Prediction and Analysis of COVID-19 with Probabilistic Graphical Models

L. Enrique Sucar[1] **Jonathan Serrano-Pérez**[1] **Verónica Rodríguez-López**[1] **Rosa Maria Gutierrez-Rios**[2] **Luis A. Pineda**[3]

[1]Computer Science Dept., Instituto Nacional de Astrofísica, Óptica y Electrónica, Puebla, México
[2]Instituto de Biotecnología, Universidad Nacional Autónoma de Mexico, Cuernavaca, Morelos, México
[3]Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de México, Ciudad de México, México

Abstract

Data analysis using machine learning, in particular probabilistic graphical models, can provide valuable information on the COVID-19 pandemic. Recently, the Mexican National COVID-19 Data Base has been made public, including more than 6.5 million cases of persons that have been registered with symptoms of COVID-19, in a single, free, open-access information system. In this work, we report an initial exploration of the Mexican COVID-19 data base with two types of models: (i) Bayesian network classifiers and (ii) causal graphical models. We obtained several Bayesian classifiers for predicting COVID-19 based on all the relevant attributes in the data base. Additionally, the structures obtained provide interesting information regarding the most relevant attributes for predicting COVID. We also applied causal discovery algorithms to uncover causal relationships between COVID and several factors, providing additional insights of this phenomena.

1 INTRODUCTION

The COVID-19 pandemic has disrupted the social and economic state of most countries worldwide. Mexico is particularly vulnerable to the virus due to its high population with diabetes, hypertension, and obesity. It also has insufficient infrastructure and it is difficult to implement isolation policies, as half of the population lives in the informal economy and depends on a daily income.

Data analysis using machine learning can provide valuable information on the pandemic, including alternative strategies for fast diagnosis, determining the most critical sectors of the population or areas in a country, analysing the effectiveness of certain containment measures, etc. However, access to large and comprehensive data has been difficult. In the case of Mexico, recently the data base from the Health Ministry has been made public, including more than 6.5 million cases of persons that have been registered with symptoms of COVID-19.

The Mexican National COVID-19 Data Base incorporates organized and standardized epidemiological and demographic information on the evolution of the COVID-19 pandemic in Mexico; in a single, free, open-access information system. This open-access data set includes references to the patients' sex and age, their place of residence, their symptoms and comorbidities, whether they were tested and the test's result, whether they were hospitalized or died, and relevant dates, among other data.

Probabilistic graphical models have certain advantages over other machine learning techniques. Besides providing, in general, good accuracy for classification or prediction, they can give additional information about the phenomena. In this work we report an initial exploration of the Mexican COVID-19 data base with two types of models: (i) Bayesian network classifiers and (ii) graphical causal models. We developed both types of models as they have different objectives: Bayesian classifiers can predict certain variables of interest, such as if someone has COVID; causal models provide information on the causal relations between different variables which can help to understand the problem and predict the effect of certain interventions.

We considered three different variants of Bayesian network classifiers (BNCs) [Sucar, 2021]: (a) Naive Bayes classifier, (b) Semi-Naive Bayes classifier (SNBC), and (c) Bayesian network-augmented (BAN) classifier. Using the three variants we learn classifiers for predicting COVID-19 and MORTALITY based on all the relevant attributes in the data base. Additionally, the structures obtained with the SNBC and BAN, provide some interesting information regarding the most relevant attributes for predicting COVID/MORTALITY, and the dependency relations between the variables.

Graphical causal models [Pearl, 2009] go beyond Bayesian

Accepted for the 37th Conference on Uncertainty in Artificial Intelligence (UAI 2021).

networks by representing *causal relations* between variables, so they can be used to predict the results of interventions and provide explanations. We used a causal discovery algorithm, Greedy and Fast Causal Inference (GFCI) [Ogarrio et al., 2016] (described in Section 4), to uncover causal relationships between COVID, MORTALITY and several factors providing additional insights of this phenomena.

In terms of classification, the best results were obtained with the SNBC, obtaining 65% accuracy for predicting COVID-19, and 95% for predicting mortality. The final structures obtained with the SNBC show the most relevant attributes used for the predictions; and surprisingly it uses only three variables to predict mortality with excellent results.

The causal structure obtained with the causal discovery algorithm provides some interesting relations. In particular, it confirms the strong relation between *Fever* and COVID; as well as some not obvious relations, such as COVID – *Anosmia*. Besides the strong expected relation between MORTALITY and COVID, it also shows a direct relation between *Hospital Service* and COVID. In what follows, and in particular in the discussion section, a more comprehensive analysis of the results is presented.

