pHealth 2021 B. Blobel and M. Giacomini (Eds.) © 2021 The authors and IOS Press. This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0). doi:10.3233/SHTI210582

Predictive Modeling of COVID and non-COVID Pneumonia Trajectories

NIKITA MRAMOROV¹, ILYA DEREVITSKII, SERGEI KOVALCHUK ITMO University, Saint Petersburg, Russia

Abstract. Today pneumonia is one of the main problems of all countries around the world. This disease can lead to early disability, serious complications, and severe cases of high probabilities of lethal outcomes. A big part of cases of pneumonia are complications of COVID-19 disease. This type of pneumonia differs from ordinary pneumonia in symptoms, clinical course, and severity of complications. For optimal treatment of disease, humans need to study specific features of providing 19 pneumonia in comparison with well-studied ordinary pneumonia. In this article, the authors propose a new approach to identifying these specific features. This method is based on creating dynamic disease models for COVID and non-COVID pneumonia based on Bayesian Network design and Hidden Markov Model architecture and their comparison. We build models using real hospital data. We created a model for automatically identifying the type of pneumonia (COVID-19 or ordinary pneumonia) without special COVID tests. And we created dynamic models for simulation future development of both types of pneumonia. All created models showed high quality. Therefore, they can be used as part of decision support systems for medical specialists who work with pneumonia patients.

Keywords. Pneumonia, disease simulation modeling, hidden markov models, bayesian network, machine learning

1. Introduction

With appearance of the COVID-19, our lifestyle has changed. It was uncommon practice to learn and work from home remotely, to wear a mask while going outside and generally to be in isolation for a large amount of time. The world manages to cope withCOVID-19 by creating vaccines and doing researches. However, curing and diagnosing one of the most infamous consequences of the COVID-19, pneumonia, is still a challenge for science. According to [1], 4,574,089 people around the world died to the consequences of COVID-19 by 2021 which includes pneumonia as on of the major factors. With 221,134,742 registered cases of COVID-19 infection, this numbers grows each week. There are some expectations of the COVID-19 third wave exists nowadays. Many researches are made in order to explore of COVID-19 pneumonia, to predict its behavior or mutations on different types of patients. The general approach for this task is the usage of mathematical models, such as Hidden Markov Model and Bayesian Network. Considering the facts above, pneumonia nowadays is considered as a strong threat towards elderly and humanity in general, so the pneumonia modeling task has a large significance these days.

¹ Corresponding Author: Nikita Mramorov, ITMO University, Saint Petersburg, Russia; Email: nmramorov98@gmail.com

2. Literature Review

For solving disease modeling tasks one of the best approaches are Hidden Markov Models and Markov Chains. For example, Kehinde Adiguna, Ayan Adeleke modeled the infections diseases trajectories with such methods. They established that the Markov chain could be used as a mathematical model for infectious diseases and revealed that the past history would affect the future only through the present state of infectious in disease which is a strong property of the Markov Chain [2]. Such approach could be applied to another modeling task. Tilahun Ferede Asena, Ayele Goshu, Mebratu Senbeta, Derbachew Asfaw Teni applied the semi-Markov modeling approach for HIV/AIDS progression. They claim that semi-Markov models are more agile in terms of parametric distribution choice than traditional Markov Chain [3].

Another option for such task includes usage of are Bayesian Networks. This class of models is famous for its graphical modeling approach that models the conditional probabilistic relationships of certain independent variables. For example, Hasan Aykut Karaboga, Aslihan Gunel and their colleges used Bayesian network as a decision tool for predicting ALS disease. They claim that Bayesian networks are widely used medicine or biology and extremely useful in terms of ease of use of posterior probabilities especially on risk assessment studies [4]. Another study group from National University of Singapore states that the underlying causal assumptions and interpretability make Bayesian networks a decent tool for medical applications and in CAD risk prediction [5].

Scientific research of disease using Computer Science and Data Analysis approaches, such as Deep Learning, is a significant research fields these days. Currently, COVID-19 pneumonia might be determined by generalized convolutional neural network model from chest X-ray images [7]. However, deep learning approach does not only used for tasks with COVID-19 or pneumonia. Authors [8] created recurrent disease progression networks for modeling risk trajectory of heart failure. Their study explored the ability of Recurrent Neural Networks in predicting long-term trajectories of recurrent events.

Each approach mentioned above has both advantages and disadvantages: Hidden Markov Models do not consider "past" states of variables and relies on the current state of presented variable; Bayesian Networks requires expert models with conditional probabilities for model evaluation and may not have all interconnections for complex systems. Our approach was developed in order to take into account advantages of both methods and decrease negative impact of both models' disadvantages.

3. Data

Data for this research was taken from the Federal State Budgetary Institution "V.A. Almazov National Medical Research Center" of the Ministry of Health of the Russian Federation. It consists of 234 patients with different pneumonia symptoms and other diagnosis. Among patients there are 63 % male and 37 % female correspondingly. In terms of pneumonia types the most common is nosocomial type.

Data consists of patients with COVID and non-COVID pneumonia. Average age for non-COVID pneumonia is 66, for COVID pneumonia is 71. P-value is more than 0.05 so our hypothesis is considered valid.

The suggested method in this article implies to figure out dependencies and connections for COVID and non-COVID pneumonia. Patients were divided into groups wherever any of them have COVID pneumonia or not. Such patients were explicitly marked by literals "U07.1" in their anamnesis. We propose a thesis that COVID pneumonia disease flow lies a greater pressure on patient health than non-COVID one. A comparison of pneumonia features that are considered sever according to clinical recommendations for COVID and non-COVID patients had been made. To properly estimate the thesis, we used Chi-Squared statistical test to compare two samples from separated COVID pneumonia and non-COVID pneumonia patient's data. Finally, the statistical patterns have been obtained and used for building disease simulation models.

