

ADVANCED MACHINE LEARNING MODELS IN PREDICTION OF MEDICAL CONDITIONS

A Dissertation
Submitted to
the Temple University Graduate Board

In Partial Fulfillment
of the Requirements for the Degree of
DOCTOR OF PHILOSOPHY

by Branimir Ljubic
May 2021

Examining committee members:

Dr. Zoran Obradovic, Advisory Chair, Department of Computer and Information Sciences
Dr. Slobodan Vucetic, Department of Computer and Information Sciences
Dr. Xinghua Mindy Shi, Department of Computer and Information Sciences
Dr. Daniel Rubin, External Member, School of Medicine

ABSTRACT

The primary goal of Machine learning (ML) models in the prediction of medical conditions is to accurately predict (classify) the occurrence of a disease, or therapy. Many ML models, traditional and deep, have been utilized for the prediction of disease diagnosis, or prediction of the most optimal therapeutic approach. Almost all categories of medical conditions were subject to ML analysis. When creating predictive ML algorithms in medicine, it is pivotal to consider what problems are intended to be solved and how much and what types of training data are available. For challenging prediction (classification) problems, the understanding of disease pathogenesis makes the selection of an adequate ML model and accurate prediction more likely.

The hypothesis of the research was to demonstrate that the optimal and adequate selection of model inputs as well as the selection and design of adequate ML methods improves the prediction accuracy of occurrence of diseases and their outcomes. The effectiveness and accuracy of created deep learning and traditional methods have been analyzed and compared. The impact of different medical conditions and different medical domains on optimal selection and performance of ML models was also studied. The effectiveness of advanced ML models was tested on four different diseases: Alzheimer's disease (AD), Diabetes Mellitus type 2 (DM2), Influenza, and Colorectal cancer (CRC).

The objective of the first part of the thesis (AD study) was to determine could prediction of AD from Electronic medical records (EMR) data alone be significantly improved by applying domain knowledge in positive dataset selection rather than setting naïve filters. Selected Clinically

Relevant Positive (SCRIP) datasets were used as inputs to a Long-Short-Term Memory (LSTM) Recurrent Neural Network (RNN) deep learning model to predict will the patient develop AD. The LSTM RNN method performed significantly better when learning from the SCRIP dataset than when datasets were selected naïvely. Accurate prediction of AD is significant in the identification of patients for clinical trials, and a better selection of patients who need imaging diagnostics.

The objective of the DM2 research was to predict if patients with DM2 would develop any of ten selected complications. RNN LSTM and RNN Gated Recurrent Units (GRU) models were designed and compared to Random Forest and Multilayer Perceptron traditional models. The number of hospitalizations registered in the EMR data was an important factor for the prediction accuracy. The prediction accuracy of complications decreases over time. The RNN GRU model was the best choice for EMR type of data, followed by the RNN LSTM model. An accurate prediction of the occurrence of complications of DM2 is important in the planning of targeted measures aimed to slow down or prevent their development.

The objective of the third part of the thesis was to improve the understanding of spatial spreading of complicated cases of influenza that required hospitalizations, by constructing social network models. A novel approach was designed, which included the construction of heatmaps for geographic regions in New York state and power-law networks, to analyze the distribution of hospitalized flu cases. The methodology constructed in the study allowed to identify critical hubs and routes of spreading of Influenza, in specific geographic locations. Obtained results could enable better prediction of the distribution of complicated flu cases in specific geographic regions and better prediction of required resources for prevention and treatment of hospitalized patients with Influenza.

The fourth part of the thesis proposes approaches to discover risk factors (comorbidities and genes) associated with the development of CRC, which can be used for future ML models to

predict the influence of risk factors on prognosis and outcomes of cancer and other chronic diseases. A novel social network and text mining model was developed to study specific risk factors of CRC. Identified associations between comorbidities, CRC, and shared genes can have important implications on early discovery, and prognosis of CRC, which can be subject to predictive ML models in the future.

Prediction ML models could help physicians to select the most effective diagnostic, preventive and therapeutic choices available. These ML models can provide recommendations to select suitable patients for clinical trials, which is very important in searching for medical solutions in health emergencies. Successful ML models can make medicine more efficient, improve outcomes, and decrease medical errors.

To all my family and friends

To all future readers

ACKNOWLEDGEMENTS

I would like to recognize with the greatest admiration my advisor, academic, Dr. Zoran Obradovic for accepting me into the Ph.D. program, for advices, valuable opinions, support, and for providing me with the highest level of scientific expertise in big data analytics and machine learning. I would also like to express my gratitude to academic Dr. Obradovic for funding my research and education.

I am very grateful to Dr. Slobodan Vucetic, for giving me support and sharing with me his high level of expertise in computer science and machine learning, which helped me to learn more and successfully complete my projects. I would also like to thank Dr. Vucetic for being a member of my dissertation committee.

I would like to express my gratitude to Dr. Xinghua Mindy Shi for serving on my dissertation committee and providing valuable feedback and insights that improved my dissertation. I express my gratitude to Dr. Mindy Shi for helping me to learn the techniques of computational biology. I would like to express my admiration to Dr. Daniel Rubin for serving on my dissertation committee and providing knowledgeable medical inputs that improved my dissertation.

I am grateful to Dr. Eduard Dragut for teaching me the management of databases and cloud computing, for invaluable help in navigating administrative procedures, and for supporting

me as his teaching assistant. I would also like to thank to Dr. Ben He, Dr, Richard Biegel, Athanasia Polychronopoulou, Dr. Xiqui (Cindy) Li for their support while I was working with them as a teaching assistant.

I have very special thanks to my colleague Martin Pavlovski for his friendly support and professional help. I also have special thanks to dear colleagues and friends Djordje Gligorijevic, Jelena Gligorijevic, Dusan Ramljak, and Ivan Stojkovic for supporting my professional work during my first two years in the Ph.D. program.

Thanks to my co-authors and friends Xi Hang Cao, Martin Pavlovski, Marija Stanojevic, Ameen Abdel Hai, Shoumik Roychoudhury, Wilson Diaz, Stefan Obradovic, Djordje Gligorijevic, Jelena Gligorijevic, Richard Nair, Glass Lucas, Jumanah Alshehri, Daniel Polimac, Avrum Gillespie, Daniel Rubin, and researchers from the KAUST group.

Thanks also, to my other colleagues and friends Jovan, Danijel, Chao, Nouf, Noor, Nima, Hussain, and other people I might leave out for sharing their scientific and friendly opinions with me.

I would also like to thank all my family, sister Vesna, son Robert, wife Luann, mother Mladena, father Vencel, and others, as well as to my friends, Aleksandar Aleksic, Aleksandar Mikic, and many others for their encouragement and support.

TABLE OF CONTENTS

| | |
|--|------------|
| ABSTRACT | ii |
| ACKNOWLEDGMENT | vi |
| LIST OF TABLES | xi |
| LIST OF FIGURES | xiv |
| CHAPTER | 1 |
| 1. INTRODUCTION..... | 1 |
| 1.1. Search and selection criteria of relevant articles | 2 |
| 1.2. Literature review..... | 4 |
| 1.2.1. Traditional models..... | 5 |
| 1.2.2. Deep learning models..... | 10 |
| 1.3. Interpretation..... | 16 |
| 1.4. Conclusion..... | 19 |
| 2. INFLUENCE OF MEDICAL DOMAIN KNOWLEDGE ON DEEP LEARNING FOR ALZHEIMER’S DISEASE PREDICTION..... | 21 |
| 2.1. Introduction..... | 21 |
| 2.2. Methods..... | 23 |
| 2.3. Results..... | 29 |
| 2.4. Discussion..... | 36 |

| | |
|---|-----------|
| 2.5. Conclusion..... | 39 |
| 3. PREDICTING COMPLICATIONS OF DIABETES MELLITUS USING ADVANCED MACHINE LEARNING ALGORITHMS..... | 40 |
| 3.1. Objective..... | 40 |
| 3.2. Background and significance | 42 |
| 3.3. Materials and methods | 43 |
| 3.4. Results..... | 50 |
| 3.5. Discussion | 56 |
| 3.6. Conclusion | 59 |
| 4. SOCIAL NETWORK ANALYSIS FOR BETTER UNDERSTANDING OF INFLUENZA | 61 |
| 4.1. Objective | 61 |
| 4.2. Background and significance | 63 |
| 4.3. Materials and methods | 67 |
| 4.4. Results | 71 |
| 4.5. Discussion | 84 |
| 4.6. Conclusion | 87 |
| 5. COMORBIDITY NETWORK ANALYSIS AND GENETICS OF COLORECTAL CANCER | 89 |
| 5.1. Introduction | 89 |
| 5.2. Materials and Methods | 91 |
| 5.3. Results | 95 |
| 5.3.1 Comorbidity analysis | 95 |
| 5.3.2. Comorbidity networks | 100 |

| | |
|--|------------|
| 5.3.3. Genes associated with CRC and comorbidities | 104 |
| 5.4. Discussion | 112 |
| 5.5. Conclusion | 116 |
| 6. CONCLUSION | 117 |
| BIBLIOGRAPHY | 120 |

LIST OF TABLES

| | | |
|------|--|----|
| 1.1. | Supervised ML methods applied in prediction of medical conditions. Absolute numbers of articles in PubMed (total and in the last two years) and percentage of the total number of publications are presented | 5 |
| 1.2. | Reviewed articles, classified by ML model types. First authors, publication year, and medical topics described in the publications are presented..... | 15 |
| 1.3. | The number of analyzed articles by types of medical conditions..... | 16 |
| 2.1. | Datasets used in the experiments and the number of patients in each of them..... | 26 |
| 2.2. | Prediction of AD by LSTM RNN trained on the SCRP dataset using the drugs, the measurements, the condition domain, and their ensemble (c.m.d.), with the fixed size of the testing dataset (234 patients) and different sizes of the training dataset. The evaluation metric – AUPRC..... | 34 |
| 2.3. | Results of experiments with SCRP datasets with different numbers of visits (2, 3, or 4) for drugs, measurements, and conditions domains as well as for an ensemble of all 3 domains (c.m.d.)..... | 35 |
| 3.1. | Datasets used in experiments and their sizes..... | 50 |
| 3.2. | Presented are results of predicted accuracy (and standard deviation) that each of the ten complications of DM2 will develop within a 9 years period after the first DM2 diagnosis using HCUP EMR data (diagnoses domain). This period varies between 1 month and 9 years for individual patients. Results are presented for patients who had at least 2, 3 or 4 visits between the first DM2 diagnosis and before each of ten complications was diagnosed. The first column: names of complications; second column: number of patients in positive cohorts for each of complications; third column: Bi-directional GRU RNN classifier; fourth column: 1-way LSTM RNN classifier, fifth column: RF classifier, sixth column: MLP classifier. Bold are the best accuracy results for each experimental setting. Italic Bold are the best overall accuracy results for each of ten complications..... | 51 |
| 3.3. | Accuracy, Sensitivity and Specificity for Bi-directional GRU RNN models in the 4-visits scenario for all 10 complications of DM2..... | 52 |

| | | |
|------|--|-----|
| 3.4. | Experiments conducted on DR datasets with 2 and 4 hospitalizations in order to test changes in accuracy results with the decrease of the training dataset size. The first column: the size of training dataset; second column: number of patients in positive cohorts for each of experimental settings; third column: Bi-directional GRU RNN classifier; fourth column: 1-way LSTM RNN classifier, fifth column: RF classifier, sixth column: MLP classifier. Bold are the best accuracy results of the each type of experiments..... | 53 |
| 4.1. | Distribution of population per zip codes (20 the largest) in the state of New York from census data for 2010..... | 74 |
| 4.2. | Number of hospitalized flu patients in the state of New York for the period 2003-2012 (HCUP data)..... | 74 |
| 4.3. | Percentages of affected population for 20 zip codes with the highest percentages of hospitalized flu patients..... | 77 |
| 4.4. | Zip code with the highest node degrees | 81 |
| 5.1. | Comorbidities in early and advanced stages of CRC in females, older than age 50. 1a: Ranked comorbidities occurred in the early stages. 1b: Ranked comorbidities occurred in the advanced stages of CRC. The prevalence of the occurrence of comorbidities is shown..... | 97 |
| 5.2. | Comorbidities in early and advanced stages of CRC in males, older than age 50. 2a: Ranked comorbidities occurred in the early stages. 2b: Ranked comorbidities occurred in the advanced stages of CRC. The prevalence of the occurrence of comorbidities is shown..... | 98 |
| 5.3. | Ranked comorbidities in CRC patients, age 50 and younger: Males (3a), and females (3b). The prevalence of the occurrence of comorbidities is shown..... | 99 |
| 5.4. | APC genes associated with CRC and comorbidities. Presented are numbers of abstracts in PubMed where these associations were described in 3 or more abstracts..... | 106 |
| 5.5. | TP53 genes associated with CRC and comorbidities. Presented are numbers of abstracts in PubMed where these associations were described in 20 or more abstracts..... | 107 |
| 5.6. | KRAS genes associated with CRC and comorbidities. Presented are numbers of abstracts in PubMed where these associations were described in 4 or more abstracts..... | 108 |
| 5.7. | MLH1 genes associated with CRC and comorbidities. Presented are numbers of abstracts in PubMed where these associations were described in 3 or more abstracts..... | 109 |

| | | |
|-------|--|-----|
| 5.8. | TGFBR2 genes associated with CRC and comorbidities. Presented are numbers of abstracts in PubMed where these associations were described in 2 or more abstracts..... | 110 |
| 5.9. | PPARG genes associated with CRC and comorbidities on PubMed. Comorbidities that are associated with CRC and PPARG in 10 or more abstracts are given..... | 111 |
| 5.10. | Ninety-six genes associated with CRC and present in both sources: PubMed and DisGeNET..... | 111 |

LIST OF FIGURES

| | | |
|------|--|----|
| 1.1. | Number of publications describing Machine learning applications in Medicine indexed in PubMed | 4 |
| 2.1. | LSTM RNN deep learning system designed in this research: an input layer (patient timeline), an embedding layer, a sequence modeling layer, and an output layer (disease risk score)..... | 27 |
| 2.2. | The naïve model results, AUPRC score obtained by LSTM RNN for condition and measurement domains separately and for their ensemble (c.m.) when using the naïve AD dataset selection for patients with at least 4 visits..... | 30 |
| 2.3. | The average precision/recall comparison of models using condition, measurement, and an ensemble of two domains (combined)..... | 31 |
| 2.4. | AUPRCs of AD predictions by LSTM RNN trained on the SCRP dataset using each of three domains (condition, drug, measurement) separately and as an ensemble (c.m.d)..... | 32 |
| 2.5. | The average precision/recall comparison of LSTM RNN models trained on the SCRP dataset using condition, measurement, and drugs domains and ensemble information of all three domains (c.m.d.)..... | 33 |
| 2.6. | Prediction of AD by LSTM RNN model, trained on the SCRP dataset using the drugs domain in different splits of the dataset for training and testing..... | 35 |
| 3.1. | Each row (Fig. 3.1a) represents one patient (Pt). Different colors in each row represent different hospitalizations. Each hospitalization contained one or more diagnoses (d) and sometimes procedures (p). Since the procedures domain did not produce good results, we dropped them (Fig. 3.1b) and performed analyses on the diagnoses domain only..... | 43 |
| 3.2. | The proposed deep learning models: one-way RNN LSTM (Fig. 2a) and bi-directional RNN GRU (Fig. 2b)..... | 46 |
| 3.3. | Prediction accuracy (RNN GRU model) that patients with DM2 will develop Diabetic Retinopathy after a minimum of 2 hospitalizations Fig. 3a), and after at least 4 hospitalizations (Fig. 3b). The results are presented by intervals when | |

| | |
|--|-----|
| Retinopathy developed: within 1 year, after 1, 2, 3, 4-5 and 6-8 years from the diagnosis of DM2..... | 54 |
| 3.4. Predicted risk probabilities of development of each of ten complications in patients with DM2 (HCUP SID California data)..... | 55 |
| 4.1. Bar plot – Number of patients infected by the influenza virus during the 2003-2012 period (monthly distribution), who were hospitalized in the state of New York..... | 71 |
| 4.2. Heat map of NY state - Distribution of hospitalized flu patients by the zip code (2003-2012), shows that the highest concentration of hospitalized flu patients were in five big urban areas (Albany, Buffalo, New York City, Rochester and Syracuse). The heatmap also shows that routes of spreading follow highways (in particular Highways 81, 86 and 90) as high frequency routes of travelling between places..... | 72 |
| 4.3. Heatmaps of a) Albany area, b) NY City area, c) North side of NY state, d) Buffalo area. Heatmaps show that the distribution of hospitalized flu patients by the zip code in period between 2003-2012, was highly concentrated in five big cities (Albany, Buffalo, New York City, Rochester and Syracuse). Heatmaps show that the routes of distribution follow highways (highways 81, 86 and 90, in Albany area highways 87 and 9 and in Buffalo area, highway 190 toward Niagara Falls)..... | 73 |
| 4.4. a) Plot of correlation between 1,471 zip codes with respect to number of hospitalized patients with flu complications in those zip codes between 2003 – 2012 b) Plotted the estimate of the power law exponent, the log–log data, fitted line at $\gamma = 2.5935$ | 79 |
| 4.5. The power law type network representation of hospitalized flu cases in the State of New York between 2003 and 2012. based on correlation between zip codes. Nodes correspond to zip codes that are linked, based on strength of calculated correlations. Zip codes (nodes in the network) with the highest node degrees are: 10465 (Bronx) with the degree of 86, then 11226 (Brooklyn) with the degree of 84 and 10027 (NY City) – degree 80..... | 80 |
| 4.6. Geographic location of 20 zip codes with the highest nodes degrees (hubs) in the state of New York. a) the whole NY State, b) NY City area –20 top hubs are in this area. Zip codes (dots on maps) with the highest node degrees are: 10465 (Bronx) with the degree of 86, 11226 (Brooklyn) with the degree of 84 and 10027 (NY City) – degree 80..... | 83 |
| 5.1. Networks of comorbidities in the early stages of CRC in females older than age 50. Comorbidity networks, based on: a) ϕ -correlation and b) RR in females. Nodes represent the top 100 comorbidities (ICD9 codes). Correlations greater than 0.32 ($\beta = 2$) and RR greater than 5.99 ($\beta = 5$) were applied for construction of edges..... | 100 |

| | | |
|------|--|-----|
| 5.2. | Networks of comorbidities in the advanced stages of CRC in females older than age 50. Comorbidity networks, based on: a) ϕ -correlation and b) RR in females. Nodes represent the top 100 comorbidities (ICD9 codes). Correlations greater than 0.30 ($\beta = 2$) and RR greater than 8.99 ($\beta = 5$) were applied for construction of edges..... | 101 |
| 5.3. | Networks of comorbidities in the early stages of CRC in males older than age 50. Comorbidity networks, based on: a) ϕ -correlation and b) RR. Nodes represent the top 100 comorbidities (ICD9 codes). Correlations greater than 0.32 ($\beta = 2$) and RR greater than 8.99 ($\beta = 5$) were applied for edges..... | 102 |
| 5.4. | Networks of comorbidities in the advanced stages of CRC in males older than age 50. Comorbidity networks, based on: a) ϕ -correlation and b) RR. Nodes represent the top 100 comorbidities (ICD9 codes). Correlations greater than 0.30 ($\beta = 2$) and RR greater than 8.99 ($\beta = 5$) were used for edges..... | 103 |

CHAPTER 1

INTRODUCTION

Over the last decade, there has been significant growth in the amount of medical data generated by the adoption and integration of electronic health records (EHR).¹ This growth in EHR data coincided with rapid development of ML techniques and computing power to analyze Big Data in medicine, which could contribute to improved medical solutions, and better and more efficient healthcare.²⁻⁵ ML is a branch of Artificial intelligence (AI), aimed to make the computer learn from past experiences and make predictions by recognizing patterns in medical data.⁵⁻⁷ ML can be classified into three categories: unsupervised, supervised, and reinforcement learning (RL). This paper focuses on supervised ML techniques, where a function that maps an input to output is inferred from labeled training data.