The rest of the paper is organized as follows. Section 2 presents an overview of the Mexican COVID-19 data base. The results obtained with Bayesian classifiers are described in Section 3, and those with causal graphical models in Section 4. An analysis of the the main insights derived from the experiments are discussed in Section 5, and we finalize with some conclusions and directions for future work.

2 MEXICAN COVID-19 DATABASE

The Mexican COVID-19 Database (DB) was produced from a combined effort of the Mexican Ministry of Health and the National Autonomous University of Mexico to make available the data collected nationwide about the COVID-19 pandemic. The data is collected by the Respiratory Viral Epidemiological Surveillance System (*Sistema de Vigilancia Epidemiológica de Enfermedades Respiratorias Virales*, SISVER) consisting on 5,186 health units of the three levels of the health care system. The DB contains over 97 useful variables, including the general references of the patient, such as sex, age, place of residence; symptoms and comorbidities; and also data on testing, hospitalization, and deaths. The system includes around 6.5 million individual records to the present date. The data has been subject to a careful curation process, is weekly updated, and can be accessed through a set of queries of general interest, and also downloaded in full for research purposes at `http://covid-19.iimas.unam.mx`.

Mexico has a great diversity of socio-economical and geographic conditions, and the original data is input and collected at the local, state and nation level with great difficulty.

Table 1: Variables selected, grouped by categories, for learning the classifiers and causal models.

Category	Variables
patient-data	GENDER, AGE, CITY, NATIONALITY, PATIENT TYPE, INDIGENOUS, JOB, HOSPITAL SERVICE, CONTACT BIRDS, CONTACT PIGS, CONTACT COVID
symptoms	FEVER, COUGH, ODINOGY, DYSPNOEA, IRRITABILITY, DIARRHEA, CHEST PAIN, CHILL, HEADACHE, MYALGIA, ARTHRALGIA, DISCOMFORT, RHINORRHEA, POLYPNEA, VOMITING, ABDOMINAL PAIN, CONJUNCTIVITIS, CYANOSIS, SUDDEN SYMPTOMS, ANOSMIA, DYSGEUSIA
comorbidities	DIABETES, COPD, ASTHMA, IMMUNOSUPPRESSION, HYPERTENSION, HIV-AIDS, OTHER COMORBIDITIES, ENDOCARDITIS, OBESITY, CHRONIC KIDNEY, SMOKING
diagnosis and treatment	ANALGESIC, ANTIVIRAL, ANTIPYRETICS, VACCINATED
objective class	COVID19, MORTALITY

The curation process has to face strong challenges, such as bias, due to the fact that only people that presented symptoms and attend a health unit are included, and there is no data about asymptomatic people; uncertainty about the reliability of the information due to the contingencies at the original collection points; missing values in the data; data integrity and unaccounted deaths; in particular, there is no record of people that never sought medical assistance in the health care system and died; neither of people that was discharged and died afterwards. Nevertheless, the data is reliable in a great degree and open for scientific research.

2.1 DATA SET PRE-PROCESSING

The data set contains 97 variables, some of them can be considered as attributes while others as classes for classification purposes. First, we identify *COVID19* and *MORTALITY* as the class variables. Then, a *naive* feature selection was carried out; that is, unique identifiers, redundant variables, variables that contain dates and variables with a high proportion of missing values are removed, this results in 47 variables that will be used as attributes. The variables used for learning the classifiers and causal models, clustered by categories, are summarized in Table 1.

The following steps were applied to the dataset: Instances with missing values are removed. The attribute *AGE* was discretized {*1: age lower than 60, 2: age greater or equal than 60*}. Only the instances that take the values {YES, NO} from the following {*MYALGIA, COUGH, HEADACHE, RHINORRHEA, ODINOGY, DIARRHEA, DYSPNOEA, CHEST PAIN, ANOSMIA, FEVER, IRRITABILITY, CHILL, ARTHRALGIA, DISCOMFORT, POLYPNEA, VOMITING, ABDOMINAL PAIN, CONJUNCTIVITIS, CYANOSIS, SUDDEN SYMPTOMS, DYSGEUSIA, DIABETES, COPD, ASTHMA, IMMUNOSUPPRESSION, HYPERTENSION, HIV-AIDS, COMPLICATION, ENDOCARDITIS, OBESITY, CHRONIC KIDNEY, SMOKING*} are considered (other values could be *unknown*, etc.).

Table 2: Description of the datasets for the classification problems. For COVID19: (P) Positive, (N) Negative. For MORTALITY: (A) Alive, (D) Death.

