ALFABETO: Supporting COVID-19 hospital admissions with Bayesian Networks

Giovanna **Nicora**[1], Antonio Lo **Tito**[2], Antonella **Donatelli**[2], Giovanni **Callea**[2], Carla **Biasibetti**[2], Maria Vittoria **Galli**[2], Federico **Comotto**[3], Chandra **Bortolotto**[4], Stefano **Perlini**[5,6], Lorenzo **Preda**[2,4] and Riccardo **Bellazzi**[1]

[1]*Dep. of Electrical, Computer and Biomedical Engineering, University of Pavia, Italy*

[2]*Joint Institute for Nuclear Research, 6 Joliot-Curie, Dubna, Moscow region, 141980, Russian Federation*

[2]*Unit of Radiology, Department of Clinical, Surgical, Diagnostic, and Pediatric Sciences, University of Pavia*

[3]*Laife Reply, Milan, Italy*

[4]*Dep. of Radiology, Fondazione IRCCS Policlinico San Matteo, Pavia*

[5]*Dep. of Internal Medicine and Therapeutics, University of Pavia*

[6]*Emergency Department, I.R.C.C.S. Policlinico San Matteo Foundation*

Abstract

The ongoing pandemics of coronavirus disease has accelerated the implementation of machine learning methods (ML) to support clinical decisions. Within this context, we present the ALFABETO project, whose aim is to aid clinicians during COVID-19 patients hospital admission through the application of ML approaches exploiting clinical and chest x-ray features. Yet, non linear ML classifiers are often perceived as not easily interpretable by users, thus hampering trust in ML predictions. Moreover, these ML models, such as Neural Networks or Random Forest, are not able to include pre-exisisting knowledge about a specific domain and are not designed to find causal relationships between variables. For these reasons, we wanted to investigate if Bayesian Networks were able to properly describe the hospital admission decision process. Bayesian Networks are probabilistic graphical models representing a set of variables and their conditional dependencies. The network structure was derived both from existing medical knowledge and from patients data collected during the first wave of the pandemic. While being explainable, we show that the Bayesian network has similar performance when compared to a less explainable ML model and that was able to generalize well across COVID-19 waves.

Keywords
Bayesian networks, COVID-19 hospitalization prediction, Clinical decision support system.

1. Introduction

As of October 2021, the ongoing pandemics of coronavirus disease 2019 (COVID-19) has caused more than 200 million confirmed cases and 4 million deaths [1]. COVID-19 illness severity varies greatly: many patients experience mild or no symptoms, while some need long or short hospitalization. Since the beginning of the pandemics, Artificial Intelligence (AI) approaches have been identified as useful approaches to support clinicians [2]. These tools hold the promise to effectively support different types of decisions, from hospital admission to therapeutic strategies, and research hospitals have worked to integrate them in clinical practice. Nevertheless,

---

building ML tools able to generalize over time and/or on patients coming from different hospitals can be challenging, since ML inherently suffers from dataset shifts and poor generalization ability across different population [3]. As the pandemic evolves, several sources of data shifts are arising, from new variants highly transmissible to new treatment protocols. Additionally, ML classifications of widely used algorithms, from Neural Networks to Gradient Boosting, are often perceived as opaque. Current research on AI Explainability (XAI) aims at making ML predictions more transparent for the user. Towards this direction, different XAI approaches have been developed, many of these providing explanations of single ML predictions by highlighting the important features that lead the classifier to its final decision. Yet, as recently stated, in order to reach explainability in medicine we need to promote causability [4]. However, explanations derived from current XAI methods provide spurious correlations rather than cause/effects relationships, leading to erroneous or even biased explanations [5]. In this context, we present the ALFABETO Sars-CoV2 project (**AL**l **FA**ster **BE**tter **TO**gheter), whose aim is to develop an AI-based pipeline integrating data from diagnostic tools and clinical features to support clinicians during the triage of COVID-19 patients within the Policlinico San Matteo University Hospital, located in Pavia (Italy). The ML-based component will suggest clinicians whether the patient can be treated at home, or he/she needs to be hospitalized. In particular, we have developed a Bayesian Network (BN), a probabilistic graphical model that allows to model the conditional dependencies of a set of variables. As a consequence, BNs are particularly suitable to model pre-existing domain knowledge and the automated reasoning process of human experts [6]. We were therefore able to model existing medical evidence and suggest potential cause/effect relationships between clinical variables and hospital admission. We evaluated whether the BN predictive performances are comparable with those of a widely used but less explainable ML model, e.g. Random Forest. We also tested the generalization ability of the models across different pandemic waves.