Supervised ML models are classified in traditional and deep learning approaches.⁵⁻⁷ The most frequently used traditional ML models in medicine are: decision trees (DT),⁸ random forest (RF),⁹ and other ensemble methods,¹⁰⁻¹⁴ single and multi-layer perceptron (MLP),^{15,16} Bayesian learning (BL),¹⁷ support vector machines (SVM),¹⁸ k-nearest neighbors (k-NN),¹⁹ linear regression (LR),²⁰ and logistic regression (LogR).²¹ Deep

learning models are inspired by the biological neural networks, where each layer of the network learns higher order features of the previous layer. Different types of neural networks have been designed, such as: deep neural networks (DNN) including deep belief networks,²² convolutional neural networks (CNN),²² recurrent neural networks (RNN - long short-term memory (LSTM) and gated rectified unit (GRU)),²³ etc.

We performed a comprehensive systematic literature review of recent publications that have used either traditional or deep supervised ML techniques in healthcare. We examined whether the method is appropriate for the selected medical prediction task, whether the model is generalizable, and whether it could be used by clinicians to improve the quality of care.³⁻⁷ Since this is a fast-growing scientific field, we included articles published within the last two years.

1.1. Search and selection criteria of relevant articles

Many ML models have been utilized for the prediction of diagnosis, or recommendation of the most optimal therapeutic approach. We developed the following protocol to review recent supervised prediction ML models in medicine.

We conducted a systematic review of articles that described supervised prediction ML models published in the last two years (01/01/2019 –12/31/2020). In our research, we considered the classification of supervised ML models on traditional and deep models. The following traditional supervised ML models were included in the review: DT, RF and other ensemble methods, the perceptron, BL models, SVM, k-NN, LR, and LogR. Deep learning models included in the review were: DNN, CNN, and RNN.

We searched PubMed/MEDLINE, as the most relevant source of healthcare-related topics, to identify articles describing applications of supervised traditional and deep predictive ML models in medicine. Initially, we searched PubMed using MeSH (Medical Subject Headings) major terms. If we did not find enough literature for the specific ML model, we applied a broader search using terms that appear in titles of publications, or in the case of bagging and boosting ensemble models and RNN we used combinations of MeSH and title/abstract searches. This searching approach extracted articles where the reviewed ML model was a major part of the article. The search strategy used keywords indicating “ML model” AND “prediction/detection/classification” AND “medical conditions (therapies, outcomes)”. In the example of deep learning methods, we divided searches into 3 groups: classic DNN, CNN, and RNN. To extract DNNs we used the following search query: ("deep learning"[Mesh] not "CNN" not "RNN" not "LSTM" not "GRU"). For CNN models the query was: (CNN[Title]) OR (convolutional neural networks[Title]). And for RNN model the query was: (RNN[Title/abstract] OR recurrent neural networks[Title] OR LSTM[Title] OR GRU[Title] OR long short term memory[Title] OR gated rectified unit[Title]).

We included articles that presented original research published in English. The search results were sorted according to types of described ML models. For each of the reviewed ML models, we selected a representative sample. We tried to include most of the human body systems. Priority was given to the newest published research in the case of multiple papers describing similar predictive ML approaches. We performed a systematic review of

the literature with the objective to analyze how useful and meaningful are described predictive ML models from the point of real applicability in medical practice.

1.2. Literature review

We identified 12,335 articles (MeSH Major topic) describing ML implementation in medicine (Figure 1.1.). 6,407 articles were published in the last 2 years.

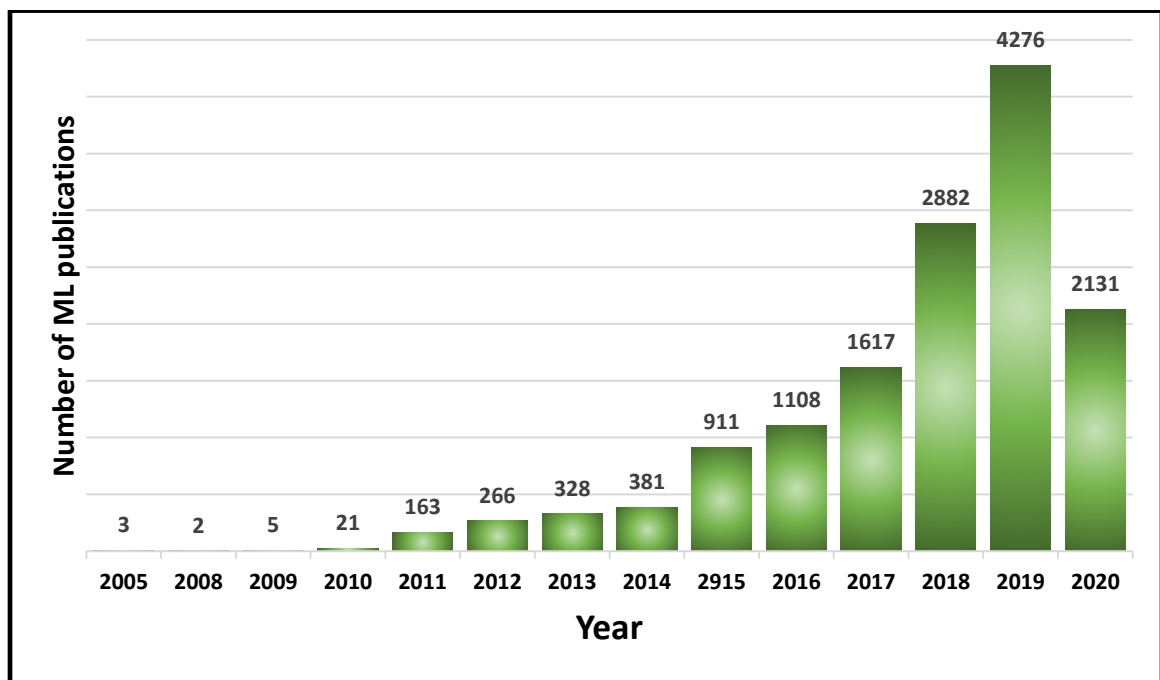


Figure 1.1. Number of publications describing Machine learning applications in Medicine indexed in PubMed as of 12/31/2020.

50 articles were included in the review. In Table 1.1. we present the method of search for each of the ML models and the number of publications retrieved using that particular search method. Most of the published ML models in medicine in the last two years were different types of deep learning models (74%).

| ML method | Type of ML | Type of PubMed Search | No of articles total | Articles in the last 2 years |
|---------------------|-------------|-----------------------|----------------------|------------------------------|
| Linear regression | Traditional | MeSH Major topic | 2,703 (16%) | 106 (2%) |
| SVM | Traditional | MeSH Major topic | 2,360 (14%) | 378 (7%) |
| DT | Traditional | MeSH Major topic | 1,919 (11%) | 139 (3%) |
| Logistic regression | Traditional | MeSH Major topic | 1,865 (11%) | 100 (2%) |
| Bayesian | Traditional | Title search | 1,072 (6%) | 61 (1%) |
| RF | Traditional | Title search | 767 (4%) | 266 (5%) |
| Ensemble - boosting | Traditional | Title search | 460 (3%) | 246 (5%) |
| k-NN | Traditional | Title search | 264 (1.5%) | 47 (0.9%) |
| Perceptron | Traditional | Title search | 247 (1%) | 38 (0.7%) |
| Ensemble - bagging | Traditional | Title search | 71 (0.4%) | 34 (0.6%) |
| DNN | Deep | MeSH Major topic | 2,207 (13%) | 1,880 (34%) |
| CNN | Deep | Title search | 1,862 (11%) | 1,338 (25%) |
| RNN | Deep | Title search | 1,532 (9%) | 828 (15%) |

Table 1.1. Supervised ML methods applied in prediction of medical conditions. Absolute numbers of articles in PubMed (total and in the last two years) and percentage of the total number of publications are presented.

1.2.1. Traditional models

Decision trees. DTs are classification methods that adopt a top-down strategy, where each node represents a classification question and the branches are partitions of the data into different classes.⁸

The terminal nodes of DTs represent different classes. DTs were developed for prediction of diabetes mellitus type 2 (DM2), and essential hypertension (EH),²⁴ by creating visually guided classification trees to facilitate the feature selection (four different datasets, sizes 547 – 12,447). The prediction accuracy of DM2 and EH in different scenarios varied between 0.58 and 0.87. DTs predicted coronary artery disease with the accuracy about 0.91 on a dataset of 303 patients,²⁵ and hospital-acquired pneumonia (accuracy \approx 0.91) among

185 schizophrenic patients.²⁶ A DT algorithm was proposed to identify pre-treatment clinical predictors of survival in rectal cancer (100 examples).²⁷ Predicted accuracy of survival rates for five and seven years were 0.76 and 0.71. The authors pointed out the necessity to carefully analyze the classification error in order to choose the most important predictor variables. Presented DT models have questionable generalization potential, since most of them were developed on small samples usually from one site. Some models required a lot of feature engineering.²⁴

Random Forrest. RF is a supervised ensemble learning method that creates many decision trees to predict the outcome.⁹ RF models have been constructed to predict hepatotoxicity on a dataset of 346 samples, with the accuracy prediction result up to 0.71.²⁸ This may provide a basis for improved safety evaluation in drug discovery and the risk assessment of environmental pollutants. Clinical utility would be better with a more predictive model. A retrospective review of patients undergoing abdominal hernia was the basis for RF modeling to predict surgical approach and determine the importance of different socioeconomic variables in selecting the type of surgery (Area under the receiver operating characteristic (AUROC) \approx 0.82).²⁹ Data were obtained from a single institution (559 patients), which limits the generalizability of findings. These models also need to encounter individual preferences of doctors who administer treatments. Psychotic and depressive symptom clusters in dementia were predicted using RF on the EHR records of 4,003 patients with dementia (AUROC \approx 0.80).³⁰ An RF model boosted by the AdaBoost algorithm was utilized to predict the severity of COVID-19 cases and the possible outcome, recovery, or death by using a patient's geographical, travel, health, and demographic data

(3397 patients, accuracy 0.94). The model revealed a positive correlation between patients' gender and deaths, finding that men are more likely to die.³¹ All presented RF models need further improvements, possibly combinations with deep learning models to find a relevant application in practice.

Ensemble Methods. Ensemble ML algorithms (boosting, bagging) combine a few weak learning models and turn them into a better learning algorithm.¹⁰⁻¹⁴ The most famous boosting algorithm is AdaBoost. Bagging and boosting ensemble methods were compared in automated electromyographic (EMG) signal (2,400 instances) classification to diagnose neuromuscular disorders. The AdaBoost with RF ensemble method achieved an accuracy of 0.99.³² The model required extensive feature engineering and it's not obvious how will it generalize on other datasets. The bagging ensemble method improves reproducibility of both cortical and subcortical functional parcellation of the human brain neuroimaging (more than 300 samples).³³ AdaBoost was used for identifying Alzheimer's disease (AD) from MRI scans (Alzheimer's disease Neuroimaging Initiative (ADNI) database of about 3,000 images) with the accuracy higher than 0.90,³⁴ and to differentiate colorectal neoplasia from normal tissue (AUROC up to 0.95 on 64 samples from 16 patients).³⁵ All three models use imaging data as inputs and need additional research work, including comparison to deep learning models, before their potential application in clinical practice.

Perceptron and Multilayer Perceptron. The perceptron is based on a threshold function that learns weights for features and processes one example at a time. It could be used as a single-layer perceptron or as a multilayer perceptron.^{15,16} Modeling of the spread of the COVID-19 infection using a MLP was designed on a dataset (20,706 examples) operated

by the Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE).³⁶ This is one of many papers about COVID-19, intended to predict the spread of the infection, but the model lacks generalization ability. MLP was applied in a diagnosis of breast cancer subtypes, using MRI images (704 images) with AUROC of 0.86.³⁷ The study used imaging data to distinguish between benign and malignant breast lesions. The challenge remains how to effectively incorporate this model into everyday oncology.

Bayesian ML approach. Bayesian ML algorithms calculate probabilities for hypotheses. The class having maximum probability is assigned as the most suitable class.¹⁷ A Bayesian approach was implemented in drug discovery on more than 2,000 compounds to predict drug binding targets (accuracy ≈ 0.90).³⁸ Computational approaches have the potential to significantly reduce the research work needed for drug target identification. A Bayesian methodology was also implemented in the prediction of risk of coronary artery disease on 303 patients and achieved AUROC ≈ 0.92 .³⁹ Presented Bayesian models will need to be tested on more data samples, to confirm clinical applicability.

Support Vector Machine. SVM model applies an optimization problem that attempts to find a separating hyperplane with as large a margin as possible.¹⁸ SVM was trained on 318 samples to distinguish neurodegenerative movement disorders such as Parkinson's Disease (PD) from healthy subjects, and from other movement disorders (precision of 0.81 and recall of 0.89).⁴⁰ In this study, DNN and RF were applied to the same task, with DNN achieving the best results. Models were trained on retrospective data at a single site, with high data quality and none of the different classifiers outperformed the others, which are some of limitations of these models. SVM models have also been used for diagnosis of

early breast cancer using PET images (116 samples, accuracy up to 0.85, AUROC 0.89),⁴¹ and detection of atrial fibrillation (AF) using ECG data (79 AF and 336 non-AF cases, accuracy 0.97-1.00).⁴² Both models use images as inputs. These models need to be tested on bigger data and compared to deep learning since recent literature shows that deep learning models yield better predictions than SVM models on imaging data.

K-Nearest Neighbors. K-NN algorithm takes into account k-neighboring points when classifying a data point and assigns the class by finding the most prominent class among the k-nearest data points.¹⁹ A k-NN method was implemented in the assessment of the aggressiveness of prostate cancer in 99 patients using zonal-specific features from MRI (AUROC 0.83-0.98).⁴³ A k-NN classifier was also trained for a microwave dielectric property-based classification of renal calculi (105 samples, accuracy \approx 0.98).⁴⁴ The benefits of the proposed method include the decrease of diagnostics time and equipment costs. The model should be tested on datasets that include more diverse data of renal calculi. Presented k-NN models used imaging data as inputs. They should be tested on bigger datasets and compared with deep learning models.

Linear regression. LR algorithm performs a regression task and predicts a specific value based on an independent variable.²⁰ A linear model was created to predict the impact of the duration of exposure (number of days) to COVID-19 on mortality rates (more than 270,000 patients).⁴⁵ Another application of linear models was developed to model the cost of care for children with cystic fibrosis (73 patients).⁴⁶ Multiple regression and LR analysis were successfully applied to predict the number of weekly deaths due to COVID-19 in India

(606 patients).⁴⁷ LR models are good options for time to event type of predictions and for predictions of the exact numbers of patients, or the cost of care.

Logistic regression. LogR uses the logistic function to binary classification and estimates the probability of the event.²¹ Multivariate LogR (as well as RF and XGBoost) models were created to identify and rank clinical features for prognosis of mortality risk in patients with COVID-19 (292 patients, AUROC \approx 0.95).⁴⁸ The model needs to be tested on larger multi-center data, and it needs to be enriched with facts about COVID-19 discovered from the newest studies. A study investigated the application of LogR and RNN LSTM models in capturing clinical risk factors for outcome prediction of 575 patients with aneurysmal subarachnoid hemorrhage (AUROC 0.89).⁴⁹ Longitudinal clinical and imaging risk factors were used as inputs. Since the LSTM RNN model achieved higher accuracy, it is likely a better choice in this type of study. A general logistic model was developed to predict the disease risk using multiple single-nucleotide polymorphism (SNP) risk factors (57 patients).⁵⁰ The authors claimed that similar logistic model-based algorithms can be established for common complex diseases (e.g., diabetes, cardiovascular disease and cancers) reportedly linked to multiple SNP risk factors. This claim has yet to be tested.

1.2.2. Deep learning models

Deep Neural Networks. DNN is a multilayer neural network with an input layer, hidden layers and an output layer. They are feedforward networks that use a nonlinear function of a linear combination of the inputs multiplied by the coefficients (weights, bias). DNN learns weights so the output from the network correctly classifies the example. The back

propagation algorithm is a standard approach to train DNNs.²² Recently, researchers increasingly used Deep learning methods for multiple aspects of medicine. A compartmental model enhanced with deep learning methodology predicted the dynamics of the COVID-19 epidemic in the U.S, using the JHU CSSE data repository.⁵¹ The model predicted the number of actively infected cases between 3.2-3.3 million on August 16-18, 2020. The real number of infected cases on August 16-18, 2020 was about 2.5 million (CDC data), so the model was not accurate. A mathematical spread model of COVID-19 with time-dependent parameters using DNN method was designed, to predict the dynamics of the pandemic outbreak on JHU CSSE data and Korea Centers for Disease Control and Prevention data.⁵² The authors estimated parameters and outcome variables of the susceptible-infected-recovered (SIR) model using DNNs for South Korea. Predictions of different aspects of COVID-19 epidemic are popular, but considering the current state of the pandemic, it is difficult to confirm that these models work accurately in reality. A DNN was designed for the prediction of the behavior of engineered RNA elements which are capable of detecting small molecules, proteins, and nucleic acids.⁵³ This work shows that DNN approaches could be used for predictions in RNA synthetic biology, but more data are needed for the training of DNNs, as well as improvement of DNN architectures in order to improve predictions. A DNN performed an automatic diagnosis of the 12-lead ECG recordings and outperformed cardiology residents in recognizing six types of abnormalities, with F1 scores above 0.80 and specificity over 0.99.⁵⁴ Additional studies could test whether DNN effectively diagnose different ECG abnormalities, including myocardial infarction. DNN classifiers were trained to predict the risk of developing

coronary heart disease (CHD) on a dataset with 25,990 patients (accuracy ≈ 0.86).⁵⁵ The authors used the reconstruction error-based feature selection and claimed that by using two DNN classifiers, they improved the performance of the model. DNN, RF, and a simple statistical test were used to predict COVID-19 infections from full blood counts only (598 samples), without knowing the history of the patients (accuracy up to 0.91).⁵⁶ It appears that these DNN models are more of theoretical work at this stage of development, without evident clinical implementation.