Dataset for:	#Instances	#Atts.	# P/A	# N/D
COVID19	2,963,824	47	1,303,818	1,660,006
MORTALITY	5,482,335	47	5,287,744	194,591

Finally, for the classification problem of COVID-19 (*COVID19* class) only instances associated to {*POSITIVES-COVID-19, NEGATIVE-COVID-19*} are considered, while for the classification problem of mortality (*MORTALITY* class) only the instances with a valid value for the class are considered. A summary of the data set after pre-processing for both classification problems is shown in Table 2. The data set of MORTALITY is highly unbalanced, ~96.45% are associated to *alive* and ~3.55% to *death*.

3 BAYESIAN NETWORK CLASSIFIERS

This section presents briefly the Bayesian classifiers that are trained to predict if a person is infected with COVID-19 and to predict MORTALITY. The implementations of the classifiers are from the toolkit *PGM_PyLib*[1] [Serrano-Pérez and Sucar, 2020]. For a detailed description of the methods see Sucar [2021].

3.1 NAIVE BAYES CLASSIFIER (NBC)

The Naive Bayes Classifier (NBC) is based in the assumption that all the attributes are independent given the class variable. So, each attribute A_i is conditionally independent of all other attributes given the class (C):

$$P(A_i|A_j, C) = P(A_i|C), \forall j \neq i \tag{1}$$

In this way, the probability of each class given the attributes can be written as:

$$P(C|\mathbf{A}) = P(C)P(A_1|C)P(A_2|C)...P(A_n|C)/P(\mathbf{A}) \tag{2}$$

Where $\mathbf{A}$ is the short representation of $A_1, A_2, ..., A_n$ with n attributes, and $P(\mathbf{A})$ can be considered as a normalization constant.

Therefore, the classification problem, based on equation 2, can be formulated as:

$$Arg_C Max[P(C|\mathbf{A}) = P(C)\ P(A_1|C)\ P(A_2|C)...\ P(A_n|C)/P(\mathbf{A})] \tag{3}$$

That is, the class C that maximizes 3 will be returned as the prediction for a new instance.

[1] Available at `https://github.com/jona2510/PGM_PyLib`

3.2 BAYESIAN NETWORK AUGMENTED BAYESIAN CLASSIFIER (BAN)

While the NBC assumes that all attributes are independent given the class, there are models that incorporate dependencies between the attributes. Bayesian network augmented Bayesian classifiers (BANs) include a dependency structure among attributes that can be any directed acyclic graph.

The posterior probability for the class given the attributes, considering the conditional probability of each attribute given the class and its parent attributes, is:

$$P(C|\mathbf{A}) = P(C)\ P(A_1|Pa(A_1), C)\ P(A_2|Pa(A_2), C)...\ P(A_n|Pa(A_n), C)/P(\mathbf{A}) \tag{4}$$

Where $\mathbf{A}$ is the short representation of $A_1, A_2, ..., A_n$ with n attributes, $P(\mathbf{A})$ can be considered as a normalization constant and $Pa(A_i)$ is the set of parent attributes of A_i.

In this way, the classification problem based on equation 4 can be formulated as:

$$Arg_C Max[P(C|\mathbf{A}) = P(C)\ P(A_1|Pa(A_1), C)\ P(A_2|Pa(A_2), C)...P(A_n|Pa(A_n), C)/P(\mathbf{A})] \tag{5}$$

That is, the class C that maximizes 5 will be returned as the prediction for a new instance.

3.2.1 Semi-Naive Bayesian Classifiers (SNBC)

The main idea of the Semi-Naive Bayesian Classifier is to eliminate attributes with *low* mutual information with the class, and eliminate or *join* attributes which are not independent given the class in order to improve the performance of the Naive Bayes classifier, *accuracy* is the evaluation measure to be improved (see section 3.3.1 for more details). Once the previous structure modifications are performed, a NBC is trained for predicting the class of new instances.

3.3 EXPERIMENTAL CONFIGURATION

The training procedure for the classifiers is described below:

• **BAN**: The Chow-Liu algorithm [Chow and Liu, 1968] was used to get the dependency structure between attributes given the class variable.
• **SNBC**: *validation* was set to 0.8, that is, 80% of the training set is used to train a local classifier and the other 20% is used to evaluate the performance of the classifier. *Epsilon* and *omega* were set to 0.005 and 0.02 respectively, that is, attributes with mutual information lower than *epsilon* are removed; and two attributes are combined or one attribute is removed if their conditional mutual information given the class is greater than *omega*.

Furthermore, *smooth* was set to 0.1 in order to avoid zero probabilities and the *prior* probabilities are used in the prediction phase for the three classifiers. Finally, 5-folds cross

validation was performed, so the results are the averages of the 5 folds.