4. Methods

The aim of this research is to explore COVID and non-COVID pneumonia trajectories and make its comparison based on statistical models, such as Bayesian Networks and Hidden Markov models. Figure 1 describes the algorithm for finding COVID and non-COVID pneumonia disease trajectories.



Figure 1. The diagram of the search algorithm for the best models.

There are three main steps in our research: data preprocessing, model engineering and models validation. Data preprocessing mostly consists of extracting data from clinical anamneses. Each patient has records of its every health manipulation.

5. Modeling and Results

This section refers to models analysis, interpretation and validation. Each model has produced unique set of trajectories for both COVID and non-COVID pneumonia. Authors applied modularity maximization techniques in order to locate the nodes groups and search for patterns and common trajectories of the disease.

First step involves building Hidden Markov Models for non-COVID and COVID pneumonia using Viterbi algorithm [6] (see Figure 2).



Figure 2. COVID Hidden Markov Model.

For both graphs cluster analysis had been applied. Each graph has three clusters with different features. Every cluster describes the trajectory of COVID pneumonia. Purple one describes COVID pneumonia directly as it has both COVID and pneumonia features within it. Largest nodes are coronary artery disease, chronic heart failure, dyspnea, infiltrate. Most of them describes one of the most common features of non-COVID pneumonia. Second cluster, colored in red, marks nodes, such as light fluid, infusion therapy, pus, invasive or non-invasive AV, fever, vasopressor support and resuscitation department as nodes with highest degree, hence, the most frequent one in COVID pneumonia trajectory. Last cluster consists of mostly features connected with breathing process, such as atrial fibrillation, bronchoscopy, vesicular breathing etc.

Next step involved creation of Bayesian Network models for both COVID and non-COVID pneumonia. For learning graph structure from the dataset we used Hillclimb algorithm due to its simplicity and fast convergence speed, and for parameter learning we chose to use BDeu score (see Figure 3).



Figure 3. COVID Bayesian Network.

COVID pneumonia model has 5 clusters, contains most of the features of non-COVID pneumonia. The key difference from the non-COVID model here is the existence of severe COVID pneumonia cluster. It contains respiratory failure, polyorganic insufficiency, invasive or non-invasive AV, mucous sputum. Those features are considered severe according to clinical recommendations.

5.1 Models Validation

This subsection involves model validation through task of making binary classification for a given set of patients whatever type of pneumonia (COVID or non-COVID) they have and evaluating its precision. For better comparison we offer four another classifiers that would challenge two build models: XGBoost linear model, Random Forest Classifier, Logistic Regression model.

Results of predictions are shown in Table 1.

Table 1. Predictions				
Method	Precision, %	Recall, %	Accuracy, %	F1-score, %
Hidden Markov Model	91	82	92	95
Bayesian Network	92	54.5	53.1	68.6
XGBoost linear model	88	51.2	53.4	66.6
Random Forest Classifier	91.5	90.3	91.4	95.5
Logistic Regression	92	32.5	38.2	49

Bayesian Network and Logistic Regression models reach the highest score among the models in terms of precision. However, in general, Hidden Markov Model shows better metrics on par with Random Forest Classifier.

6. Conclusions

In this paper authors proposed two dynamical models of simulating COVID and non-COVID pneumonia based on Hidden Markov Model and Bayesian Network architectures. Every model has shown COVID and non-COVID pneumonia unique trajectories, however Hidden Markov model turned out to be more preferable than the Bayesian Network due to its performance on models validation. Such process also provided that both models are able to detect most significant features of COVID and non-COVID pneumonia and might be useful for future research and medical analysis. Further work might include comparison of the built models with graphs made by experts for final validation and future tuning.

Acknowledgements

The reported study was funded by RFBR according to the research project № 20-31-90135.

References

- [1] WHO Coronavirus (COVID-19) Dashboard, https://covid19.who.int/, 2021
- [2] Kehinde Adiguna, Ayan Adelekeb, Oluwasesan Adewusic, Oluwafunmilola Olubiyid, Omobolaji Halide, Bayowa Babalolaf. Modelling Infectious Diseases Using Markov Chain, American Scientific Research Journal for Engineering, Technology, and Sciences (ASRJETS) 61(1):280, August 2020
- [3] Tilahun Asena, Ayele Goshu, Mebratu Senbeta, Derbachew Teni, Semi-Markov Modelling of HIV/AIDS Disease Progression, December 2020
- [4] Hasan Aykut Karaboga, Aslihan Gunel, Senay Vural Korkut, Ibrahim Demir and Resit Celik. Bayesian Network as a Decision Tool for Predicting ALS Disease, Brain Sciences 11(2):11, January 2021
- [5] Qiongyu Chen1, Guoliang Li, Bin Han, Chew Kiat Heng, and Tze-Yun Leong. Coronary Artery Disease Prediction with Bayesian Networks and Constraint Elicitation, September 2006
- [6] G.D. Forney. The viterbi algorithm, Proceedings of the IEEE (Volume: 61, Issue: 3, March 1973), Codex Corporation, Newton, MA, USA, 1973
- [7] Adi Alhudhaif, Kemal Polat, Onur Karaman, Determination of COVID-19 pneumonia based on generalized convolutional neural network model from chest X-ray images, 2021
- [8] Xing Han Lu, Aihua Liu, Recurrent disease progression networks for modelling risk trajectory of heart failure, 2021