Convolutional Neural Networks. CNN networks use a special kind of linear mathematical operation called convolution. The hidden layers of a CNN typically consist of a series of convolutional layers.²² Deep CNNs show great potential for melanoma subtypes and localization diagnosis on dermoscopic image datasets (780 images) and achieved AUROC ≈ 0.93 .⁵⁷ Improvements in the accuracy of this model could be achieved by adding more training images of mucosal and subungual sites. Data-augmentation deep models (DADLM) that enhance the learnability of CNNs and Convolutional Long Short-Term Memory (ConvLSTM) deep learning models, improve the accuracy of COVID-19 detection.⁵⁸ The study used 50 images (X-ray and CT) and presented two data augmentation techniques based on simple image transformations. This model needs more reliable data to confirm its performance. The fast-track COVID-19 classification network (FCONet) was developed to diagnose COVID-19 pneumonia in CT images (3993) and differentiate it from non-COVID-19 pneumonia and non-pneumonia diseases with ≈ 0.99 accuracy.⁵⁹ A CNN was also utilized to classify solid, lipid-poor, contrast enhancing renal masses using enhanced CT images (143 patients) with the accuracy ≈ 0.99 , and AUROC

≈ 0.82 ,⁶⁰ and for automated prediction of breast cancer risk (92 histopathological images, F1 score 0.73).⁶¹ These CNN models are examples of deep learning that rely on medical imaging. More research that uses EHR, or medical text data in addition to imaging, could contribute to faster and cheaper diagnostics. Deep CNN has shown an accurate gene prediction in metagenomics fragments with an accuracy of 0.91 (seven million training samples).⁶² The potential limitation is whether CNN-based models will be able to identify correct features when sequence errors are present. Deep CNN has shown great success in decoding motor preparation of upper limbs from time-frequency maps of EEG signals.⁶³ A deep learning architecture was applied to early diagnosis of glaucoma (301 images, AUROC 0.92),⁶⁴ and for early diabetic retinopathy detection,⁶⁵ on retinal fundus images (40 images, AUROC 0.94). Further studies with larger datasets, adding post-processing methods, and improved optimized deep ML architectures could increase the accuracy of these models.

Recurrent Neural Networks. RNNs, are a type of neural networks that allow an analysis of temporal heterogenous medical data. LSTM or GRU units effectively model the irregular visiting patterns in the long sequence of events in EHR. They can process input of any length, while a model size does not increase with the size of the input.²³ RNN models are good choices in scenarios with temporal heterogenous medical data, because of LSTM or GRU units that effectively model the irregular visiting patterns in the long sequence of events in EHR, and effectively model nonlinear relationships which exist in EHR. RNNs and the magnetic induction system were integrated to detect a wide range of human motions.⁶⁶ The benefit of LSTM RNN for sequence classification is the ability to support

multiple parallel temporal input data from different sensor modalities.⁶⁶ LSTM and GRU RNN models were developed to predict complications of DM2 (two million patients with DM2 diagnosis), with the prediction accuracy up to 0.84. They outperformed traditional ML models in prediction accuracy (up to 0.76) of 10 selected complications of DM2.⁶⁷ An RNN approach was used for predicting hemoglobin levels in patients with end-stage renal disease (7,739 patients) and produced Mean Absolute Error (MAE) of 0.54.⁶⁸ Further research is needed to incorporate the dialysis and laboratory information. RNNs were designed for monitoring of depth of anesthesia based on features of EEG signals (20 patients).⁶⁹ RNN models were utilized for predicting onset of sepsis on 30,000 samples and achieved AUROC 0.81.⁷⁰ LSTM RNN models predicted AD from conditions, measurement and drugs domain, on about 2,600 patients. A successful application of the drugs domain in the prediction of AD was presented (area under the precision recall curve (AUPRC) 0.99).⁷¹ Additional research with the drugs domain is required to develop comprehensive clinically applicable ML solutions. Researchers leveraged a big dataset to build an RNN to predict trends of future glucose levels and adverse glycemc events in DM1 (27,466 samples, MAE 0.05).⁷² Utilizing the historical information in EHR, an RNN LSTM method was designed to predict heart failure (365,446 patients, AUROC 0.89).⁷³ The performance of the presented RNN models could be improved by training them on better quality hospital datasets and further optimization of deep learning models. We summarized the reviewed ML models in table 1.2.

| ML Method | Author and year | Topic |
|---|------------------------------|--|
| DT | Soguero-Ruiz C, et al 2020 | Diabetes Mellitus, Hypertension |
| DT, SVM | Joloudari JH, et al 2020 | Coronary artery disease |
| DT, K-NN, SVM, RF, Bayes | Kuo KM, et al 2019 | Pneumonia – hospital acquired |
| DT | De Felice F, et al 2020 | Rectal cancer |
| RF | Chavan S, et al 2020 | Liver toxicity |
| RF | Tracy BM, et al 2020 | Hernia repair |
| RF | Mar J, et al 2020 | Dementia, Neuropsychiatric symptoms |
| RF, AdaBoost | Iwendi C, et al 2020 | COVID-19 |
| Ensemble, bagging, boosting, RF, AdaBoost | Yaman E, et al 2019 | Neuromuscular disorders |
| Ensemble, bagging | Nikolaidis A, et al 2020 | Functional parcellation of the human brain |
| Ensemble, AdaBoost | Saravanakumar S, et al 2019 | Alzheimer’s disease |
| Ensemble, AdaBoost | Li S, et al 2020 | Colorectal cancer |
| MLP | Car Z, et al 2020 | COVID-19 |
| MLP | Leithner D, et al 2020 | Breast cancer |
| Bayesian | Madhukar NS, et al 2019 | Drug target identification |
| Bayesian | Gupta A, et al 2019 | Coronary artery disease |
| SVM, DNN | Varghese J, et al 2020 | Movement Disorders |
| SVM | Satoh Y, et al 2020 | Breast cancer |
| SVM | Lown M, et al 2020 | Atrial fibrillation |
| K-NN | Jensen C, et al 2019 | Prostate cancer |
| K-NN | Saçlı B, et al 2019 | Renal calculi, Kidney disease |
| LR | Verma V, et al 2020 | COVID-19 |
| Linear models, Bayesian | Levy JF, et al 2019 | Cystic Fibrosis |
| LR, Multiple regression | Ghosal S, et al 2020 | COVID-19 |
| LogR | Ma X, et al 2020 | COVID-19 |
| LogR | Tabaie A., et al 2020 | Aneurysmal Subarachnoid Hemorrhage |
| LogR | Long C, et al 2019 | Genetics, Risk factors |
| DNN | Deng Q 2020 | COVID-19 |
| DNN | Jung SY, Et al 2020 | COVID-19 |
| DNN | Angenent-Mari NM, et al 2020 | Genetics |
| DNN | Ribeiro AH, et al 2020 | ECG diagnosis, Heart diseases |
| DNN | Amarbayasgalan T, et al 2019 | Coronary heart disease |
| DNN, RF | Banerjee A, et al 2020 | COVID-19 |
| CNN | Winkler JK, et al 2020 | Melanoma |
| CNN, DNN, LSTM | Sedik A, et al 2020 | COVID-19 |
| CNN, DNN | Ko H, et al 2020 | Chest CT Image, Pneumonia |
| CNN, DNN | Oberai A, et al 2020 | Renal tumor, CT scan |
| CNN, DNN | Wetstein SC, et al 2020 | Breast tumors |
| CNN | Al-Ajlan A, et al 2020 | Gene prediction |
| CNN | Mammone N, et al 2020 | EEG signals, Motor upper limb |
| CNN | Muramatsu C. 2020 | Glaucoma, Retinal fundus images |
| CNN, DNN | Hatanaka Y. 2020 | Retinopathy |
| RNN | Golestani N, et al 2020 | Human activity recognition |
| RNN, RF, MLP | Ljubic B, et al 2020 | Diabetes mellitus, complications |
| RNN | Lobo B, | End-Stage Renal Disease, Hemoglobin |
| RNN | Li R, et al. 2020 | Monitoring Depth of Anesthesia |
| RNN | Scherpf M, et al 2020 | Sepsis |
| RNN LSTM | Ljubic B, et al 2020 | Alzheimer’s disease |
| RNN | Mosquera-Lopez et al 2020 | Diabetes mellitus type 1 |
| RNN | Maragatham G, et al 2019 | Heart failure |

Table 1.2. Reviewed articles, classified by ML model types. First authors, publication year, and medical topics described in the publications are presented.

The number of articles by types of medical conditions and topics is presented in table 1.3.

| Types of diseases by systems or therapy | Number of reviewed articles |
|--|------------------------------------|
| Nervous System and Sense Organs | 12 |
| COVID-19 | 9 |
| Neoplasms (Oncology) | 8 |
| Circulatory System | 8 |
| Genomics, proteomics, molecules | 3 |
| Endocrine, Nutritional and Metabolic | 3 |
| Respiratory System | 3 |
| Genitourinary System | 2 |
| Infectious and Parasitic Diseases (w/o COVID-19) | 1 |
| Digestive System | 1 |
| Mental Disorders | 1 |
| Surgical procedures | 1 |
| New drugs, drug therapy | 1 |

Table 1.3. The number of analyzed articles by types of medical conditions.

1.3. Interpretation

This review shows that almost all categories of diseases were subject to ML analysis. Most ML models in Medicine represent good software solutions with high prediction accuracy, but only handful of models could find an implementation in medical practice. Neurological conditions are the most common medical system subject to ML model applications.^{30,32--34,40,49,63-66,69,71} The most frequent type of data used in these applications were imaging data. Images consist of spatially coherent pixels in a local region, meaning that pixels close to each other share similar information. Deep learning architectures (especially CNN) produce higher accuracy predictions from image inputs than from EHR type of datasets, which are often heterogeneous. Another medical discipline extensively used in ML analysis is oncology. Accurately predicting the development of cancers or complications of cancers could indicate earlier diagnosis and therapeutic approaches that would improve

outcomes.^{27,35,37,41,43,57,60,61} Majority of these ML applications use imaging data (most often histologic type) for classification of malignant versus benign tumors. Cardiovascular conditions and DM are among the most common medical conditions used in predictive analysis.^{24,25,39,42,49,54,55,67,72,73} The challenges with this type of predictions are often related to limitations of data availability. Insurance claims data was frequently used but often lack important clinical information such as laboratory results and medications.^{67,74}

COVID-19 disease is one of the most researched medical conditions used for ML predictions recently. Many traditional and deep ML models were utilized with the goal to help to detect COVID-19 infections, complications, or outcomes.^{31,36,45,47,48,51,52,56,58,59}

The performance of predictive ML models in medicine depends on multiple factors. For challenging prediction problems, the understanding of the disease is likely to lead to more accurate prediction. Physicians have to be better motivated to use ML developments, which is not always easy to achieve since they perceive this activity as something that decreases their time with patients.¹ Since, many physicians use computers daily, better presented benefits of ML prediction models could increase their adoption in medicine. Evidence-based medicine requires statistical analysis of medical data, and ML is a form of that analysis. Some form of ML should become a part of statistics teaching in medical school, to prepare future physicians for meaningful adoption of medical ML models. To make ML more meaningful in clinical practice, we should focus on tasks that physicians need help with, and where the results of ML could help physicians to improve their decisions. The computers that physicians use for EHR could also be used for ML models. Additionally, ML is relatively inexpensive compared to basic science and large-scale clinical research.

Traditional ML methods do not always achieve accuracy that would convince medical doctors of the benefits of the proposed predictive models. Prediction accuracy of 70-90% is generally a good result in terms of performances of ML models but may not be high enough to suggest clinically meaningful improvements in practice. Traditional models have the advantage of simplicity and interpretability but suffer from somewhat worse accuracy.^{40,56,67}

The most successful and meaningful application of deep learning ML models was achieved in the imaging field.^{54,57-61,64,65} Analyses of CT scans, X-rays, Doppler ultrasound, histopathological images obtained high accuracy results, which often outperform medical experts. RNN models capture the temporal nature of EHR, imaging and other medical data to predict diseases, complications, and outcomes.⁶⁶⁻⁷³ Deep learning models produce higher accuracy but suffer from issues of interpretability and instability.^{15,75} We certainly need to consider the interpretability of deep ML models to make them efficiently applicable in medical practice.⁷¹ Combinations of traditional and deep learning models could address challenges of interpretability and accuracy.^{40,56,67} Many datasets are small and do not have enough samples for the implementation of deep learning models. In those cases, traditional ML models are the only option.

To build effective ML models, we have to understand how to select relevant features to train ML models. Computational methods that use optimization function to automatically select useful features have been developed.^{76,77} In addition to automatically selected features, we often have to use medical domain knowledge to identify useful features that could help in the improvement of predictions.^{67,71} It's not often obvious to determine which

of the input features contributed to the achieved accuracy. If analyses point toward certain features as the most important for obtaining the model performance, the next challenge is how to quantify the relevance of those features.

An important dilemma is the selection of the most optimal ML models for a specific task. We have to consider which specific medical problem we want to solve. We need to determine how much and what type of data are available, how much data are missing, do we have temporal information included. Implementation of ML in prediction of medical conditions using EHRs and other non-imaging data as a cheaper source of data could achieve meaningful results at a lower cost. We would need to weigh positive and negative factors in each of the options before we select the most optimal model for the given task.

Predictive ML models could potentially help to build CDSS to make better medical decisions. These models can provide recommendations to select suitable patients for clinical trials.⁷¹ Domain knowledge and collaborations between physicians and ML experts can improve the prediction performance of ML models in medicine and facilitate implementation in clinical practice.

1.4. Conclusion

Prediction ML models could help clinicians to select the most effective diagnostic and therapeutic choices available. Successful ML models can make medicine more efficient, improve outcomes and decrease medical errors. We predict that ML models will continue to develop, and they will be applied more broadly in clinical practice.

In the following chapters, applications of advanced ML models on four selected diseases are presented. We analyzed advanced ML models on the examples of: Alzheimer's disease, Diabetes Mellitus type 2, Influenza, and Colorectal cancer.

CHAPTER 2

INFLUENCE OF DOMAIN KNOWLEDGE ON DEEP LEARNING FOR ALZHEIMER'S DISEASE PREDICTION

2.1. Introduction

According to the National Institute of Aging, more than 5.5 million Americans are diagnosed with AD.⁷⁸ Estimates indicate that AD may rank third as a cause of death for older people, only second to heart disease and cancer.^{79,80} Identification of individuals at risk for developing AD is imperative for testing therapeutic interventions.⁸¹ Many researchers have presented an overview of the classification of Mild Cognitive Impairment (MCI).⁸² Early diagnosis could help with the recruitment of patients to participate in clinical trials and the testing of possible new drug therapies for AD. Several studies indicate that the use of imaging for early detection of AD is imperative to early diagnosis.⁸³ Diagnostic imaging can be very costly and the question is what is the cost-effectiveness of imaging in AD.⁸⁴

Heterogeneous structures of Electronic Medical Records (EMR) data pose a challenge for machine learning (ML) algorithms.⁸⁵⁻⁸⁷ For this study, we used a repository of heterogeneous ambulatory EMR data, collected from primary care medical offices spread

over the U.S. These data are typical of ambulatory EMR data from medical practices and not data from clinical trials. The data for this study was sourced from IQVIA and EMR vendors which were then mapped into the Observational Medical Outcomes Partnership (OMOP) format.

ML algorithms utilizing Magnetic Resonance Imaging (MRI) for prediction of AD have been developed.⁸⁸⁻⁹⁰ An application of RNN models was designed to differentiate AD patients from healthy control individuals using neuroimaging data.⁹¹ Longitudinal EMRs were used to study the progression of chronic diseases like AD.^{92,93} LSTM RNN can effectively predict AD progression by fully leveraging the temporal and medical patterns derived from patient's office visits.

Our research goal was to implement LSTM RNN deep learning configuration to predict AD diagnosis using EMR data alone (without relying on diagnostic imaging).^{94,95} The study objective was to show that selection of relevant input datasets is important for overall LSTM RNN model predictive performance. We wanted to determine if applying medical domain knowledge in data preprocessing and positive dataset selection significantly improves the prediction of AD comparing to the naïve model. The most current health problem with Coronavirus (COVID19) which severely affected the entire world, proves the importance of relevant medical data in the creation of analytical and predictive models as well as in the application of adequate preventive health measures.⁹⁶

Furthermore, we attempted to efficiently apply the drugs domain in prediction of AD. The objective was also to evaluate the contribution of individual clinical domains as well as the ensemble of few domains to the prediction of AD.

An accurate early prediction of AD could help with the recruitment of patients for clinical trials, which could help find new drug therapies for AD. Also, the contribution of predictions of AD is a better selection of patients who need imaging diagnostics for differential diagnosis of AD.

2.2. Methods

Our comprehensive methodology comprises relevant inputs selection using medical domain knowledge, and construction of the LSTM RNN deep learning model. We used the ambulatory EMR database to predict the occurrence of AD. Data are in the OMOP common data model (CDM) format. The terminologies used to describe the clinical conditions vary from database to database. In the OMOP concept, data contained in different types of observational databases are transformed into a common format. The OMOP CDM provides a common data standard to analyze multiple data sources concurrently.⁹⁷ The OMOP concept allows the evaluation of individual clinical domains separately as well as an ensemble of different domains.

Patients with AD diagnosis in the EMR database were extracted using the OMOP code for AD. We found 24,734 AD patients in the EMR database. This dataset does not contain patients treated with experimental therapies for dementia or AD. It contains only medications that are approved by the FDA and available for everyday use. The measurement domain includes labs measured in ambulatory settings such as blood tests and vital signs. It does not contain data obtained in clinical settings such as biological results about CSF biomarkers related to tau and amyloid. The condition domain data

contain only basic diagnoses, not including specific cognitive measurements such as episodic memory deficits according to psychometric tests. We did not use any imaging data.