3.3.1 Evaluation Measures

Different evaluation measures are used to evaluate the performance of the classifiers. Let TP, TN, FP and FN be true positives, true negatives, false positives and false negatives, respectively. The evaluation measures are described below:

- **Accuracy**: also know as *exact-match*. Return the percentage of instances correcly predicted.
- **Precision** = $\frac{TP}{TP+FP}$
- **Recall** = $\frac{TP}{TP+FN}$
- **F1-score** = $2 * \frac{precision*recall}{precision+recall}$

The ***macro average*** of precision, recall and F1-score are reported in the section of results, that is, the average of the individual evaluation of each value in the class.

3.4 RESULTS

This section presents the results for both classification problems. Table 3 summarizes the results for COVID-19 and Table 4 for MORTALITY. Furthermore, because the MORTALITY dataset is highly unbalanced, the SNBC internal evaluation measure was replaced by the *F1-score* in order to avoid a classifier that always predicts the majority class.

Figure 1 shows the dependency structure for the SNBC for predicting COVID, including the attributes selected in the optimization process. At the end, eight attributes remained, although one is the combination of two variables: *Patient type* and *Hospital Service*[2]. Figure 3 depicts the dependency structure obtained between the attributes for the BAN classifier. In this case, the class variable, COVID, is connected to all the attributes (not shown in the graph for clarity).

The resulting structure obtained with the SNBC for predicting MORTALITY is shown in Figure 2. Only three variables remain to predict MORTALITY. Figure 4 shows the structure obtained by the BAN for predicting MORTALITY; here also the class is connected to all the attributes.

4 GRAPHICAL CAUSAL MODELS

This section presents the graphical causal models for the Mexican COVID19 data. First, we introduce the graphical causal models and the algorithms for their learning. Next, we describe the graphical causal models discovered.

[2]These two attributes are combined into a single variable given they are not conditionally independent given the class, as a result of the SNBC learning algorithm, see Sucar [2021], Chapter 4. *Patient type* indicates if the patient was hospitalized or not, and *Hospital Service* in which type of hospital.

Table 3: Results for classification problem of COVID-19, standard deviation in parentheses. In bold the best scores for each measure.

	NBC	BAN	SNBC
Accuracy	0.6396 (0.01)	0.6396 (0.01)	**0.6467 (0.015)**
Precision	0.6338 (0.01)	0.6338 (0.01)	**0.6438 (0.014)**
Recall	0.632 (0.009)	**0.6321 (0.009)**	0.6293 (0.02)
F1-score	0.6323 (0.01)	**0.6323 (0.01)**	0.6267 (0.025)

Table 4: Results for classification problem of MORTALITY, standard deviation in parentheses. In bold the best scores for each measure.

	NBC	BAN	SNBC
Accuracy	0.9435 (0.017)	0.9435 (0.017)	**0.9549 (0.011)**
Precision	0.6923 (0.034)	0.6922 (0.034)	**0.7145 (0.034)**
Recall	**0.9176 (0.008)**	0.9174 (0.008)	0.8546 (0.032)
F1-score	0.7536 (0.038)	0.7535 (0.038)	**0.761 (0.028)**

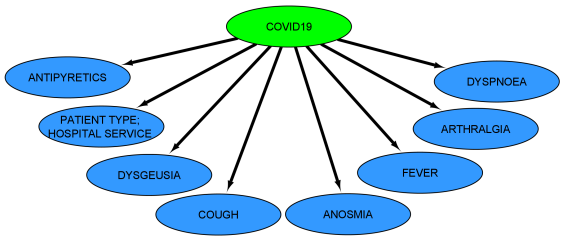

Figure 1: Dependency structure generated by the SNBC classifier for predicting COVID19. In green the class node, COVID19.

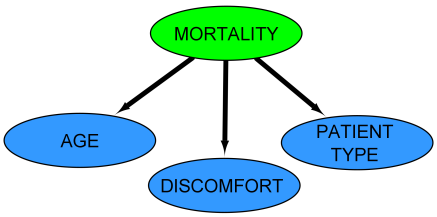

Figure 2: Dependency structure generated in the SNBC classifier for predicting MORTALITY. In green the class node, MORTALITY.

4.1 OVERVIEW OF CAUSAL DISCOVERY TECHNIQUES

A graphical model has a causal interpretation when its structure contains a directed edge $X \rightarrow Y$ if there is an intervention that fixes X to a specific value and changes the probability distribution of Y [Spirtes et al., 2000].