2. Materials and Methods

2.1. Datasets

During the first wave of the COVID-19 pandemics in Italy (March 2020-May 2021), we gathered data from 660 COVID-19 patients treated at the IRCCS Policlinico S. Matteo hospital, an excellence center that is known to have successfully treated the fist diagnosed COVID-19 patients in western countries. Half of these patients were hospitalized, while the remaining showed a better prognosis, and were treated at home. For each patient, we collected clinical features, such as age, gender, and evidence of comorbidities. Deep Learning was used to extract features from chest radiographs (RX) images through the X-RAIS platform, developed by Reply^{tm}. X-RAIS is a deep network able to analyze different types of medical images and to extract relevant information for diagnosis. In this context X-RAIS transforms the RX image of a patient into 5 numerical clinically relevant features: Consolidation, Infiltration, Edema, Effusion and Lung Opacity. These 5 features, together with 19 clinical features, will represent the input of a ML model that will predict whether a patient should be hospitalized (class 1) or not (class 0). We randomly selected 90% of the patients as training set. The remaining 10% of patients will be kept for testing and selecting the best performing model. During the third wave (March-May 2021),

462 additional patients experienced the triage. In this case, 68% of patients were hospitalized. The third wave set was exploited as validation set.

2.2. Bayesian Network design and implementation

To implement the BN, we first designed a graph based on our pre-existing knowledge. This graph contains relationships between few variables that may represent the clinicians reasoning process during triage. The graph is represented in Figure 1a: the node label as "Treatment (Home vs Hospital)" is the target node representing our outcome of interest, i.e. whether the patient should be hospitalized or not. To make this decision, we assume that the clinician would evaluate at least the age, the gender (male patients are more likely to incur more severe consequences from the infection) and whether the patient has breathing difficulties. The target node depends on these 3 variables, and we also assume a direct dependency between age and breathing difficulties. We then enrich the structure of this graph with the remaining collected variables, by using the hill climbing search algorithm applied on the training data: starting from the constraints represented in Figure 1a, this method implements a greedy local search and performs single-edge manipulations that maximally increase a score of fitness. The search terminates once a local maximum is found [7]. The resulting graph is shown in Figure 1b: notably, the Boolean feature indicating whether the patient has more than 2 comorbidities ("ComorbiditiesGreaterThan2") is explicitly linked to comorbidities nodes, such as the presence of cancer or cardiovascular diseases. Interestingly, the outcome is not directly linked to the node "ComorbiditiesGreaterThan2", but it can be linked to the presence of comorbidities through patient's age. Some DL features directly depend on the target node. BN is implemented in Python 3.7, using the `bnlearn` package.

Table 1
Predictive performance of the Bayesian Network (BN) and the Random Forest (RF) on the Test set.

	AUC	PRC	Accuracy	Precision	Recall	Specificity	F1 score
BN	0.80	0.85	0.76	0.82	0.78	0.73	0.79
RF	0.76	0.84	0.71	0.78	0.72	0.69	0.75

Table 2
Predictive performance of the Bayesian Network (BN) and the Random Forest (RF) during the third wave (95% Confidence Interval).

	AUC	PRC	Accuracy	Precision	Recall	Specificity	F1 score
BN	[0.76 0.84]	[0.86 0.92]	[0.67 0.75]	[0.86 0.92]	[0.61 0.69]	[0.78 0.85]	[0.71 0.79]
RF	[0.79 0.86]	[0.88 0.94]	[0.71 0.79]	[0.86 0.92]	[0.67 0.75]	[0.78 0.85]	[0.75 0.82]
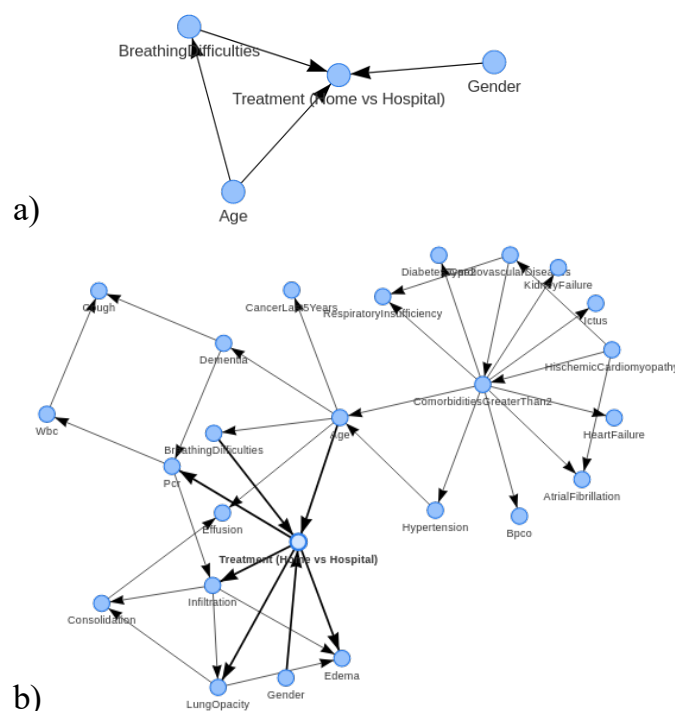
Figure 1: Directed Acyclic Graphs to model COVID-19 hospital admissions. a) Simple graph designed from existing evidence. b) Graph learned from data, starting from the simple graph in a)