In the initial experiment, we built positive datasets by setting certain filters, calling this approach *naïve*. We filtered out conditions (diagnosis) that appeared less than 30 times in the whole dataset which are rare diseases that have small information value and represent noise. We also filtered out patients who had less than four visits. The number of visits was set to four or more to capture the temporal nature of patients' histories.⁶⁷

The negative dataset was selected randomly from the entire ambulatory EMR database and it consists of patients who had 4 visits, who didn't develop AD, and who were born before 1950. By selecting this age group, we made the age distribution of patients in the negative dataset almost identical to the age distribution of patients in positive datasets. The average age in positive datasets when AD was diagnosed was 80.2. About 51% of patients in positive datasets were older than 80 at the time of AD diagnosis. The interval between the last visit before AD and the visit when AD was diagnosed varies between a few days and many years.

Further, we applied medical domain knowledge to build positive datasets. Development of AD could roughly be classified into three stages: preclinical stage, MCI stage, and clinical AD stage. Since the OMOP vocabulary does not have a hierarchical organization like ICD coding, we had to search for terms that match ICD codes for MCI. We included the following conditions from OMOP vocabulary to define the pre-AD stage: Mild Cognitive

Impairment, Memory impairment, Organic mental disorder, Amnesia, Forgetful, Cognitive disorder. The final dataset represents the union of the above-mentioned conditions.

Initially, in the naïve setup, we had cases that some patients had for example: knee injury, common cold, and flu in their medical history, and the naïve model was predicting the occurrence of AD. We were instructed by clinicians that data should have some relevance to AD to be accepted by medical experts.

We applied medical domain knowledge and constructed datasets with MCI stage included, to avoid prediction of AD only from unrelated diseases. All selected patients had the MCI condition in the pre-AD stage. None of the visits designated to the pre-AD stage contained the OMOP code for AD diagnosis. We selected all data over a continuous period before the first AD visit date in our positive datasets. The first AD visit date is the date when the patient was diagnosed with AD for the first time. We didn't include visits after the first AD diagnosis, to avoid data leak. Experiments were conducted using the following three domains: conditions, measurements, and drugs. We selected patients who had at least 4 visits and data in each of the three domains to be able to compare prediction results for AD among these domains. This preprocessing resulted in the final relevant dataset of 2,324 patients. We will call this dataset **Selected Clinically Relevant Positive (SCRIP)** dataset.

One SCRIP positive dataset was constructed for each of the domains.

Experiments with patients who had data in different combinations of two domains were also conducted. We present datasets used in experiments in Table 2.1.

| Dataset | Number of patients |
|--|---------------------------|
| Naïve model, patients with 4 visits, all studied domains | 2,600 |
| SCRP model, 4 visits, all studied domains | 2,324 |
| SCRP model, 3 visits, all studied domains | 3,199 |
| SCRP model, 2 visits, all studied domains | 3,568 |
| SCRP, 4 visits, conditions and drugs domains | 3,726 |
| SCRP, 4 visits, conditions and measurements domains | 3,846 |
| SCRP, 4 visits, conditions domain | 6,418 |

Table 2.1. Datasets used in the experiments and the number of patients in each of them

We also conducted experiments by excluding particular conditions that were used to define the pre-AD stage and making different combinations of relevant conditions.

We developed a model in Python (PyTorch library) that uses LSTM RNN to model a sequence of medical codes and their temporal associations. RNN LSTM contains hidden units that can analyze sequences of events, like EMR. RNN LSTM is suitable for longitudinal datasets with different sizes of time intervals between events which is the case in our application.

The created model consists of an input layer, an embedding layer, a sequence modeling layers, and an output layer (Figure 2.1.).

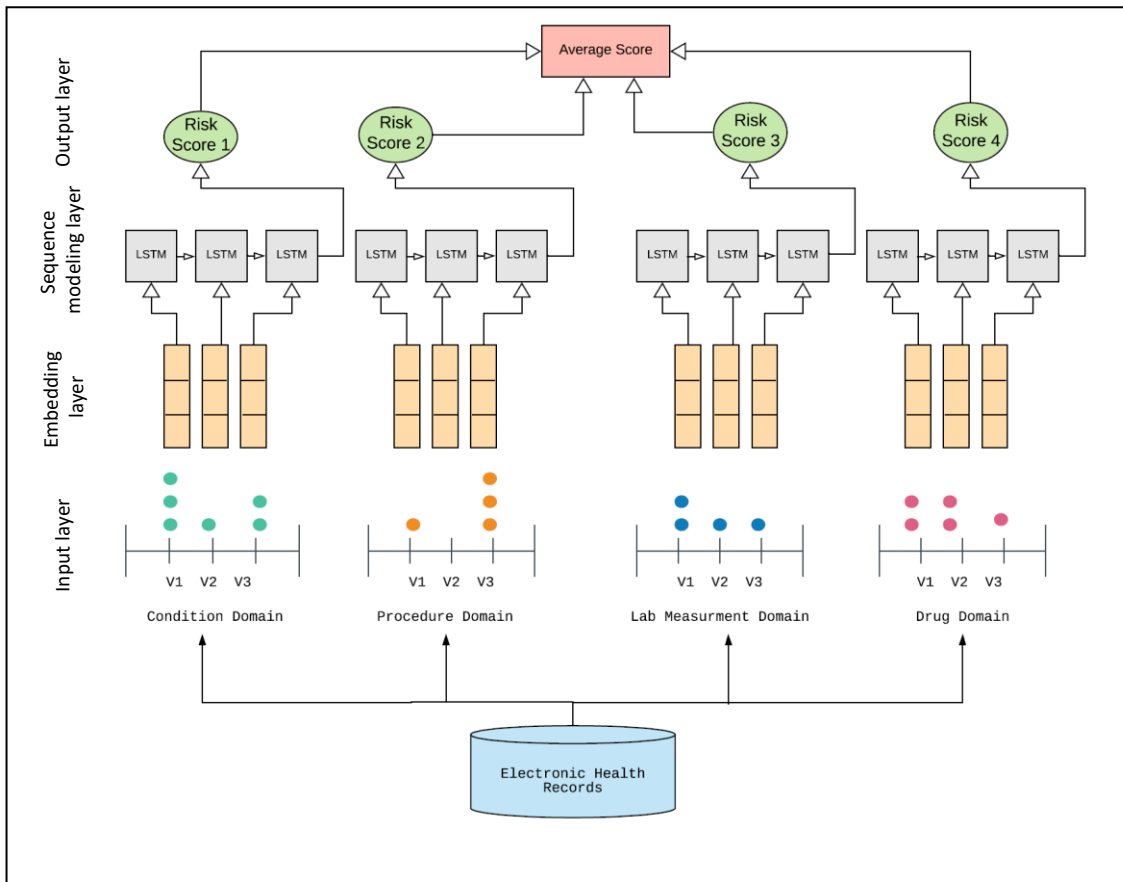


Figure 2.1. LSTM RNN deep learning system designed in this research: an input layer (patient timeline), an embedding layer, a sequence modeling layer, and an output layer (disease risk score).

We generated and presented a sequence of visits for each patient to the input layer. Initially, a dictionary was created to map all distinct OMOP codes into indices (integers). Each visit was then represented with a *one-hot (feature) vector* which contains a value of one at the indices corresponding to the OMOP codes from that visit. Note that the dimensionality of the visit vectors depends on the domain that is being considered for learning (for instance, if the patients' sequences of visits are extracted from the *condition domain*, the dimension

of the visit vectors will equal the number of distinct OMOP *condition codes*). The total number of different OMOP codes for conditions (diagnoses) in the starting dataset was 7880. After applying filters in the final naïve setup, the number of different OMOP codes for conditions was 1445, for measurements 782, procedures 545, and for drugs 828. In the SCRP setup, after applying all filters, the number of different OMOP codes for conditions was 1325, for measurements 752, procedures 495, and for drugs 145.

After processing, patients' records were in the form of sequences of one-hot encoded visit vectors. In the embedding layer, the bags of indices were converted to bags of embedding. The next layer was an RNN sequence modeling layer with LSTM cells and the last layer was the output layer.

First, we conducted all experiments using the naïve positive datasets and then all experiments using the SCRP datasets. We performed experiments with different training : testing ratios with both types of positive datasets. We randomly selected 90% of patients for training and the remaining 10% of patients for testing (234 patients in the SCRP dataset, and 260 patients in the case of the naïve dataset). Furthermore, in the next group of experiments, we kept the size of the testing dataset fixed at 10%. We were decreasing the size of the training dataset, from the starting 2,090 patients to 1,800, then to 1,500, 1,250, and 1,000 patients to identify at what size of the dataset the performance of our model starts decreasing. We also performed experiments with different split ratios of training and testing datasets. We used 80:20, as well as 70:30 ratio in the random split for training and testing datasets.

In the training phase, the AD positive to negative ratio was 1:2 (majority sub-sampling) and in the testing phase, the AD positive to negative ratio was 1:9 in experiments with the naïve dataset and between 1:5 and 1:25 in experiments with SCRP datasets. The ratio 1:9 positive to negative corresponds to the percentage of AD in the general population (10-15%). In SCRP datasets we used ratios between 5-25% in the testing phase to simulate different possible subpopulations. Negative examples that complemented positive datasets, were selected randomly from the preprocessed dataset of patients, that we described earlier. For each domain, we had at least 20 trials. In all experiments number of hidden LSTM units was set to 100, the number of epochs was 100 and the batch size was 100. We used Adam optimizer.

The objective was to predict whether the patient will be diagnosed with AD at the next visit. The evaluation metric was Area Under the Precision-Recall Curve (AUPRC). A precision-recall curve is a plot of the precision (y-axis) and the recall (x-axis) for different thresholds. We used the Softmax function to output a risk score for AD.

We ran our LSTM RNN code separately for each of the three domains. In the end, we combined the results of each domain into an ensemble model (average of the outputs from best performing domains).

2.3. Results

In the *naïve* model, after applying filters described in Section 2, 2,600 patients with at least 4 visits were in the positive dataset. Risk score predictions in the naïve model in the form of AUPRCs for models trained separately for two domains (conditions, measurements) are

shown in Figure 2.2. The best results were obtained with the measurements (labs) domain. An ensemble of two domains led to the improvement of predictions of risk scores of AD development. Drugs and procedures domains had low accuracy in the naïve model. The precision/recall comparison of models using condition, measurement, and an ensemble of two domains (c.m.) is presented in Figure 2.3.

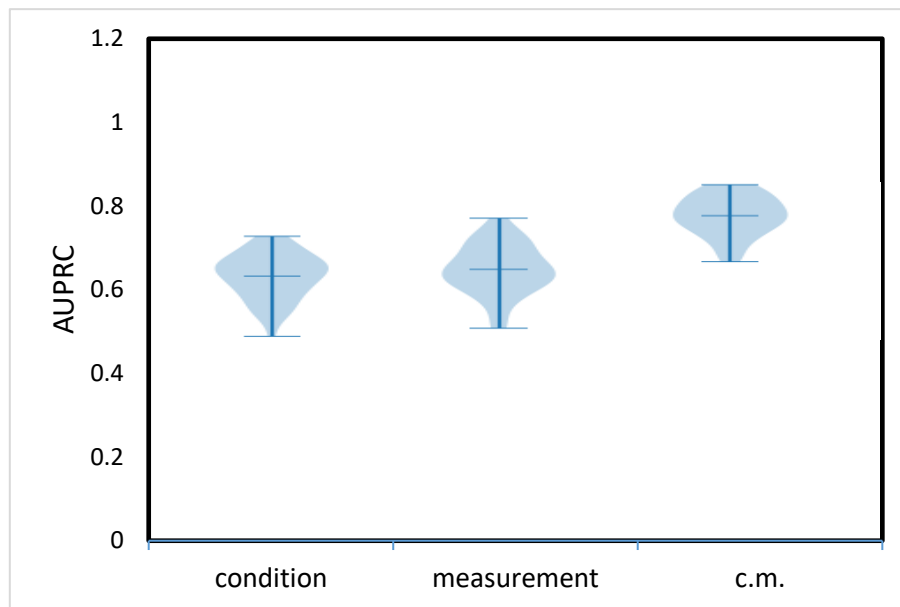


Figure 2.2. The naïve model results, AUPRC score obtained by LSTM RNN for condition and measurement domains separately and for their ensemble (c.m.) when using the naïve AD dataset selection for patients with at least 4 visits.

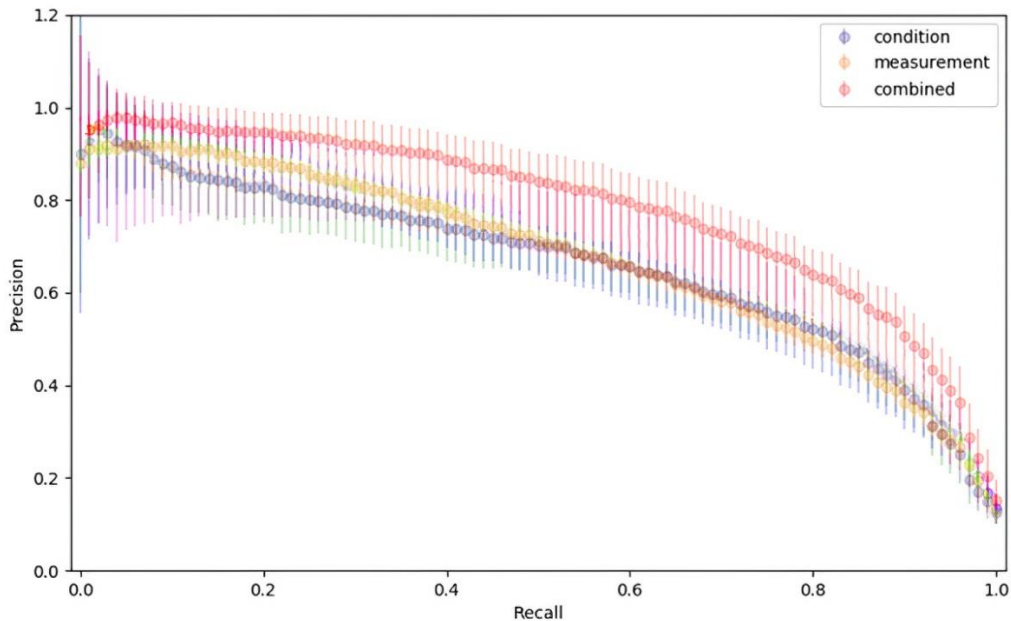


Figure 2.3. The average precision/recall comparison of models using condition, measurement, and an ensemble of two domains (combined).

The final **SCRIP dataset** with all three domains and at least 4 visits, resulted in 2,324 patients. We ran the LSTM RNN model using the condition, measurement, and drug domains separately. Initially, we used 10% of data for testing and 90% for training of our model. AD prediction based on the condition domain was similar to the naïve model. However, the measurement (labs) domain produced significantly better results than the naïve model (AUPRC 0.98-0.99).

We achieved a successful application of the drugs domain in our model. Selected drugs (145 distinct drugs) were used to treat different conditions, considered as pre-AD, and they were administered before patients had been diagnosed with AD. These were standard drugs already used in everyday medical practices, and not experimental drugs. We excluded

Acetylcholinesterase Inhibitors from this list. Results of AD prediction using only the drugs domain achieved AUPRC of 0.98-0.99. The selection of relevant drugs that are given to treat one of pre-AD conditions and symptoms is the factor that provided the difference in predictions of AD, comparing to the naïve dataset. This approach decreased the number of considered drugs from 828 (*naïve* model) to 145 in the *SCRIP* model, which reduced noise and allowed improved prediction results.

After we developed the LSTM RNN model for each of domains separately, we joined results into an ensemble model where outputs of three single domain-based models were averaged. Results of the ensemble prediction of all three domains achieved out-of-sample AUPRC above 0.99.

For *SCRIP* dataset experiments, in Figure 2.4. we present results of risk scores for the development of AD for each of the studied domains and for the ensemble model (c.m.d) using information from all three domains. The procedure domain had low accuracy and we will not show those results.

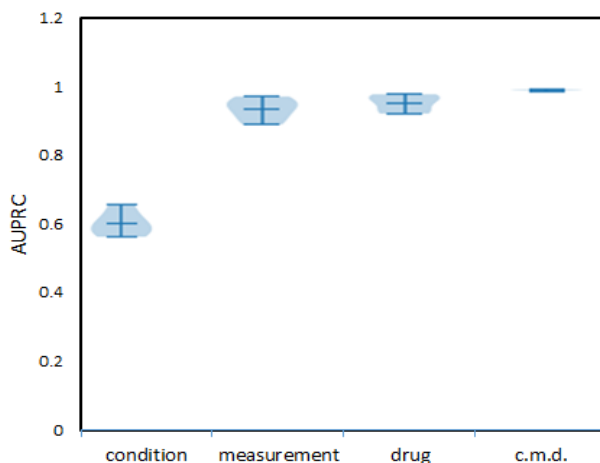


Figure 2.4. AUPRCs of AD predictions by LSTM RNN trained on the *SCRIP* dataset using each of three domains (condition, drug, measurement) separately and as an ensemble (c.m.d).

Figure 2.5. shows Precision-Recall curves for three domains and for the ensemble of three domains for the SCRP dataset. The ensemble of three domains (c.m.d.) achieved the best AD prediction results.

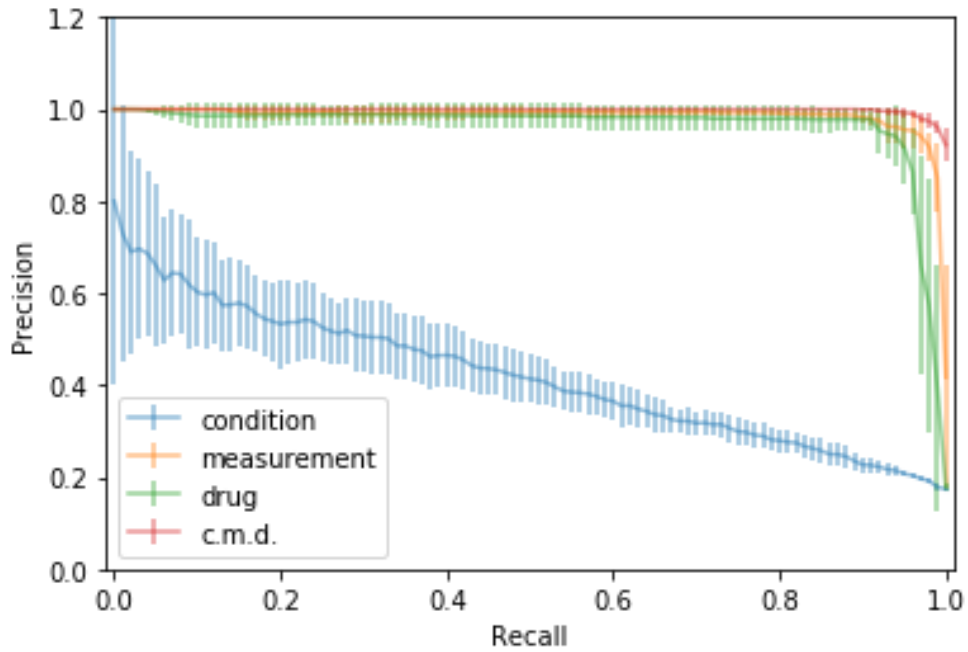


Figure 2.5. The average precision/recall comparison of LSTM RNN models trained on the SCRP dataset using condition, measurement, and drugs domains and ensemble information of all three domains (c.m.d.)