For inferring structures of graphical causal models, causal discovery methods analyze observational data and rely upon some of the following assumptions: i) there is no common

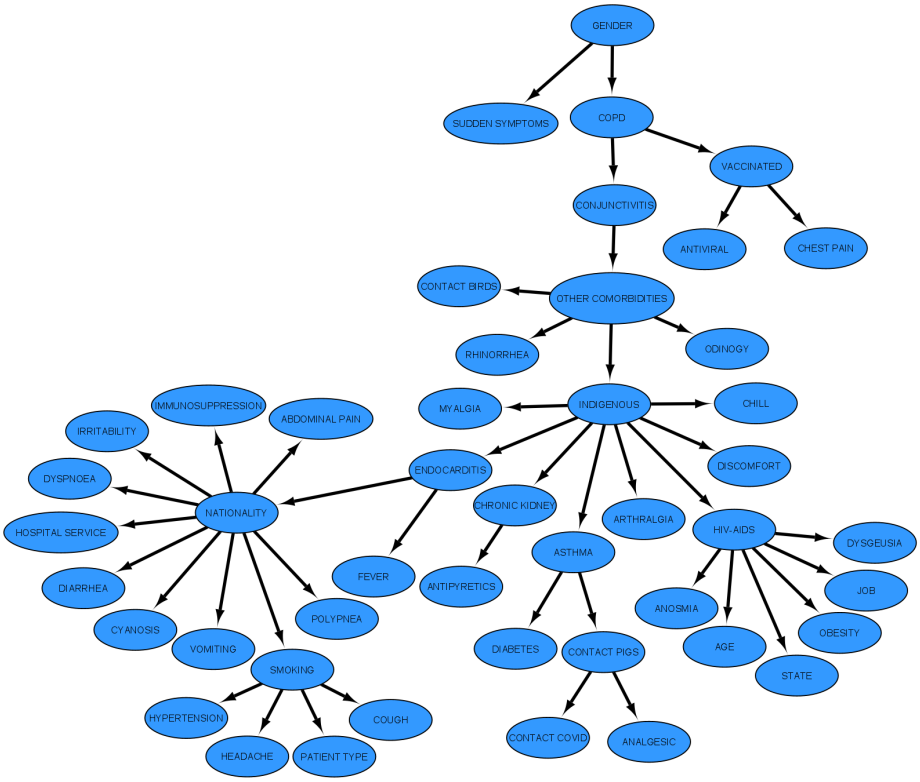

Figure 3: Dependency structure generated by the BAN classifier for COVID19. The node COVID19 is not shown in the graph for clarity, but all nodes are descendants of it.

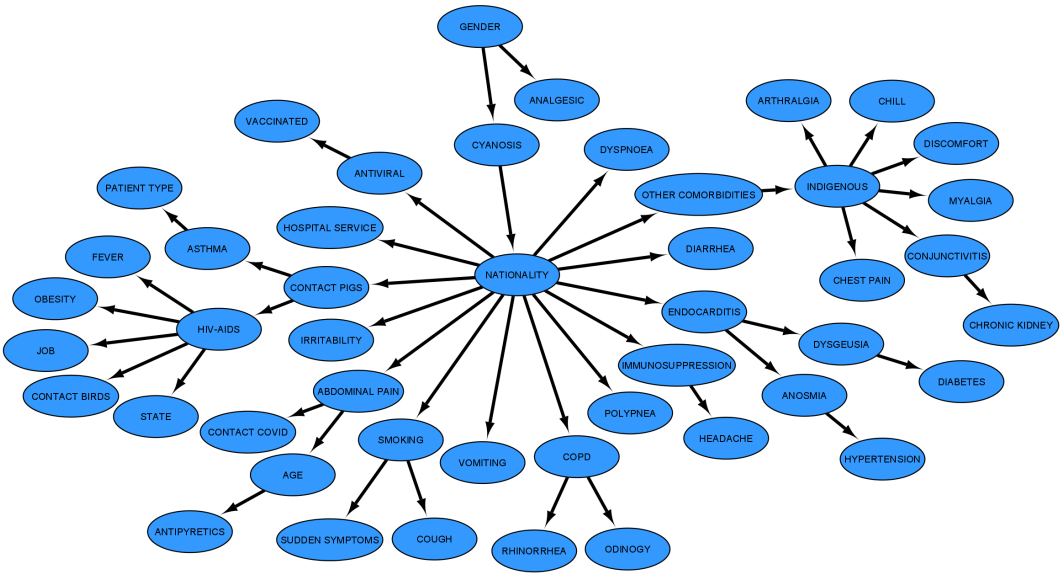

Figure 4: Dependency structure generated by the BAN classifier for MORTALITY. The node MORTALITY is not shown in the graph for clarity, but all nodes are descendants of it.

unobserved direct cause of two observed variables (Causal Sufficiency), ii) every conditional independence entailed in the causal structure is also in its associated probability distribution (Causal Markov Assumption) and, iii) every conditional independence entailed in the probability distribution of the graphical model is also in its structure (Causal Faithfulness Assumption).