3. Results

Here, we report predicted performance of the BN in Figure 1b, whose structure is based both evidence and from data. The simplest network, based only on evidence (Fig. 1a) shows good recall (i.e. the ability to correctly classify hospitalization) on the Test Set (85%), but low specificity (around 20%) and it was excluded from the analysis. We trained and tested three additional models: a regularized Logistic Regression, Gradient Boosting and Random Forest (RF). We show the performance of the RF only, since it outperforms the other two models on test data. RF is a widely applied ensemble classifier, that works by training several decision trees through bagging. Table 1 reports BN and RF classification performance on 66 patients of the Test set, in terms of various metrics, such as Area Under the ROC Curve (AUC) and Area Under the Precision-Recall Curve (PRC). Performances are quite similar, but BN shows slightly higher values for all the metrics. To test whether the error rates of the two approaches are significantly different, we apply the McNemar's Test [8]. P-value is 0.6, and we cannot reject the null hypothesis, i.e. the two classifiers have the same error rates. In Table 2 we can observe the 95% confidence interval prediction ability on the 462 third wave patients, where RF has slightly higher performance.

Also in this case the p-value of the McNemar's test calculated on the confusion matrix is high (0.7), and the estimated confidence intervals overlap. In comparison with Test results, both BN and RF show lower recall, but higher specificity. We examine the RF features importance by computing the mean decrease impurity The most important feature for classification is the protein C reactive value (Pcr),which was also directly linked to the outcome by the BN structure learning algorithm. Pcr levels usually increase when an inflammation is occurring. In RF, Pcr is followed by four DL-extracted features (LungOpacity, Edema, Consolidation and Infiltration). All these features, except for Consolidation, have a direct link to the outcome (1b). Age, gender and breathing difficulties are placed in the 8th, 9th and 10th positions.

4. Discussion

The need for clinical decision support systems implementing ML is increasing. These approaches are able to detect useful and hidden patterns in data, that can be exploited to support knowledge discovery and/or to implement automatic and highly accurate classifiers. Explainability of the classification process is needed to safely integrate ML within clinical practice, yet the majority of high-performing classifiers are perceived as *black-box*. Additionally, by learning entirely from training data, most of them prevent the integration of existing medical knowledge and evidence. Here, we show the development of a Bayesian Network whose aim is to predict hospital admission of COVID-19 patients. BN allows us to: 1) develop a model that is explainable by design and 2) combine known evidence about variable dependency with information encoded in patients data. The resulting structure can be inspected by clinicians to understand the classification process. The BN is able to generalize well during the third wave, despite some population variables, such as age, changed in comparison with patients of the first wave, used for training. Moreover, BN predictive ability is similar to a completely data-driven and less interpretable approach (RF). Future works will explore new networks configuration, with additional medical knowledge, and the exploration of potential causal relationships between variables.

References

[1] WHO Coronavirus (COVID-19) Dashboard, ???? URL: https://covid19.who.int.

[2] A. Alimadadi, S. Aryal, I. Manandhar, P. B. Munroe, B. Joe, X. Cheng, Artificial intelligence and machine learning to fight COVID-19, Physiological Genomics 52 (2020) 200–202. doi:10.1152/physiolgenomics.00029.2020, publisher: American Physiological Society.

[3] C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, D. King, Key challenges for delivering clinical impact with artificial intelligence, BMC Medicine 17 (2019) 195. doi:10.1186/s12916-019-1426-2.

[4] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, H. Müller, Causability and explainability of artificial intelligence in medicine, WIREs Data Mining and Knowledge Discovery 9 (2019) e1312. doi:10.1002/widm.1312.

[5] Y.-L. Chou, C. Moreira, P. Bruza, C. Ouyang, J. Jorge, Counterfactuals and Causability in

Explainable Artificial Intelligence: Theory, Algorithms, and Applications, arXiv:2103.04244 [cs] (2021). URL: http://arxiv.org/abs/2103.04244, arXiv: 2103.04244.

[6] S. Thirumuruganathan, M. Huber, Building Bayesian Network based expert systems from rules, in: 2011 IEEE International Conference on Systems, Man, and Cybernetics, 2011, pp. 3002–3008. doi:10.1109/ICSMC.2011.6084157, iSSN: 1062-922X.

[7] M. Scutari, C. E. Graafland, J. M. Gutiérrez, Who learns better Bayesian network structures: Accuracy and speed of structure learning algorithms, International Journal of Approximate Reasoning 115 (2019) 235–253. URL: https://www.sciencedirect.com/science/article/pii/S0888613X19301434. doi:10.1016/j.ijar.2019.10.003.

[8] T. G. Dietterich, Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms, Neural Computation 10 (1998) 1895–1923. URL: https://doi.org/10.1162/089976698300017197. doi:10.1162/089976698300017197, publisher: MIT Press.