Changes of AUPRC scores when different ratios of training and testing datasets were used in experiments are shown in Table 2.2. We kept the fixed size of the testing set and slowly decreased the size of the training set.

| No. of patients in training dataset | No. of patients in testing dataset | AUPRC Drugs | AUPRC Measurements | AUPRC Conditions | AUPRC c.m.d. |
|-------------------------------------|------------------------------------|-------------------|--------------------|-------------------|-------------------|
| 2,090 | 234 | 0.985 ± 0.040 | 0.986 ± 0.048 | 0.651 ± 0.031 | 0.991 ± 0.038 |
| 1,800 | 234 | 0.982 ± 0.048 | 0.985 ± 0.050 | 0.650 ± 0.040 | 0.990 ± 0.033 |
| 1,500 | 234 | 0.869 ± 0.053 | 0.889 ± 0.060 | 0.648 ± 0.043 | 0.908 ± 0.035 |
| 1,250 | 234 | 0.860 ± 0.049 | 0.866 ± 0.052 | 0.647 ± 0.037 | 0.871 ± 0.042 |
| 1,000 | 234 | 0.810 ± 0.058 | 0.814 ± 0.059 | 0.645 ± 0.052 | 0.835 ± 0.055 |

Table 2.2. Prediction of AD by LSTM RNN trained on the SCRIP dataset using the drugs, the measurements, the condition domain, and their ensemble (c.m.d.), with the fixed size of the testing dataset (234 patients) and different sizes of the training dataset. The evaluation metric - AUPRC

When training on 80% and testing on 20% of available data, results of risk score prediction of AD development were almost identical to the initial setup (90:10 split ratio) for all three domains as well as for the ensemble of domains. However, when using 70% of the SCRIP data for training and the remaining 30% of data for testing, a decrease of AUPRC for drugs and measurements domains was evident (AUPRC for the drugs domain was 0.87 and AUPRC for measurements was 0.9).

The influence of different splits of the dataset on training and testing sets is illustrated in Figure 2.6. for the drugs domain-specific LSTM RNN model. It appears that for each of the scenarios that we attempted when our model started dropping accuracy the critical size of the dataset for the training phase was about 1,500 patients.

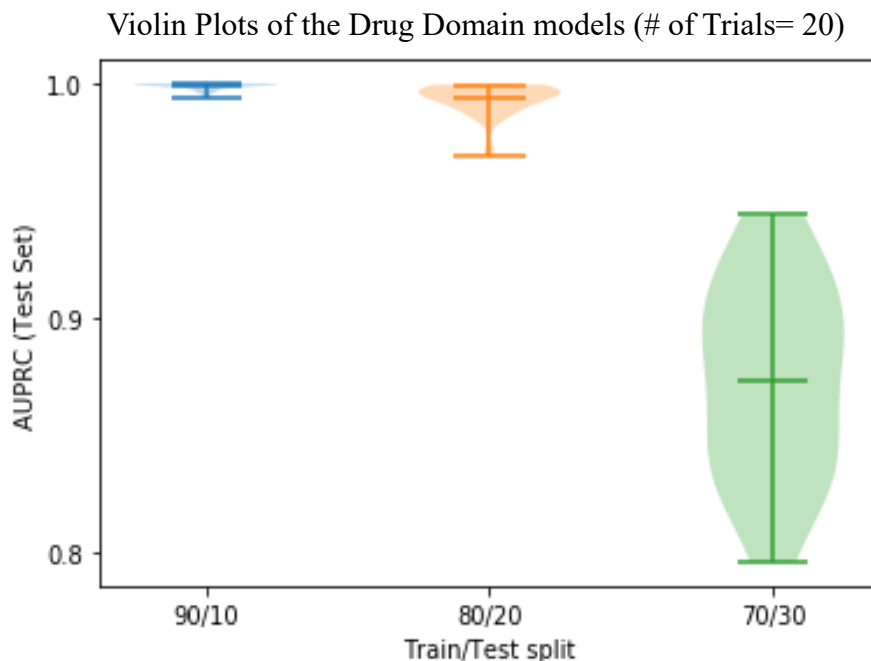


Figure 2.6. Prediction of AD by LSTM RNN model, trained on the SCRIP dataset using the drugs domain in different splits of the dataset for training and testing.

Experiments that estimated the effect of increased coverage on prediction quality as well as different numbers of visits (2, 3, or 4) are presented in Table 2.3.

| Datasets Positive cohorts | No. of patients | AUPRC Drugs | AUPRC Measurements | AUPRC Conditions | AUPRC c.m.d. |
|---------------------------|-----------------|----------------------|----------------------|----------------------|----------------------|
| 4 visits | 2,324 | 0.985 ± 0.040 | 0.986 ± 0.048 | 0.651 ± 0.031 | 0.991 ± 0.038 |
| 3 visits | 3,199 | 0.974 ± 0.042 | 0.973 ± 0.039 | 0.640 ± 0.032 | 0.982 ± 0.035 |
| 2 visits | 3,568 | 0.893 ± 0.046 | 0.908 ± 0.048 | 0.601 ± 0.030 | 0.924 ± 0.042 |

Table 2.3. Results of experiments with SCRIP datasets with different numbers of visits (2, 3, or 4) for drugs, measurements, and conditions domains as well as for an ensemble of all 3 domains (c.m.d.)

We conducted additional experiments by making different combinations of relevant conditions and tested the contribution of different pre-AD conditions to the accuracy results. In these experiments, the AURPC score dropped at least 10-15%, because the size of the training dataset was reduced to less than 1,500 AD positive patients which were not sufficiently large to conduct experiments with combinations of different relevant conditions.

2.4. Discussion

This research demonstrates that our comprehensive method which incorporates SCRP positive datasets based on the application of clinical domain knowledge and design of LSTM RNN deep learning models produced excellent prediction accuracy of AD. The way the inputs were selected proves to be important for the overall quality of the model.

We identified 125 papers on PubMed (MeSH thesaurus vocabulary search) directly describing the application of various ML techniques in diagnostics and prediction of AD. Aghili and collaborators used LSTM RNN models on longitudinal imaging data (MRI, PET) to predict AD and distinguish it from normal control cases.⁹¹ Albright studied whether serial MRI, PET, biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD and achieved prediction results of the model's average AUC=0.866.⁹⁰ Marzban and colleagues constructed CNN deep learning network using MRI images and achieved prediction results of the onset of AD (ROC-AUC between 0.84 – 0.94).⁹⁸ Arbabshirani and colleagues presented a review of more than 200 studies describing ML techniques in the prediction of

AD and other brain diseases, using imaging.⁹⁹ Moore designed the path signature model to predict AD, using imaging data and considering conversion to AD stage from healthy individuals or from individuals with MCI.¹⁰⁰ Martinez-Murcia and colleagues implemented data analysis of AD, based on deep convolutional autoencoders. The MRI imaging-derived markers could predict clinical variables with correlations above 0.6, achieving a classification accuracy of over 80% for the diagnosis of AD.¹⁰¹ MRI images were utilized for the construction of ML models for a diagnosis of MCI and AD, with achieved accuracy up to 75.51% (CNN networks). The prediction that patients with stable MCI will develop AD achieved ROC-AUC of 92.5%.¹⁰²⁻¹⁰⁵ Carpenter and Huang analyzed ML methods (naïve Bayes, kNN, SVM, random forest, and neural networks) for virtual screening (VS).¹⁰⁶ They presented a workflow for applying ML-based VS to the search for potential therapeutic drugs for AD. VS is important in the drug development process because it performs efficient searches over millions of compounds increasing chances for potential AD drug discovery. Huang and collaborators presented a nonlinear supervised sparse regression-based random forest framework to predict longitudinal AD clinical scores.¹⁰⁷ Ford et al, created ML models (logistic regression, naïve Bayes, SVM, random forest, and neural networks) to predict early dementia using EHR data. They included data on 93,120 patients, with a median age of 82.6 years, and achieved ROC-AUC of 74%.¹⁰⁸

We opted to use an LSTM RNN deep learning approach on heterogeneous EMR data alone (without relying on diagnostic imaging). None of the papers currently indexed in PubMed uses only EMR data for the prediction of AD. We hypothesized that for a large number of patients less costly and easier to obtain EMR data have sufficient representational power

to incorporate temporal heterogeneous information into the successful prediction of AD. This was a more difficult task than prediction from imaging data, because of the irregular temporal nature of EMR data. Furthermore, EMR data are cheaper and more abundant than expensive imaging MRI and PET data.

The increasing adoption of EMR systems has enabled significant opportunities for the development of ML algorithms in health care.^{86,108} EMR patient records provide a valuable resource for creating diagnostic support algorithms for the detection of dementia and AD.^{86,108} Tang and collaborators attempted to predict differential diagnoses of the top 25 most common conditions in the MIMC-III dataset.⁸⁶ They developed traditional and sequential (LSTM RNN) models, using ICD9 codes. The advantage of our model, comparing to Tang and collaborators is in the usage of the OMOP data concept which provides the ability to include more data from different sources and create ML models separately for different clinical domains and combine them into an ensemble model. Our approach allows the evaluation of performances of ML models on individual clinical domains as well as on combinations of different clinical domains, and subsequently the selection of the most optimal ML models. Our results suggest that special consideration is required to identify a medically relevant positive dataset as the optimal input to the LSTM RNN prediction model. The conditions domain was crucial in the selection of relevant information that contributed to the creation of datasets with patients who had MCI conditions before AD was diagnosed. The drugs domain was successfully applied in our model. Other researchers usually attempt to apply all drugs contained in datasets.⁸⁶ Tang and colleagues attempted to apply all drugs from the MIMIC dataset in their models and

they did not achieve good prediction results of diseases using this domain.⁸⁶ Most drugs are not relevant to diseases of interest and contribute to bad prediction as noise. Our research shows that the optimal approach with the drugs domain is to apply class-by-class of drugs and evaluate the contributions of each class and combinations of classes to predictions of diseases, which is different than the model described by Tang, where they attempted to apply all drugs.⁸⁶ Although we achieved good results with the drugs domain, the main contribution of this study is in the approach on how to use this domain. The measurements domain provided good predictions in all approaches. Considering the number of doctor's visits, RNN models achieved the best results when data include 4 ambulatory visits vs. relying on fewer visits data.⁶⁷

2.5. Conclusion

Significance of accurate and early prediction of AD could be found in the identification of patients for clinical trials, which can possibly result in the discovery of new drugs for the treatment of AD. Also, the contribution of predictions of AD is a better selection of patients who need some form of imaging diagnostics for AD. This research sets the framework for future analyses of disease-relevant temporal heterogeneous EMR data. Further research is necessary for the evaluation of different groups of drugs, measurements, and conditions and their contribution to the successful prediction of AD.

CHAPTER 3

PREDICTING COMPLICATIONS OF DIABETES MELLITUS USING ADVANCED MACHINE LEARNING ALGORITHMS

3.1. Objective

Diabetes Mellitus type 2 (DM2) is a chronic, metabolic disease and affects almost 100 million people all over the world including over 30 million in the US where it is the seventh leading cause of death.¹⁰⁹⁻¹¹¹ In the last 20 years, the number of adults diagnosed with DM2 has more than doubled.¹¹² The incidence of DM2 continues to rise and has quickly become one of the most prevalent and costly chronic diseases worldwide.¹¹³ Increased levels of glucose in the blood can cause many health complications over time.¹¹⁴ Management of DM2 requires a multidimensional approach.¹¹⁵⁻¹¹⁷ Identification of people at high risk of progression of DM2 enables targeted prevention.^{118,119}

Multiple computer science, especially machine learning (ML) applications have been developed to help with DM2 detection, management, prevention of progression, and improvement of patients' quality of life.¹²⁰ We designed deep and traditional ML models to predict development of complications in patients diagnosed with DM2. The Healthcare Cost and Utilization Project (HCUP) State inpatient databases (SID) Electronic Medical

Records (EMR) data for the period of 9 years were used for experiments. They contain diagnosis, procedures and time of patients' visits. We developed models based on a one-way Recurrent Neural Network Long Short-Term Memory (RNN LSTM) and bi-directional RNN Gated Recurrent Units (GRU) to capture the temporal nature of EMR data. Traditional models such as, Random Forest (RF) and Multilayer Perceptron (MLP) were used for comparison.

To evaluate prediction performance of different approaches we selected ten well described complications of DM2: Angina Pectoris (AP), Atherosclerosis, Ischemic Chronic Heart Disease (ICHHD), Depressive Disorder (DD), Diabetic Nephropathy, Diabetic Neuropathy, Diabetic Retinopathy (DR), Hearing Loss (HL), Myocardial Infarction (MI), and Peripheral Vascular Disease (PVD).

Following were the objectives of our study:

- Predict if these complications will develop along the course of DM2 (in our study within 9 years from DM2 diagnosis).
- Analyze how many hospitalizations between the diagnosis of DM2 and the diagnosis of each of ten complications were the most optimal for deep learning or traditional models to produce the best prediction accuracy.
- Test if deep learning RNN models are superior to traditional ML models in accuracy of predictions on the EMR heterogeneous temporal data.
- Analyze how the prediction accuracy of complications would change over time period of 9 years.

Timely and accurate prediction of complications could help with implementation of more specific and targeted measures, which would potentially prevent or slow down their development. Consequently, slowing down the development of complications would save significant economic resources needed for their treatment.

3.2. Background and significance

Patients with DM2 suffer many life-threatening complications including macrovascular like stroke, coronary artery disease, and/or microvascular complications: retinopathy, neuropathy, nephropathy and others. DM2 represents the most common etiology of extremity pain and diabetic neuropathy.^{121,122} Diabetic nephropathy continues to be a chronic and devastating complication of DM2.¹²³ Diabetes and depression occur together frequently.¹²⁴ DM2 appears to impair auditory function.¹²⁵ A close link exists between DM2 and cardiovascular diseases (CVD).¹²⁶ ML methods were proposed (SVM, RF, Linear Regression, Naive Bayes) to predict diabetic complications.¹²⁷ ML was used for forecasting future glucose fluctuations in the blood.¹²⁸ Deep learning LSTM neural networks and probabilistic modeling were designed for prediction of diabetes.^{129,130} ML models K-Nearest Neighbors, Naive Bayes, SVM, Decision Tree, Logistic Regression and RF were also proposed for prediction of onset of Diabetes.^{131,132} Clinical Risk Prediction with limited EMR and challenges of deep learning in Medicine were analyzed.^{133,134}

3.3. Materials and methods

We conducted experiments on hospital discharges data for 9 years (2003-2011) obtained from the HCUP, SID of California database. The studied dataset contains time of hospitalizations (visits) and ICD9 codes of diagnoses and procedures. The HCUP data were preprocessed and all patients with the diagnosis of DM2 were extracted (1,910,674 patients), using adequate SQL and Python queries. Original data were rearranged to create a table (matrix). Every row represents one patient (Pt). Each row contained a patient's hospitalizations in the order in which these visits occurred (Figure 3.1). Different colors in each row represent different hospitalizations. Within each hospitalization patients had one or more diagnoses (d) and sometimes procedures (p). Because the procedures domain did not produce good accuracy, we performed detailed analyses on the diagnoses domain only.

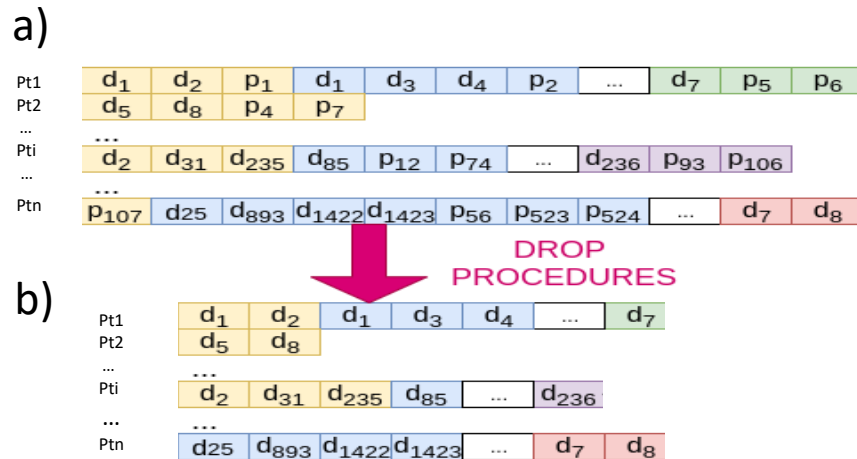


Figure 3.1. Each row (Fig. 3.1a) represents one patient (Pt). Different colors in each row represent different hospitalizations. Each hospitalization contained one or more diagnoses (d) and sometimes procedures (p). Since the procedures domain did not produce good results, we dropped them (Fig. 3.1b) and performed analyses on the diagnoses domain only.

In the first group of experiments patients who had the index complication diagnosed after at least 2 hospitalizations from the first DM2 diagnosis were extracted as the positive class. The same number of DM2 patients who didn't develop the index complication were randomly selected for the negative class, using a population based sample.¹³⁵

In the second group of experiments patients were selected for the positive class if the complication appeared after at least 3 visits from the first DM2 diagnosis. In the third group of experiments patients who developed the studied complication after at least 4 hospitalizations from the DM2 diagnosis were selected for the positive class. From each of these three datasets we randomly selected matching pairs (by minimum number of hospitalizations) of positive and negative cohorts. Thirty balanced datasets were created, one for each of the ten complications in each of the three groups of experimental settings. The average number of visits per patient was 4.07 with a standard deviation 5.08. All the hospitalizations starting from the hospital visit in which patients were diagnosed with the complication that we were predicting were excluded from the positive cohort, in order to avoid data leak. After this adjustment, the average number of hospitalizations for patients with a positive label was very similar to the average number of visits for patients with a negative label.

Diseases that appeared rarely or too frequently in the selected datasets don't contribute to the prediction as they don't have high informative value. All diseases which appeared more than 200,000 or less than 50 times among all the patients were deleted. After this preprocessing, 1,023 ICD9 disease codes were used to represent the patients' hospital visits. We applied Singular Value Decomposition (SVD) to reduce dimensionality of

visits.¹³⁶ This dimensionality reduction method uses matrix decomposition to transform features and select only features with the highest variance because those are the most informative characteristics.