Fast Causal Inference (FCI) [Spirtes et al., 1999] is a constrained-based causal discovery method that performs conditional independence tests for recovering causal structures. FCI does not assume causal sufficiency. It searches for causal structures with unobserved variables under the causal Markov and faithfulness assumptions. Starting with a fully connected undirected graph and increasing the size of conditional variables in each iteration, FCI applies conditional independence tests for deciding which edges to delete. It stops when there are no subsets of variables in which to perform independence tests. Then it identifies if the statistical relation between two variables is due to an unobserved variable. FCI outputs a partial ancestral graph (PAG) that includes the following types of edges: 1) $X \rightarrow Y$ when X is a cause of Y, and Y is not a cause of X. 2) $X \leftrightarrow Y$ when X is not a cause of Y, and Y is not a cause of X. There is an unobserved variable causing X and Y. 3) $X \circ\!\!\rightarrow Y$ when X is a cause of Y (Y is not a cause of X), or there is an unobserved variable causing X and Y. 4) $X \circ\!\!-\!\!\circ Y$ when (exactly one of the following holds): i) X is a cause of Y, ii) Y is a cause of X, iii) there is an unobserved variable causing X and Y, or iv) both (i) and (iii), or v) both (ii) and (iii).

Fast Greedy Equivalence Search (FGES) [Ramsey, 2015] is a modification of the scored-based method Greedy Equivalence Search (GES) [Chickering, 2002]. In the same form as GES, under causal sufficiency, causal Markov, and faithfulness conditions, FGES optimizes some operations for discovering causal structures of high dimensionality. FGES uses a score function for evaluating potential causal structures and returns that structure with the highest score. First, starting with an empty graph, FGES adds edges until the score function reaches a local maximum. After that, it deletes those edges that also could improve the score function. Finally, FGES returns a partial directed acyclic graph, called Markov equivalence class (MEC), representing a set of equivalent graphical causal models with the same probability distribution. These MECs include a directed edge if it appears in all equivalent causal structures, and an undirected edge if it appears in some of them.

The Greedy and Fast Causal Inference (GFCI) [Ogarrio et al., 2016] combines FGES and FCI. First, GFCI applies FGES to find a Markov equivalence class (MEC) that is undirected in the next step. Then, it uses FCI to remove false edges and correct the orientation of those edges in the output of FGES.

4.2 RESULTS

With the help of the *Causal-Cmd* tool version 1.2.2 [Spirtes et al., 1990] of the Center for Causal Discovery, we discovered the causal relations between the variables in the Mexican COVID19 database. For our experiments, we considered the variables in Table 1 and the instances for *COVID19* and *MORTALITY* described in Section 2.1. We applied the GFCI method using the Bayesian Dirichlet equivalent and uniform score (BDeu Score) and the Chi-square test, with an $alpha = 0.01$, prior equivalent sample size = 10, and maximum degree = 4.

In Figures 5 and 6, we present the causal relations discovered by GFCI[3]. Figure 5 depicts mainly the causal relations among symptoms, *COVID19*, and *MORTALITY*. We found that *COVID19* is a direct cause of *MORTALITY* and that *FEVER* is the only direct cause of *COVID19*. Some of the other symptoms form causal paths with *COVID19*, for example, $CHEST\ PAIN \rightarrow CHILL \rightarrow ODINOGY \rightarrow HEADACHE \rightarrow COUGH \rightarrow FEVER$. We also found that *COVID19* has effects over the use of antipyretic, analgesic, and antiviral treatments.

In Figure 6, we present the causal relations among *COVID19* and *MORTALITY* with comorbidity and patient- data variables. It can be observed in this figure that any comorbidity variable has direct causal relations with *COVID19*. Some of the comorbidity variables, having symptoms as intermediaries, form causal paths with *COVID19*. For example, $COVID19 \rightarrow ANOSMIA \rightarrow DYSGEUSIA \rightarrow OBESITY \rightarrow DIABETES$. We also found that *AGE*, *GENDER*, *VACCINATED*, have uncertain relations with other variables, and *JOB* and *HOSPITAL SERVICE* are direct and indirect causes of *MORTALITY*.

The causal discovery algorithms have limitations, such as not always determining the directions of the causal links, and can be affected by unobserved co-factors; so the results should be take with caution.

5 DISCUSSION

The causal relationships observed in Figure 5 show results consistent with the clinical pictures observed in patients with $COVID19$ disease (see `https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html`). Where a significant percentage presents fever, dry cough, and tiredness, where the latter is not reported as such in the database, but which could be interpreted as an attack on the general state (*DISCOMFORT*), which does not seem to have a direct causal relationship in the built network. The *CHILLS* derive in two routes associated with the symptoms. The first is

[3] Although a single causal structure is obtained, this is depicted in two parts for clarity.

