho_{l,j} = 1) w(\rho_{l,\cdot}) = \theta_j,\end{aligned}$$

and in the same manner $\eta_{j,s} + \eta_{j,-s} = \eta_j$.

A.1 Distribution of $\Theta_{j,s}/\Theta_j$

We have

$$p(\Theta_{j,s}, \Theta_{j,-s} | \eta_{j,s}, \eta_{j,-s}, \alpha_0) = \mathrm{Dir}(\theta_{j,s}, \theta_{j,-s} | \eta_{j,s}, \eta_{j,-s}, \alpha_0 - \eta_{j,s} - \eta_{j,-s}).$$

Performing the transformation $U = \Theta_{j,s}$ and $V = \Theta_{j,s} + \Theta_{j,-s}$ (observe that $V = \Theta_j$) the Jacobian is equal to 1 and the density is given by

$$\begin{aligned} p(u, v) &= p_{\Theta_{j,s},\Theta_{j,-s}}(u, v - u|\eta_{j,s}, \eta_{j,-s}, \alpha_0), \\ &\propto u^{\eta_{j,s}-1}(v - u)^{\eta_{j,-s}-1}(1 - v)^{\alpha_0-\eta_{j,s}-\eta_{j,-s}}. \end{aligned}$$

Here $u > 0, v > u$ and $1 > v$, thus this is a valid probability distribution function.

Now transforming $X = V$ and $Y = \frac{U}{V}$, then the Jacobian is equal to x and the density is given by

$$\begin{aligned} p(x, y) &= x\ p_{U,V}(xy, x|\eta_{j,s}, \eta_{j,-s}, \alpha_0), \\ &\propto x(xy)^{\eta_{j,s}-1}(x - xy)^{\eta_{j,-s}-1}(1 - x)^{\alpha_0-\eta_{j,s}-\eta_{j,-s}} \\ &\propto x^{\eta_{j,s}+\eta_{j,-s}-1}(1 - x)^{\alpha_0-\eta_{j,s}-\eta_{j,-s}}\ y^{\eta_{j,s}-1}(1 - y)^{\eta_{j,-s}-1}. \end{aligned}$$

Then $X \perp Y$, where $X \sim \text{Beta}(x|\eta_j, \alpha_0 - \eta_j)$ and $Y \sim \text{Beta}(y|\eta_{j,s}, \eta_j - \eta_{j,s})$. Note that $\eta_j = \eta_{j,-s} + \eta_{j,s}$.

Since $Y = \frac{\Theta_{j,s}}{\Theta_j}$, then

$$\begin{aligned} \frac{\Theta_{j,s}}{\Theta_j} &\sim \text{Beta}\left(\frac{\theta_{j,s}}{\theta_j}|\eta_{j,s}, \eta_j - \eta_{j,s}\right), \\ \Rightarrow \mathbb{E}\left(\frac{\Theta_{j,s}}{\Theta_j}\right) &= \frac{\eta_{j,s}}{\eta_j}. \end{aligned}$$

APPENDIX B. POSTERIOR EXPECTATION OF THE CORRELATION

We want to approximate

$$\mathbb{E}\left(\text{Corr}(Y_j, Y_s)|\boldsymbol{y}^{(n)}\right) = \mathbb{E}\left(\frac{\Theta_{j,s} - \Theta_j\Theta_s}{\sqrt{\Theta_j(1 - \Theta_j)}\sqrt{\Theta_s(1 - \Theta_s)}}|\boldsymbol{y}^{(n)}\right).$$

First, note that it is straightforward to generate samples from

$$\begin{aligned} (\Theta_{j,s}, \Theta_{j,-s}, \Theta_{s,-j}) &\sim \text{Dir}(\theta_{j,s}, \theta_{j,-s}, \theta_{s,-j}|\eta_{j,s}, \eta_{j,-s}, \eta_{s,-j}, \alpha_0 - \eta_{j,s} - \eta_{j,-s} - \eta_{s,-j}), \\ &\sim \text{Dir}(\theta_{j,s}, \theta_{j,-s}, \theta_{s,-j}|\eta_{j,s}, \eta_j - \eta_{j,s}, \eta_s - \eta_{j,s}, \alpha_0 + \eta_{j,s} - \eta_j - \eta_s), \end{aligned}$$

and with each sample we can compute $\theta_j = \theta_{j,s} + \theta_{j,-s}$, $\theta_s = \theta_{j,s} + \theta_{s,-j}$ and then

$$\xi = g(\theta_{j,s}, \theta_j, \theta_s) = \frac{\theta_{j,s} - \theta_j\theta_s}{\sqrt{\theta_j(1 - \theta_j)}\sqrt{\theta_s(1 - \theta_s)}}.$$

Thus, generating N samples and computing $\xi_1, \ldots, \xi_N$ we can approximate

$$\mathbb{E}\left(\text{Corr}(Y_j, Y_s)|\boldsymbol{y}^{(n)}\right) \approx \frac{1}{N}\sum_{l=1}^{N}\xi_l,$$

and this is a usual Monte Carlo approximation.