Input to SVD was a matrix in which rows were all visits in a dataset and columns were all possible disease codes for that dataset. We have used a flat (one-hot encoded) representation of the “diagnoses”, and not a flat representation of “time”. Each of patients had at least few visits (rows in the matrix). Hospitalizations happened over the time period, with the maximum interval of 9 years. Although time is not specifically used as one of the features, the time component is reflected by the fact that consecutive visits were ordered by their timesteps. Each cell value in the matrix represented if a specific disease was present inside the visit. The value of each cell, therefore, was 0 or 1. Most of the cells had value 0 because only a few diseases appeared in each hospitalization. Output of SVD is a matrix which rows are patients, but columns are 50 features in a transformed space with the highest variance selected by SVD, which captures block correlations between data features.

Further, we created a matrix in which rows represented patients and columns were hospital visits ordered by timestamps. We deleted all patients who had more than 50 hospitalizations to reduce the size of a sparse matrix, but most of the patients had much less than 50 hospitalizations. If a patient had less than 50 hospitalizations, we padded their row with 0 to ensure that all rows have the same length. Then, we substituted each visit with a 50-feature vector from SVD and zero (non-existing visits) with a 0-vector of length 50. In other words, new matrix rows are patients containing concatenated feature vectors of that patients’ visits (each row has up to 50 hospitalizations * 50 features = 2,500). After

preprocessing and feature selection, the dimensions of data matrix were (Number of patients) X (Number of features) and this was the input for all the ML models that we tested in this study (RNN LSTM, RNN GRU, RF, MLP). The input for RNN (and other) models were all hospitalizations of all patients given to the model in chronological order, as the sequence ordered by timestamps for each patient.

Two types of ML models were utilized in this work: deep learning models and traditional models. The proposed deep learning models were one-way RNN LSTM and bi-directional RNN GRU (Figure 3.2).

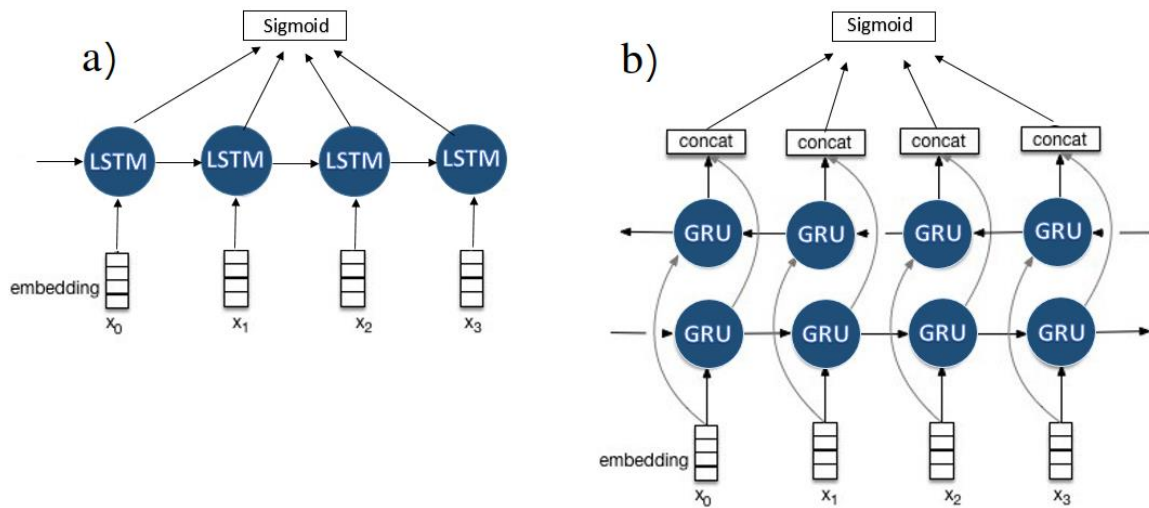


Figure 3.2. The proposed deep learning models: one-way RNN LSTM (Fig. 3.2a) and bi-directional RNN GRU (Fig. 3.2b)

RNN is a neural network where hidden neurons can analyze temporal sequential EMR data.¹³⁷ It has the same structure as the basic neural network, but neurons in the same layer are connected, allowing for neurons to learn information from its left neighbor in addition

to the current input. Therefore, RNN neurons have two sources of inputs, the present and the recent past. Learning process is described with following equations:

$$h^t = \text{relu}(b + Wh^{t-1} + Ux^t) \quad (1)$$

$$\bar{y} = \text{sigmoid}(b + \sum_t Vh^t) \quad (2)$$

To calculate value h^t of a hidden neuron t , a non-linear transformation, ReLU, is applied to weighted W value of its left hidden neuron h^{t-1} and the weighted U value of its input x^t . Prediction is calculated as a sigmoid function of weighted V sum of all hidden neurons with added bias b . Learning is achieved with back-propagation. Because of the long chain through which historical information (in forward direction) and gradient of error (in backward direction) had to be pass, RNN suffers from the vanishing gradient problem, which means that weights don't change and the model is not able to learn. To remedy this, a LSTM was invented, in which simple neurons of RNN are replaced with more complex short-term memory structure. LSTM shares the same weights across layers which reduces the number of parameters that the network has to deal with. The GRU is another solution for vanishing gradient. It substitutes the simple neuron with a gated unit which has fewer parameters than the LSTM neuron, because it lacks an output gate.¹³⁸

To model heterogeneous sequential data, we tested bi-directional GRU as a proposed method and compared it to 1-way LSTM. "Keras" Python libraries were used to implement constructed algorithms. We also compared GRU and LSTM to traditional machine learning algorithms. Our hypothesis was that both deep learning methods, RNN LSTM and RNN GRU will perform better than traditional ML models (RF and MLP) on medical temporal data like HCUP due to their ability to learn from a patient's history. In the proposed model,

we used ReLU and sigmoid activations. Also, we added a dropout between hidden and output layers, which randomly selects given percent of connections to cut. This is a well-known regularization technique that helps the model learn general pattern in data.

We used RF as a traditional model, because they have been shown as the state-of-the-art model in existing literature on predicting complications of diabetes and MLP because it is a simpler neural network model which doesn't account for time. Both were implemented with the "Scikit-learn" library in Python. The RF is a classification algorithm which consists of many decision trees.⁹ MLP is a network that consists of multiple layers of perceptrons and uses backpropagation learning. It uses a nonlinear activation function, which in addition to multiple layers distinguishes it from a linear perceptron.¹³⁹

We compared the performance of these four models in prediction of the occurrence of ten selected complications of DM2 with each of the three settings (2, 3 or 4 hospitalizations after DM2 diagnosis). The problem that we wanted to solve was a binary classification task. The evaluation metric was accuracy (3) computed as:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

where, TP is true positive, TN is true negative, FP is false positive, FN is false negative.

Further we tested what would be the minimum number of patients for RNN LSTM and GRU deep learning models to work properly and produce good prediction accuracy, after which the performance of these models starts to decrease. Two groups of experiments considering the minimum required number of patients for deep learning RNN models to

perform optimally, were conducted on DR data with at least 2 or 4 hospitalizations. We performed experiments with the entire datasets (58,641 patients with 2 and 25,468 patients with 4 hospitalizations) and then with randomly selected 5,000, 2,000, 1,000 and 500 patients in the positive cohort to discover the number of patients in the positive training cohort after which the accuracy starts dropping.

Furthermore, the prediction accuracy was evaluated for different time intervals between DM2 diagnosis and the first diagnosis of studied complications. We evaluated the accuracy of GRU RNN models for the intervals between DM2 and development of DR for: less than a year, 1 year, 2 years, 3 years, 4-5 years, and 6-8 years. Analyses were conducted on two experimental models: the complication developed after 2 or 4 hospitalizations from the initial DM2 diagnosis. Since, the entire dataset covers the period of 9 years, the maximum interval between DM2 diagnosis and the first diagnosis of DR identified in the dataset was 8 years.

Data in all experimental settings were split into 72% training, 8% validation and 20% testing. Cross-validation was used to find the best hyper-parameters values. For proposed RNN-like models we varied dropout and number of neurons in GRU/LSTM layer using random search. In the literature, dropout percent is usually between 0% and 50% and the number of units in the GRU/LSTM layer are usually selected among values 32, 64, 128, 256, and 512. For all RNN networks we used batch size 128, Adam optimizer, binary cross-entropy loss. We trained 20 epochs for RNN and tested on the epoch which had the best cross-validation accuracy. RF and MLP were trained with specific hyper-parameters as

well. In RF, maximum height of trees was bounded to 10 and number of trees was 100. MLP had 100 hidden units.

We repeated the same process for ten complications separately and we repeated all tests for datasets with a filter of the minimum of 2, 3 or 4 hospitalizations before the studied complication was diagnosed. We used t-test at the level of $p=0.05$ to check the significance of accuracy results that tested ML models produced in different experimental settings. Finally, probabilities that patients with DM2 will develop each of the studied complications (HCUP data) were calculated. Our codes are available on a public repository: <https://github.com/bljubic/diabetes-prediction>

3.4. Results

The total number of patients in the HCUP SID California dataset between 2003 and 2011, as well as the number of patients with DM2 diagnosis are shown in Table 3.1. We also present the number of patients with at least 4, 3 or 2 hospitalizations after DM2 diagnosis and before an index complication was diagnosed. These datasets were the source of data for positive and negative cohorts for all experiments.

| Dataset | Number of patients |
|--|--------------------|
| HCUP SID California (2003-2011) | 11,609,450 |
| Patients in HCUP with diagnosed DM2 | 1,910,674 |
| Patients with DM2 and 2 hospitalizations | 1,295,691 |
| Patients with DM2 and 3 hospitalizations | 930,837 |
| Patients with DM2 and 4 hospitalizations | 692,397 |

3.1. Datasets used in experiments and their sizes.

Experiments were performed separately for each of ten complications, and results for RNN deep learning models (Bi-directional GRU and 1-way LSTM) as well as traditional models (RF and MLP) are presented in Table 3.2. The evaluation metric is accuracy on out of sample data, and sizes of samples for each type of experiments are shown in the same table.

| Complication | No. of patients | Bi Direct. GRU | 1-way LSTM | RF | MLP |
|------------------------|-----------------|----------------------|----------------------|---------------|---------------|
| Angina Pectoris | | | | | |
| 4 visits | 19,589 | 0.796 ± 0.024 | 0.780 ± 0.016 | 0.717 ± 0.011 | 0.743 ± 0.013 |
| 3 visits | 26,973 | 0.789 ± 0.012 | 0.793 ± 0.019 | 0.722 ± 0.012 | 0.732 ± 0.008 |
| 2 visits | 42,459 | 0.738 ± 0.016 | 0.738 ± 0.018 | 0.701 ± 0.013 | 0.714 ± 0.009 |
| Atherosclerosis | | | | | |
| 4 visits | 32,914 | 0.756 ± 0.003 | 0.750 ± 0.015 | 0.712 ± 0.007 | 0.691 ± 0.008 |
| 3 visits | 44,688 | 0.750 ± 0.008 | 0.745 ± 0.012 | 0.704 ± 0.011 | 0.671 ± 0.008 |
| 2 visits | 62,016 | 0.713 ± 0.011 | 0.701 ± 0.018 | 0.689 ± 0.014 | 0.665 ± 0.012 |
| ICHD | | | | | |
| 4 visits | 52,959 | 0.835 ± 0.005 | 0.828 ± 0.008 | 0.759 ± 0.009 | 0.761 ± 0.017 |
| 3 visits | 81,658 | 0.814 ± 0.008 | 0.813 ± 0.007 | 0.745 ± 0.010 | 0.763 ± 0.015 |
| 2 visits | 147,718 | 0.802 ± 0.010 | 0.802 ± 0.015 | 0.744 ± 0.014 | 0.758 ± 0.015 |
| Depressive Dis. | | | | | |
| 4 visits | 56,343 | 0.820 ± 0.005 | 0.812 ± 0.008 | 0.714 ± 0.011 | 0.752 ± 0.013 |
| 3 visits | 78,732 | 0.802 ± 0.018 | 0.810 ± 0.004 | 0.739 ± 0.014 | 0.741 ± 0.015 |
| 2 visits | 135,492 | 0.773 ± 0.021 | 0.776 ± 0.019 | 0.722 ± 0.016 | 0.761 ± 0.016 |
| Hearing Impair. | | | | | |
| 4 visits | 8,576 | 0.734 ± 0.017 | 0.720 ± 0.021 | 0.691 ± 0.019 | 0.701 ± 0.021 |
| 3 visits | 12,030 | 0.743 ± 0.017 | 0.730 ± 0.020 | 0.694 ± 0.021 | 0.704 ± 0.019 |
| 2 visits | 16,884 | 0.716 ± 0.019 | 0.694 ± 0.022 | 0.680 ± 0.024 | 0.671 ± 0.023 |
| MI | | | | | |
| 4 visits | 38,380 | 0.733 ± 0.011 | 0.713 ± 0.013 | 0.691 ± 0.010 | 0.661 ± 0.016 |
| 3 visits | 52,896 | 0.723 ± 0.014 | 0.701 ± 0.012 | 0.688 ± 0.013 | 0.665 ± 0.014 |
| 2 visits | 92,961 | 0.711 ± 0.015 | 0.679 ± 0.013 | 0.663 ± 0.015 | 0.662 ± 0.017 |
| Nephropathy | | | | | |
| 4 visits | 37,982 | 0.768 ± 0.012 | 0.750 ± 0.014 | 0.699 ± 0.014 | 0.694 ± 0.024 |
| 3 visits | 52,283 | 0.766 ± 0.013 | 0.748 ± 0.013 | 0.696 ± 0.015 | 0.689 ± 0.020 |
| 2 visits | 71,053 | 0.742 ± 0.008 | 0.738 ± 0.010 | 0.695 ± 0.012 | 0.678 ± 0.015 |
| Neuropathy | | | | | |
| 4 visits | 49,060 | 0.746 ± 0.053 | 0.719 ± 0.073 | 0.671 ± 0.033 | 0.668 ± 0.039 |
| 3 visits | 69,053 | 0.738 ± 0.043 | 0.739 ± 0.068 | 0.664 ± 0.040 | 0.664 ± 0.046 |
| 2 visits | 99,825 | 0.715 ± 0.038 | 0.712 ± 0.054 | 0.660 ± 0.035 | 0.662 ± 0.055 |
| PVD | | | | | |
| 4 visits | 48,565 | 0.767 ± 0.002 | 0.744 ± 0.014 | 0.695 ± 0.006 | 0.691 ± 0.014 |

| | | | | | |
|--------------------|--------|-----------------------------|---------------|---------------|---------------|
| 3 visits | 67,686 | 0.759 ± 0.006 | 0.743 ± 0.010 | 0.708 ± 0.009 | 0.684 ± 0.010 |
| 2 visits | 93,905 | 0.738 ± 0.011 | 0.738 ± 0.014 | 0.701 ± 0.008 | 0.680 ± 0.012 |
| Retinopathy | | | | | |
| 4 visits | 27,796 | <i>0.796 ± 0.014</i> | 0.782 ± 0.001 | 0.741 ± 0.011 | 0.740 ± 0.007 |
| 3 visits | 36,221 | 0.752 ± 0.021 | 0.731 ± 0.013 | 0.698 ± 0.012 | 0.700 ± 0.011 |
| 2 visits | 58,641 | 0.728 ± 0.019 | 0.725 ± 0.014 | 0.696 ± 0.018 | 0.676 ± 0.012 |

Table 3.2. Presented are results of predicted accuracy (and standard deviation) that each of the ten complications of DM2 will develop within a 9 years period after the first DM2 diagnosis using HCUP EMR data (diagnoses domain). This period varies between 1 month and 9 years for individual patients. Results are presented for patients who had at least 2, 3 or 4 visits between the first DM2 diagnosis and before each of ten complications was diagnosed. The first column: names of complications; second column: number of patients in positive cohorts for each of complications; third column: Bi-directional GRU RNN classifier; fourth column: 1-way LSTM RNN classifier, fifth column: RF classifier, sixth column: MLP classifier. Bold are the best accuracy results for each experimental setting. Italic Bold are the best overall accuracy results for each of ten complications.

In Table 3.3, we present the accuracy, sensitivity, and specificity results for Bi-directional GRU RNN models in the 4-visits scenario for all complications of DM2, which was the model that achieved the best prediction accuracy. The results for 2 and 3 visits are omitted since they were consistent with results for 4 hospitalizations.

| Complication | Accuracy | Sensitivity | Specificity |
|------------------------|---------------|-------------|-------------|
| Angina Pectoris | 0.796 ± 0.024 | 0.862±0.019 | 0.698±0.014 |
| Atherosclerosis | 0.756 ± 0.003 | 0.791±0.012 | 0.718±0.014 |
| ICHD | 0.835 ± 0.005 | 0.886±0.014 | 0.787±0.012 |
| Depressive Dis. | 0.820 ± 0.005 | 0.848±0.009 | 0.792±0.010 |
| Hearing Impair | 0.734 ± 0.017 | 0.743±0.016 | 0.722±0.012 |
| MI | 0.733 ± 0.011 | 0.806±0.021 | 0.652±0.012 |
| Nephropathy | 0.768 ± 0.012 | 0.826±0.017 | 0.654±0.021 |
| Neuropathy | 0.746 ± 0.053 | 0.795±0.041 | 0.701±0.049 |
| PVD | 0.767 ± 0.002 | 0.774±0.005 | 0.753±0.011 |
| Retinopathy | 0.796 ± 0.014 | 0.799±0.007 | 0.792±0.018 |

Table 3.3. Accuracy, Sensitivity and Specificity for Bi-directional GRU RNN models in the 4-visits scenario for all 10 complications of DM2.