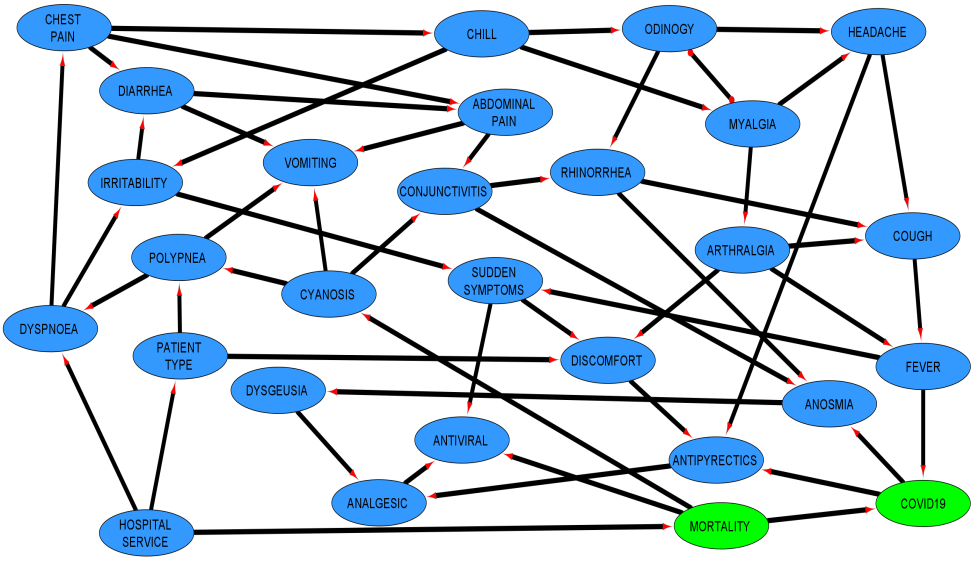

Figure 5: The causal relations between symptoms, treatments, *COVID19*, and *MORTALITY* variables of the Mexican COVID19 database discovered by the GFCI algorithm. The edges in the causal graph have the following meaning: $X \rightarrow Y$ when X is a cause of Y, and Y is not a cause of X; $X \circ\!\!-\!\!\circ Y$ when (exactly one of the following holds): (i) X is a cause of Y, (ii) Y is a cause of X, (iii) there is an unobserved variable causing X and Y, both (i) and (iii), or both (ii) and (iii).

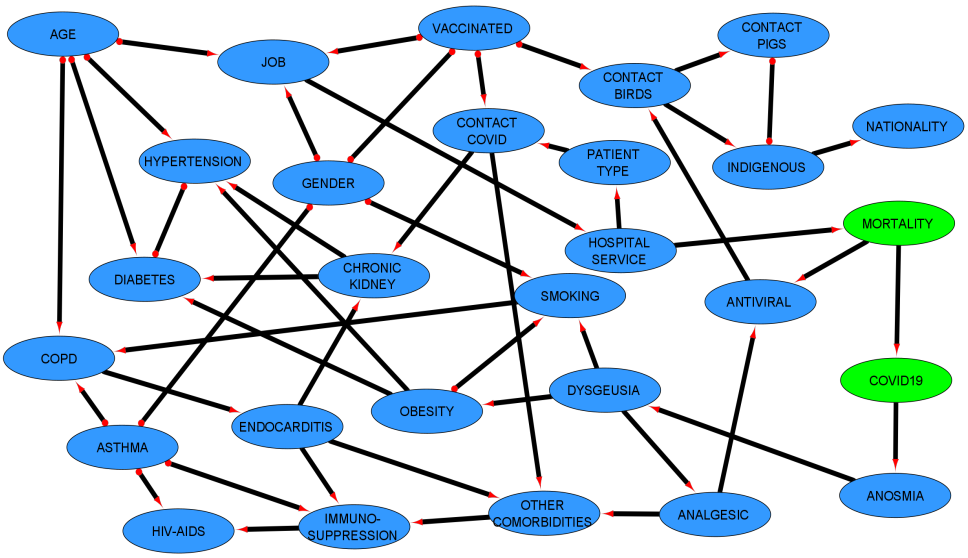

Figure 6: The causal relations between comorbidity, patient-data, *COVID19*, and *MORTALITY* variables of the Mexican COVID19 database discovered by the GFCI algorithm. The edges in the causal graph have the following meaning: $X \rightarrow Y$ when X is a cause of Y, and Y is not a cause of X; $X \circ\!\!\rightarrow Y$ when X is a cause of Y (Y is not a cause of X), or there is an unobserved variable causing X and Y; $X \circ\!\!-\!\!\circ Y$ when (exactly one of the following holds): (i) X is a cause of Y, (ii) Y is a cause of X, (iii) there is an unobserved variable causing X and Y, both (i) and (iii), or both (ii) and (iii).