Different choices of hyper-parameters were tested. For bi-directional GRU RNN the best results were achieved with the dropout parameter value 0.2, and 128 hidden GRU neurons. We tried randomly dropout parameters between 0 and 0.5 and the number of hidden units 32, 64, 128, 256, and 512 in 20 experimental runs for each type of hyper-parameters. Accuracy results varied 2% in experiments for parameters selection. The best results for LSTM RNN model were achieved with the dropout parameter value 1.9 and 128 LSTM neurons. We present the average accuracy of 20 runs, including the standard deviation. We used the same set of hyper-parameters in experiments with 2, 3 or 4 hospital visits. Changes in the prediction accuracy of deep learning (RNN) models as well as traditional models when the size of positive training cohorts decreases are presented on the example of DR in Table 3.4. The performance of both deep learning RNN models deteriorates when the training dataset size decreases, especially when the number of patients in the positive training dataset drops below 1,000. The traditional models' performance vary slightly but does not change statistically significantly.

| No. of hospitalizations | No. of patients | Bi Direct. GRU | 1-way LSTM | RF | MLP |
|-------------------------|-----------------|----------------------|----------------------|----------------------|----------------------|
| 4 | 27,796 | 0.796 ± 0.014 | 0.782 ± 0.001 | 0.741 ± 0.011 | 0.740 ± 0.007 |
| 4 | 5,000 | 0.782 ± 0.012 | 0.776 ± 0.006 | 0.738 ± 0.010 | 0.747 ± 0.011 |
| 4 | 2,000 | 0.765 ± 0.011 | 0.743 ± 0.010 | 0.752 ± 0.014 | 0.766 ± 0.012 |
| 4 | 1,000 | 0.769 ± 0.014 | 0.767 ± 0.002 | 0.752 ± 0.009 | 0.742 ± 0.008 |
| 4 | 500 | 0.745 ± 0.013 | 0.745 ± 0.008 | 0.740 ± 0.013 | 0.750 ± 0.009 |
| 2 | 58,641 | 0.728 ± 0.019 | 0.725 ± 0.014 | 0.696 ± 0.018 | 0.676 ± 0.012 |
| 2 | 5,000 | 0.715 ± 0.014 | 0.706 ± 0.015 | 0.690 ± 0.021 | 0.662 ± 0.016 |
| 2 | 2,000 | 0.707 ± 0.015 | 0.707 ± 0.011 | 0.687 ± 0.019 | 0.660 ± 0.015 |
| 2 | 1,000 | 0.700 ± 0.019 | 0.685 ± 0.015 | 0.662 ± 0.012 | 0.657 ± 0.009 |
| 2 | 500 | 0.659 ± 0.018 | 0.640 ± 0.010 | 0.650 ± 0.016 | 0.652 ± 0.011 |

Table 3.4. Experiments conducted on DR datasets with 2 and 4 hospitalizations in order to test changes in accuracy results with the decrease of the training dataset size. The first column: the size of training dataset; second column: number of patients in positive cohorts for each of experimental settings; third column: Bi-directional GRU RNN classifier; fourth

column: 1-way LSTM RNN classifier, fifth column: RF classifier, sixth column: MLP classifier. Bold are the best accuracy results of the each type of experiments.

The prediction accuracy (GRU RNN model) of development of DR within the same year when DM2 was diagnosed, and after 1, 2, 3, 4-5 and 6-8 years of diagnosis of DM2 are presented in Figure 3.3. Experiments were completed with data of patients who had at least 2 hospitalizations or at least 4 hospitalizations after DM2 was diagnosed. All other complications have similar trends of the predicted accuracy regarding the time intervals.

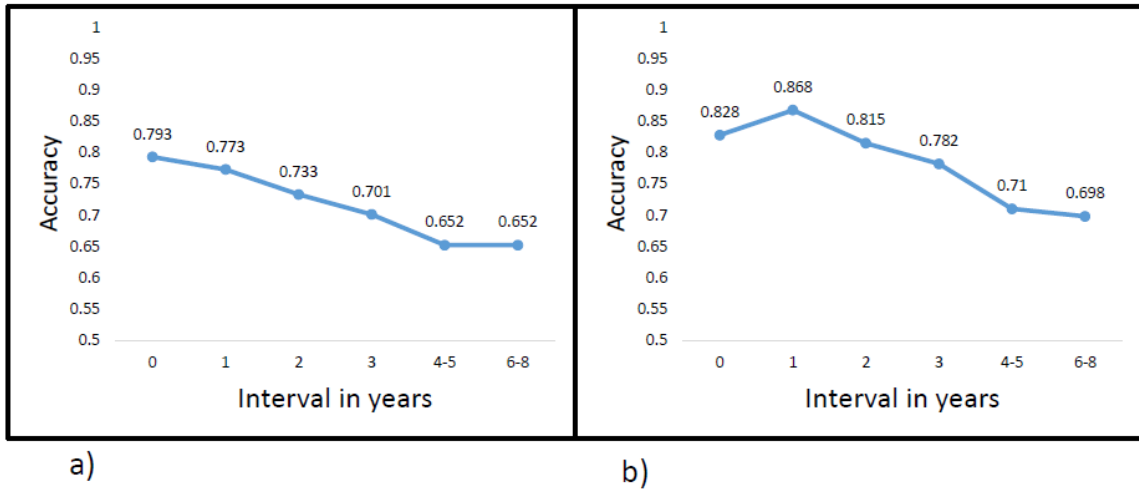


Figure 3.3. Prediction accuracy (RNN GRU model) that patients with DM2 will develop Diabetic Retinopathy after a minimum of 2 hospitalizations Fig. 3a), and after at least 4 hospitalizations (Fig. 3b). The results are presented by intervals when Retinopathy developed: within 1 year, after 1, 2, 3, 4-5 and 6-8 years from the diagnosis of DM2.

Predicted risk probabilities of development of each of ten studied complications in patients with DM2, according to HCUP data, are presented in Figure 3.4.

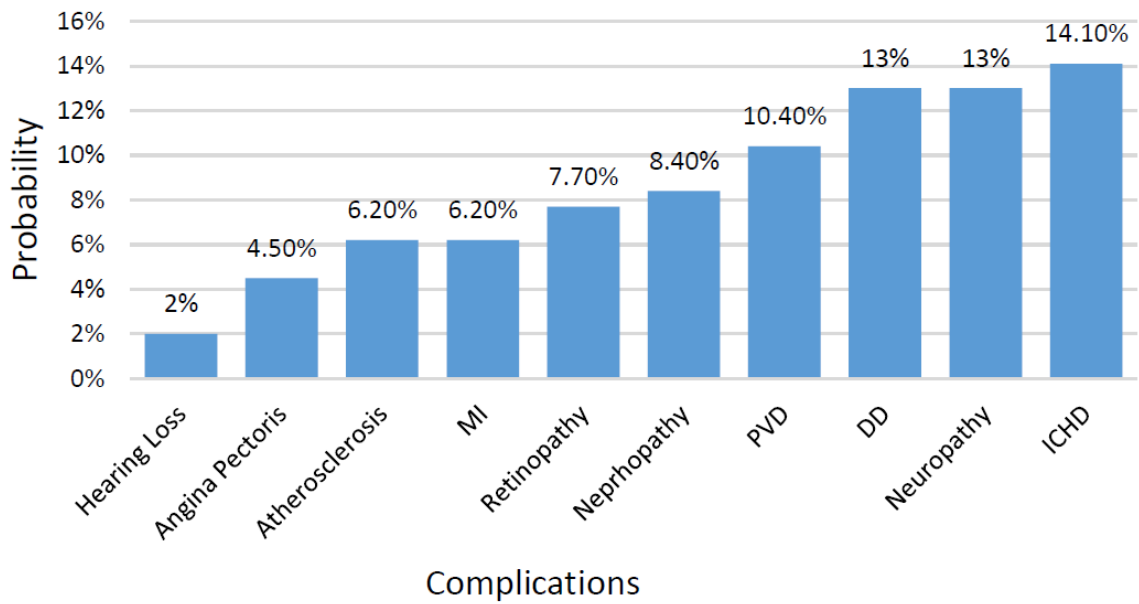


Figure 3.4. Predicted risk probabilities of development of each of ten complications in patients with DM2 (HCUP SID California data).

3.5. Discussion

In conducted experiments both deep learning algorithms were significantly more accurate than traditional models. Sarwar and colleagues reported accuracies for prediction of diabetes for the following ML models: LR 74% accuracy, SVM 77% , NB, DT 74%, RF

71% and KNN achieved 77%.¹³⁰ Ngufor and colleagues applied tree-based machine learning algorithms such as RF, gradient boosted machine (GBM), recursive partitioning, conditional inference trees and mixed-effect machine learning (MEMl) framework to predict longitudinal change in hemoglobin A1c.¹⁴⁰ Ngufor's model assumes that the number of variables which change over the time period is small. In the case of hemoglobin A1c application, there is only one continuous variable that changes longitudinally. However, in our experiments diagnoses are categorical variables (with more than 1000 categories) which change with each visit, making Ngufor's method inapplicable. In the last time, according to numerous publications, RNN based models proved to be superior to traditional models with high dimensional temporal EMR type of data. Choi and colleagues showed that RNN deep learning models performed better than traditional ML approaches on EMR temporal data.¹⁴¹ Massaro and collaborators described an application of a deep learning LSTM model as a very good choice for Diabetes prediction.¹²⁷ Zhang and colleagues applied the MetaPred model and transfer learning using CNN and LSTM RNN models in addition to traditional ML models.¹³¹

Our study focuses on detection of the most often complications of DM2, using high dimensional EMR type of data. RNN models, especially GRU, achieved state-of-the-art prediction accuracy (table 3.2) by discovering complex temporal relationships inside EMR data. A t-test ($p=0.05$ level) did not show a statistically significant difference between the two RNN models. Comparison of two traditional ML models shows that both models performed similarly. Our experiments show that an RNN GRU model is the best choice for

the high dimensional temporal EMR data. It performed significantly better than traditional models, according to a t-test at the level of $p=0.05$.

Considering the number of hospitalizations filter deep learning models achieved the best results when data include 4 hospitalizations after DM2 diagnosis vs. relying on less hospitalizations. The accuracy of our RNN GRU model with 4 hospitalizations achieved values between 73.3% (MI) and 83.5% (ICHD). Datasets with 2 hospitalizations are 2-3 times larger than datasets with 4 hospitalizations. Our results point toward the number of hospitalizations as more important factor for prediction results than the size of datasets. We did not find significant difference in prediction accuracy if the minimum number of hospitalizations was 3 instead of 4.

Analyzes of the influence of sizes of datasets indicate that about 1,000 patients are sufficient in the positive dataset for RNN models. The performance of RNN ML models decreased significantly when the size of datasets decreased to 500 patients. Traditional ML models did not show statistically significant changes in achieved accuracy if the size of datasets decreased to 500. They also did not show significant changes in accuracy with changes in the number of hospitalizations.

Figure 3.3. shows that prediction accuracy of DR decreases over time. In the case of 2 hospitalizations, the accuracy decreases steadily over time because the sizes of datasets are big enough not to affect the performance of deep learning models. In experiments with 4 hospitalizations the initial accuracy within one year is lower because that dataset was relatively small (523 patients) which is less than the optimal number of patients for the performance of the GRU model. After this initial period, the rest of the curve in Figure 3b

is similar to the curve in Figure 3a. Similar changes are noticed with all other complications.

Considering individual complications, RNN models were the most accurate when predicting Depressive disorder and ICHD. These two diseases had the largest absolute numbers of patients in positive cohorts. The prediction accuracy of ICHD was 83.5% which was significantly better than the prediction accuracy of Hearing Loss (the smallest dataset) or MI which were: 73.4% and 73.3% consequently. Individual diseases performed differently, and it is difficult to determine whether the size of datasets, comorbidities of individual complications, or perhaps time gaps between different visits influenced prediction results. Analyses of probabilities of development of ten complications show that ICHD, DD, and Diabetic Neuropathy have a higher probability of occurrence (13-14%) than all other complications, including hearing loss with the probability of occurrence of only 2%.

Results of our research with early and accurate prediction of ten frequent complications of DM2 are important for targeting high-risk patients for monitoring and intervention. They would enable the application of timely prevention measures which will postpone complications, improve quality of life and increase survival rates. Our methodology could be generally applied to prediction accuracy problems of any other disease or complications of that disease. It could be applied to predict cancer diseases, from EMR type of data, or it can be applied to predict chronic diseases, such as heart or lung diseases or complications of those chronic diseases. It can also be applied to predict acute medical conditions, such

as heart attack, stroke, acute kidney failure, the occurrence of infectious diseases (flu, coronavirus, hepatitis, etc).

Further improvement of created RNN models would improve the prediction accuracy of DM2 complications and other diseases, which could have significant clinical implications. It could become incorporated into a clinical decision support system and help clinical workers to improve the quality of healthcare. Our GRU RNN model can predict a clinical event (disease, complication) with high accuracy.

By demonstrating that the application of RNN deep learning models can make a successful prediction of clinical events we hope that our study may contribute to facilitating a wider use of ML in clinical medicine in the form of a clinical decision support system. Also, it could be applied in health care emergencies such as the current crisis with COVID19 virus to make some important and helpful predictions that will help public health experts.

3.6. Conclusion

Deep learning approaches, especially the RNN GRU model, were superior to traditional ML models with temporal EMR medical data. Conducted large scale experiments suggest that the number of hospitalizations (visits) should be three or more in the case of temporal data if deep ML models are applied. Tradeoff between the number of hospitalizations and the size of datasets should be considered, because datasets with three visits could be significantly larger than those with four visits which will require more computational resources.

Deep learning models applied on the HCUP data achieved a very good prediction accuracy with ten selected complications of DM2. Improvements in the accuracy of results might be possible if we had had data from other domains, such as labs or drugs, available.

Our study provides evidence that better understanding and management of DM2 from the aspect of the studied complications is possible when training deep learning models on appropriately preprocessed EMR data. An accurate prediction of the occurrence of complications is important in the planning of targeted measures aimed to slow down or prevent their development.

CHAPTER 4

SOCIAL NETWORK ANALYSIS FOR BETTER UNDERSTANDING OF INFLUENZA

4.1. Objective

Infectious diseases, like influenza, can have devastating consequences on populations. Influenza is associated with substantial morbidity and mortality.¹⁴² In addition to clinical impact, Influenza has significant economic impact. Starting from 2010, each year CDC estimates the burden of influenza in the U.S. The burden of influenza disease in the United States can vary widely and is determined by a number of factors including the characteristics of circulating viruses, the timing of the season, how well the vaccine is working to protect against illness, and how many people got vaccinated. CDC estimates that influenza has resulted in between 9.3 million – 49.0 million illnesses, between 140,000 – 960,000 hospitalizations and between 12,000 – 79,000 deaths annually since 2010. The estimated number of flu illnesses during the 2017-2018 season was 49 million, flu hospitalizations - 960,000, and flu deaths 79,000.¹⁴³

Standard medical treatments and vaccines are often not sufficient to stop flu infections. Equally important, is to understand how the pathogen spreads in the population.¹⁴⁴

Understanding the nature of human contact patterns is crucial for predicting future pandemics and developing effective control measures.¹⁴⁵ Explorations of spatio-temporal spread is also, very important in order to explain, are Influenza infections more spatially synchronized and widespread in populous highly connected areas, compared to smaller, more isolated ones.¹⁴⁶ Reasons for geographical trends of spreading of Influenza could be explained in terms of population size, connectivity, and demographics. Understanding the spatio-temporal spread of infectious disease is important both for the design of control strategies and to deepen fundamental knowledge about the interaction between infectious diseases dynamics and spatial mixing of the population.¹⁴⁷ It's important to locate geographic hotspots (so called hubs) for Influenza infections.¹⁴⁸ Analyses of geographic distribution of Influenza and demographic characteristics of geographic hotspots, help us to make better planning of hospital resources for complicated cases of the flu and better management of healthcare systems. Environmental factors, population sizes and, demographics, are major determinants in disease spread and potential complications, which can result in admissions to hospitals. Influenza causes many complications, that can worsen the disease, require hospitalization and complicate outcomes. Respiratory, cardiovascular, digestive system, and other complications have been studied.¹⁴⁹⁻¹⁵¹

The objective of this research is to develop a novel method, that leverages options of heatmaps and network science in explanations of the spatial distribution of patients with complications of Influenza. The goal is to visualize complicated flu cases throughout a particular geographic region. Public health experts, doctors, and other medical scientists, by using the results of the visualization tools, like heatmaps, could rapidly recognize, how

flu cases are distributed, and how they expand geographically. This knowledge would help them plan resources for earlier detection of flu and reduce the future impact of influenza.¹⁵² Further, we will utilize network science options to analyze correlations among zip codes of the NY state, considering numbers of hospitalized flu cases over the 10 years period. The goal is to show observed and calculated correlations in the network, as nodes (zip codes) linked, based on strength of correlations. That will allow calculation of nodes degrees, with the objective to find the most connected nodes and hubs, based on results of these network centrality measures. Adequate measures should be applied to isolate (treat) zip codes that represent discovered hubs, in order to decrease the number of complicated flu cases in the future. The final objective is to present the results of our novel method to health professionals and researchers, which will help them to plan adequate resources to contain flu infection and prepare appropriate hospital resources for patients with complications. Results could potentially save many lives and improve the health of the population. The experiments were conducted on the state of New York data, but the proposed method is scalable and could easily be generalized to any other geographic region in the U.S. and all over the world.

4.2. Background and significance

According to the Centers for Disease Control (CDC), seasonal influenza infects approximately 5–20% of the U.S. population every year.¹⁵³ Connections between Influenza and networks is a well-studied topic, that dates back to the mid-1980s and many papers describe the association between influenza and networks. To assess the influence of

network effects, the predictions were compared, from the detailed network model, consisting of fixed contacts of known weights, to several simplified alternatives. Spatial spread of influenza infections and geographic transmission hubs were analyzed and described in recent publications.¹⁴⁶⁻¹⁴⁸ Numerous research studies evaluated the risk and development of complications associated with influenza virus infections.^{149,150} Many of those complicated cases require hospital treatment. Researchers described influenza-associated critical illness hospitalizations.¹⁵¹ Demographic factors associated with influenza A(H1N1) infection have been studied.^{154,155} Many authors assessed the network configuration, network stability, and changes in risk configuration and risk behavior, using social network analysis and visualization techniques. The evolving science of social networks has evident potential to help researchers to explain the spread of infectious diseases.¹⁵⁶ Human social networks change over time: we typically do not meet exactly the same individuals every day. Gligorijevic and colleagues studied the importance of the confidence of predictions in longer-term forecasting in health and climate domains.¹⁵⁷ They presented an effective novel iterative method developed for Gaussian structured learning models, for propagating uncertainty in temporal graphs, by modeling noisy inputs (most of the inputs in the field of infectious diseases). Good planning of hospital resources will also need a prediction of length of hospital stay, for individual patients, in addition to prediction of numbers of potential hospitalizations. Stojanovic and collaborators described how to learn low-dimensional vector representations of patient conditions and clinical procedures in an unsupervised manner, and generate feature vectors of hospitalized patients, useful for predicting their length of stay, total incurred charges, and mortality rates.⁴ Barabasi and

Kleinberg published a robust analytical and numerical framework to mathematically model the spread of pathogens.^{158,159} Meyer developed power-law models to better capture dynamics of infectious disease spread.¹⁶⁰ He demonstrated power-law model frameworks and spatial distribution heatmaps on meningococcal bacterial meningitis in Germany and influenza virus infectious disease in Southern Germany. Many papers described social network application and geographical distribution in explaining other diseases besides human types of Influenza. Poolkhet in his study described social network analysis for assessment of avian influenza spread and trading patterns of backyard chickens in Thailand.¹⁶¹ Song and colleagues used Pearson's correlation to measure the impact of socioeconomic factors on AIDS diagnosis rates in certain geographic areas. The correlation based method discovered the complexity of contribution of socio-demographic determinants of health and geographic area based measures to AIDS diagnosis.¹⁶²

In our study, we proposed a novel method for better understanding of the geographical distribution of hospitalized Influenza cases in a specific geographical region, using the combination of geographically specific heatmaps and social network analysis. A combination of social network analysis and visualization of findings on interactive geographical heatmaps is a novelty, that provides quick and efficient information about hubs and spatial distribution of hospitalized flu patients. We calculated correlations among zip codes in the state of New York. Zipcodes represent specific geographic localities in a specific state, with very characteristic demographics in each of them. In our research, we used detailed demographics from the 2010 Census. If we consider that the particular zip code has specific demographic characteristics, then we can assume that the specific

geographical location of the zip code and demographic characteristics, affect the numbers of cases and hospitalizations of Influenza. We then find correlations of these zip codes, knowing that actually, we calculate correlations of numbers of cases affected by geographic locations and demographic characteristics of zip codes.