causally related to *MYALGIA*, *ARTHRALGIA*, and *FEVER* reported as usual symptoms in COVID patients. *MYALGIA* receives two interesting relationships: *ODYNOPHAGIA* and *CHILLS*, where the pathways lead to a *FEVER* that could be identified as general symptomatology commonly associated with the disease. It is important to highlight that given the novelty of the disease, the symptoms have been increasingly associated. This implies that a series of symptoms such as *DYSGEUSIA*, *ANOSMIA*, and *CONJUNCTIVITIS*, among others, were added to the database as health service providers have identified them as probable symptoms. Therefore, it is expected that some causal relationships could be impacted by a lesser amount of information collected over time.

The comorbidities such as diabetes, hypertension, and obesity have been correlated with increased severity of the disease and less satisfactory outcome, including higher mortality; being more so when these comorbidities are combined. Although they are importantly related to each other in the causal network, we can observe that they do not seem to impact mortality or suffer from *COVID19* directly. Furthermore, the results indicate that there is a close relationship between *HYPERTENSION* and *DIABETES*. It is noteworthy that in most cases, both the symptoms and the comorbidities are declared by the patients, so the results, although significant, should be taken with caution.

Regarding the prediction of *COVID19*, a possible improvement of our results could occur if health units receiving *COVID19* patients had access to clinical records, as currently in most of the cases the information is just what the patient reports, so it is not reliable. This could significantly improve predictions, and thereby find the known relationship between *COVID19* and comorbidities described above.

Finally, what seems to be a constant pattern in Figures 5 and 6, is that admission to a hospital has a significant relationship with the mortality of the Mexican population, which opens up interesting questions to investigate. For example, if these increases in mortality took place at the peaks of the epidemic in Mexico, which led to the saturation of care services, especially all intensive care units, which would be a response to what was observed in this study.

6 CONCLUSIONS

The Mexican COVID-19 Data Base, which contains more than 6.5 million records, offers a unique opportunity for modelers interested in understanding epidemiological aspects of the pandemic caused by the SARS-COV2 pathogens that cause the *COVID19* disease. The analyses carried out in this work show that the symptoms present direct causal relationships towards aspects such as mortality and the probability of suffering from *COVID19*. The data collected results from people who attended medical units, so there is an under-registration, which could mask the relationships between some variables, such as comorbidities. Despite this, the results presented in this work show the first effort to understand which epidemiological variables impact the evolution of the pandemic in the Mexican population.

As future work, we plan to apply other causal discovery algorithms, and to update the models as more data becomes available.

References

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3(Nov):507–554, 2002.

C. Chow and C. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14(3):462–467, May 1968. ISSN 0018-9448. doi: 10.1109/TIT.1968.1054142.

Juan Miguel Ogarrio, Peter Spirtes, and Joe Ramsey. A hybrid causal search algorithm for latent variable models. In *Conference on Probabilistic Graphical Models*, pages 368–379. PMLR, 2016.

J. Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, New York, 2009.

Joseph D Ramsey. Scaling up greedy causal search for continuous variables. *arXiv preprint arXiv:1507.07749*, 2015.

Jonathan Serrano-Pérez and L. Enrique Sucar. Pgm_pylib: A toolkit for probabilistic graphical models in python. In Manfred Jaeger and Thomas Dyhre Nielsen, editors, *Proceedings of the 10th International Conference on Probabilistic Graphical Models*, volume 138 of *Proceedings of Machine Learning Research*, pages 625–628. PMLR, 23–25 Sep 2020. URL `http://proceedings.mlr.press/v138/serrano-perez20a.html`.

Peter Spirtes, Richard Scheines, and Clark Glymour. The TETRAD project. In *Acting and Reflecting*, pages 183–207. Springer, 1990. doi: https://doi.org/10.1007/978-94-009-2476-5_13.

Peter Spirtes, Christopher Meek, and Thomas Richardson. An algorithm for causal inference in the presence of latent variables and selection bias. *Computation, causation, and discovery*, 21:1–252, 1999.

Peter Spirtes, Clark N Glymour, Richard Scheines, David Heckerman, Christopher Meek, Gregory Cooper, and Thomas Richardson. *Causation, prediction, and search*. MIT press, 2000.

L. Enrique Sucar. *Probabilistic Graphical Models: Principles and Applications*. Springer Nature, Switzerland, 2nd edition, 2021.