We took a period of 10 years, since we can draw conclusions about the distribution of Influenza cases in so long period. Based on our findings we can expect similar patterns in the next decade. In order to visualize findings, our method uses contemporary Google maps as the base for heatmaps. Researchers can clearly see the names of cities, roads, rivers, mountains. Visualization is more effective than a description of regions, because researchers who are not familiar with all places in one state, can quickly see where the cities are located, how are they connected (highways, roads), and are any geographic features (mountains, lakes...) between nearby cities that can slow down the spread of infections between 2 cities. Especially if they are not connected with direct roads.

Our novel method, proposed in this study, is a different methodology from previously published approaches. We constructed heatmaps, that show the distribution of flu patients in NY state, which enables easy and fast visualization of zip codes, that are the most affected by flu infection, as well as visualization of the most likely routes of Influenza spreading. We performed a network analysis of distributions of patients through zip codes, where nodes represent affected zip codes and links represent correlation among zip codes. The designed network allows calculation of centrality measures, aimed to provide discovery of hubs and significance of individual zip codes. These findings could help medical professionals to improve the planning of resources, needed to treat flu infections

and complications and to better allocate resources. Our approach will provide a fast and accurate understanding of Influenza in specific geographic areas, which can be the size of one or more states or more countries. Detailed heatmaps will locate regions, that need the most resources for medical intervention. Our model provides more geographic details about the distribution of flu infection, than previously published research.

We conducted our research on Healthcare Cost and Utilization Project (HCUP) data for the period of 10 years, and we recommend further study of the geographical distribution of Influenza on more different datasets in order to better understand the geographical distribution of hospitalized influenza patients and what demographic and socio-economic factors contribute to that distribution.

4.3. Materials and methods

Proposed is a novel method for better understanding of the geographical distribution of hospitalized Influenza cases in New York state, using the combination of geographically specific heatmaps and social network analysis. We analyzed data from the HCUP, the State Inpatient Databases (SID). HCUP is a family of health care databases that contain the data of State data organizations, hospital associations, private data organizations, and the Federal government. The HCUP includes the largest collection of longitudinal hospital care data in the United States and contains information on inpatient stays, emergency department visits, and ambulatory care. The SID are state-specific files that contain all inpatient care (hospital) records in participating states. The State-specific SID encompasses more than 97 percent of all U.S. hospital discharges.

Data for this project were downloaded from the HCUP - SID New York State inpatient database. We downloaded and analyzed data regarding influenza infections for the period of 10 years (2003-2012). There were 30,380 cases of influenza registered in the database. Those were patients who required hospital admission and stay, due to more severe flu or the presence of complications.

Influenza cases were depicted on the bar plot and visualized on heatmaps. We further analyzed these heatmaps with absolute numbers of hospitalized flu patients, to study activity and spatial spreading of the flu. We, also, normalized absolute numbers of patients over the population in each of the zip codes. Normalization of results helped us discover which zip codes were the most prone to flu infections. In order to conduct the study, we used demographic data from the Census Bureau from the last census conducted in 2010. Census demographic data match the period of the processed data from HCUP databases. We analyzed a percentage of the population affected by severe flu complications that required hospitalizations. We normalized the number of patients per number of people who lived in individual zip codes to obtain percentages of affected population per zip code.

Next, we constructed a power-law type network. Statistical analysis was performed to determine if the network follows power laws. A Kolmogorov–Smirnov test was conducted, at the significance level of 0.05. We created a function to estimate the exponent and to plot the log-log data and the fitted line. Networks whose degree distributions follow a power-law are called scale-free networks. The probability of observing high-degree nodes, or hubs, is very high in this type of network. Scale-free networks, also, have a large number

of small degree nodes that tend to connect among themselves and are virtually absent in a random network.

In order to utilize network analysis of the geographic distribution of flu patients, we constructed a matrix (1,471 rows and 12 columns). The rows represent 1,471 zip codes in the state of NY, from which, patients were registered in the HCUP-SID database. The columns represent 12 months. Suitable for analysis of the flu distribution, through zip codes, was the weighted signed correlation network. To form this network, we constructed a 1,471 x 1,471 matrix with the aim of calculating the correlation between zip codes. Pearson's correlation (r) was calculated between pairs (x,y) of zip codes. (formula.4).

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2} \sqrt{\sum(y_i - \bar{y})^2}} \quad (4)$$

We constructed the network from Pearson's correlations of numbers of complicated flu cases among zip codes. The network is first represented as a weighted correlation adjacency matrix.

For detailed network analysis, we chose 100 zip codes, with the largest number of patients (70 and more), due to the privacy protection of patients. The correlation was calculated among the top 100 zip codes. The adjacency matrix, $A[ij]$, encodes whether and how a pair of nodes is connected. For weighted networks, the adjacency matrix reports the connection strength between node pairs.

There are two types of weighted correlation networks: unsigned and signed. Unsigned networks have an absolute value of correlation $|cor|$ and Signed networks keep the sign of the correlation $\frac{(cor + 1)}{2}$.

We selected a signed correlation network. The nodes of such a network correspond to zip codes, and edges between them are determined by the pairwise Pearson correlations between zip codes. The network was created from the adjacency matrix and performed power transformation (normalization) of calculated Pearson's correlations. For normalization, we used signed network normalization $|(\text{corMatrix}+1)/2|^\beta$ for 100 x 100 adjacency matrix. By raising the absolute value of the correlation to a power $\beta \geq 1$ (soft thresholding), the signed correlation network emphasizes high correlations at the expense of low correlations. We tested β -values between 1 and 10, and chose the value of $\beta = 2$, as the best choice to represent the flu infection by the correlation among zip codes in the network. We used the correlation matrix after the transformation (normalization) as adjacent matrices to plot the network. The cutoff for edges, to be plotted on the network, was set to some reasonable number (smaller correlations were not plotted). The cutoff correlation of 0.9 and higher was selected to be plotted as an edge. In order to determine hubs in the network, we calculated degrees for all 100 nodes.

The associated network, based on correlation results between zip codes of patients was constructed in R, with the help of WGCNA, Statnet, and gplot packages.

Our method brings together Biomedical informatics, Medicine, and Network science, in an attempt to illuminate the nature of Influenza, in this specific population. This method can be generalized and applied to any other infectious disease and geographic region in the U.S. and in the world.

4.4. Results

Data from the HCUP–SID New York State database for the period 2003-2012 were analyzed. We studied the evolvement of flu infections that required hospitalization throughout the year, with monthly breakdown of cases (January through December), for 10 years with a total number of 30,380 inpatient cases, with influenza diagnosis. The display of monthly breakdown of the number of hospitalized cases for the flu is shown on the bar plot (Figure 4.1). The highest number of cases was registered in December (6,720). Flu virus infections were also very active in January, February and March. Out of the peak of the flu season, cases were sporadic, even in big cities.

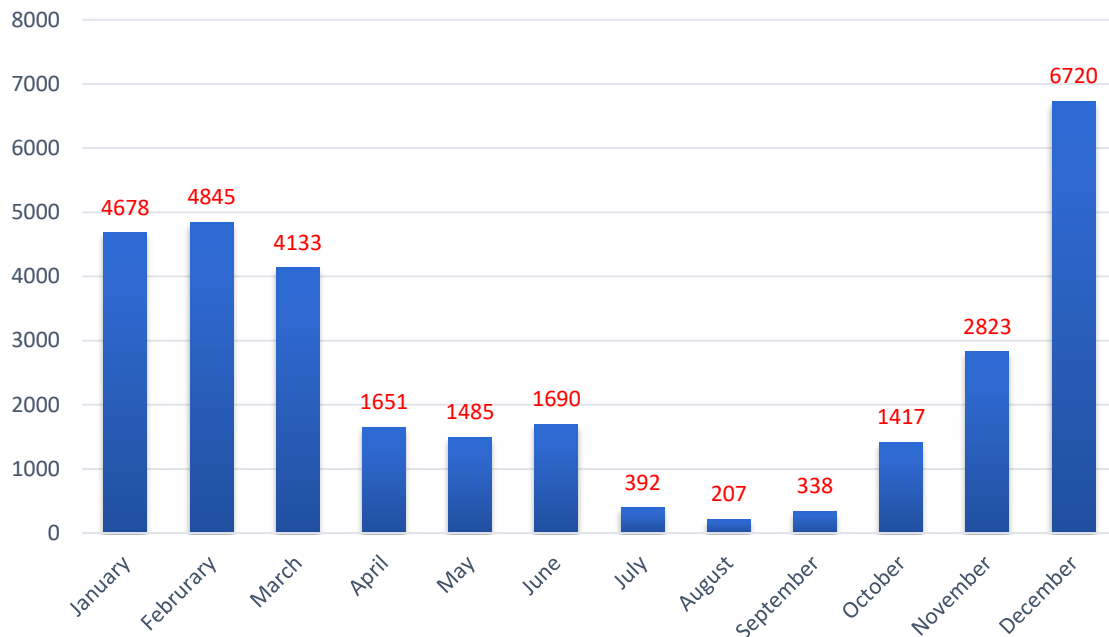


Figure 4.1. Bar plot – Number of patients infected by the influenza virus during the 2003-2012 period (monthly distribution), who were hospitalized in the state of New York.

The Heatmap (Figure 4.2) of the state of NY, which shows the distribution of hospitalized patients with flu complications throughout different zip codes, was constructed for the same period of 10 years. Dots on heatmaps show numbers of patients who reside in particular zip codes. We show only zip codes with more than 20 cases of flu (due to privacy reasons). The total number of zip codes shown on the Heatmap was 443, with the total number of patients of 29,071.

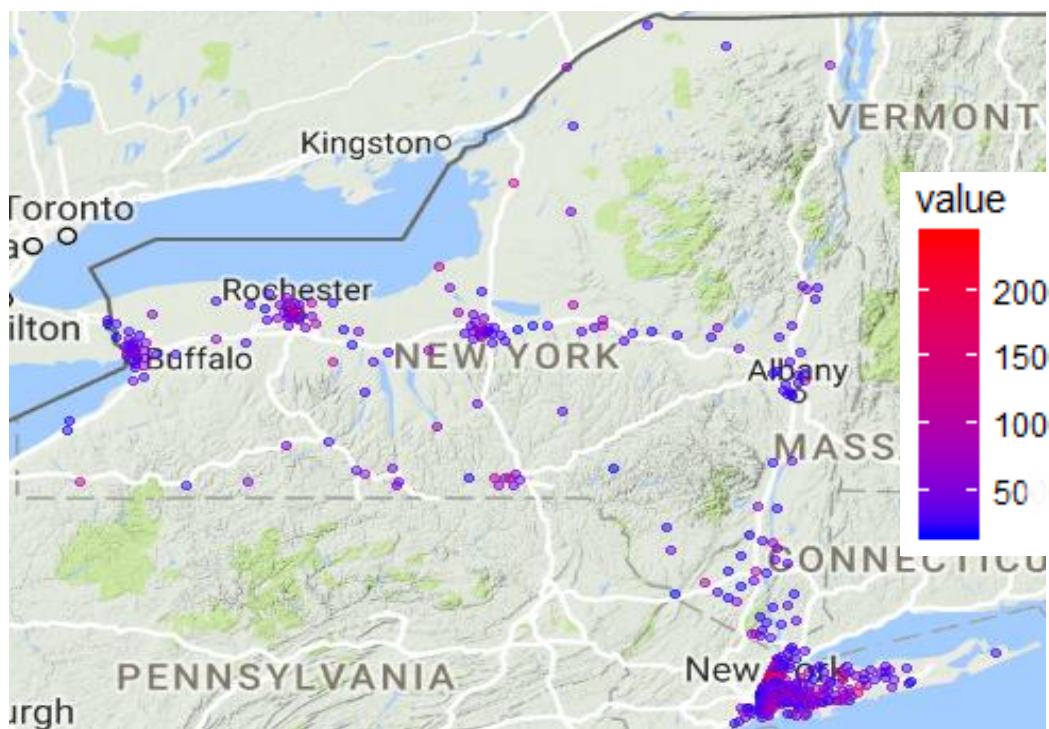


Figure 4.2. Heat map of NY state - Distribution of hospitalized flu patients by the zip code (2003-2012), shows that the highest concentration of hospitalized flu patients were in five big urban areas (Albany, Buffalo, New York City, Rochester and Syracuse). The heatmap also shows that routes of spreading follow highways (in particular Highways 81, 86 and 90) as high frequency routes of travelling between places.

The highest numbers of cases are registered in the most urban zip codes. Red color (high number of cases) on heatmaps is noticeable in big cities: Albany, Buffalo, New York City,

Rochester, and Syracuse. Also, we can observe that a lot of blue dots (that correspond to zip codes with 20-50 cases) are highly concentrated in big urban areas. An important finding on the heatmaps, is that the distribution of hospitalized patients follows highways or other big roads, which indicates that flu spreads along the routes that people use to move from place to place. Rural areas had small numbers of cases. Further, we present heatmaps of big urban areas separately (Figure 4.3): a) Albany area, b) New York City area, c) North side of NY State and d) Buffalo area. These heatmaps can be used for future predictions and healthcare planning for particular zip codes, as the areas with the highest risk for the flu infection outbreaks and spreading. Accordingly, health care providers should plan more resources to deal with sick patients in these particular areas.

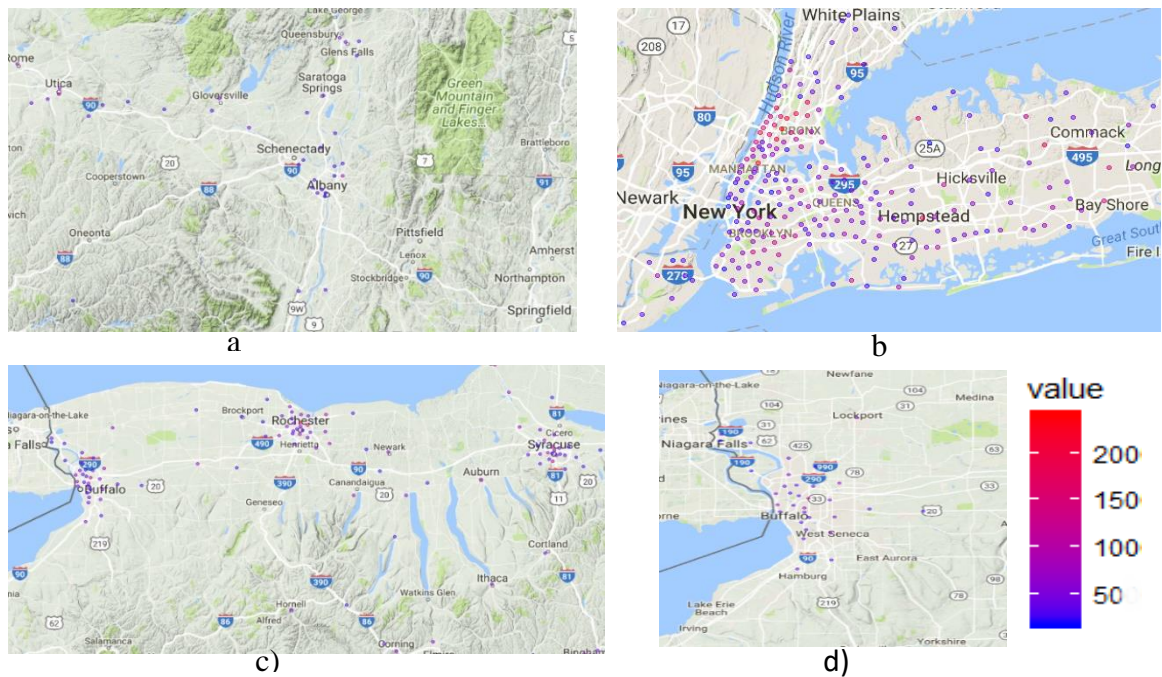


Figure 4.3. Heatmaps of a) Albany area, b) NY City area, c) North side of NY state, d) Buffalo area. Heatmaps show that the distribution of hospitalized flu patients by the zip code in period between 2003-2012, was highly concentrated in five big cities (Albany, Buffalo, New York City, Rochester and Syracuse). Heatmaps show that the routes of distribution follow highways (highways 81, 86 and 90, in Albany area highways 87 and 9 and in Buffalo area, highway 190 toward Niagara Falls)

We analyzed the distribution of population per zip codes in the state of NY. Populations of 20 the largest zip codes are shown in table 4.1.

| | | | | | | | | | | |
|------------|--------|--------|--------|-------|-------|-------|-------|-------|-------|-------|
| Zip | 11368 | 11226 | 11373 | 11220 | 11385 | 10467 | 10025 | 11208 | 11236 | 11207 |
| Population | 109931 | 101572 | 100820 | 99598 | 98592 | 97060 | 94600 | 94469 | 93877 | 93386 |
| Zip | 11219 | 11211 | 11377 | 11214 | 11234 | 10456 | 11230 | 11355 | 10314 | 11212 |
| Population | 92221 | 90117 | 89830 | 88630 | 87757 | 86547 | 86408 | 85871 | 85510 | 84500 |

Table 4.1. Distribution of population per zip codes (20 the largest) in the state of New York from census data for 2010.

We presented 20 zip codes with the largest absolute number of hospitalized patients infected with flu in the period of 2003–2012 in table 4.2.

| | | | | | | | | | | |
|------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| <i>Zip</i> | 10457 | 10458 | 10467 | 10029 | 10456 | 10032 | 11550 | 11746 | 10466 | 10463 |
| Population | 241 | 221 | 208 | 207 | 200 | 190 | 185 | 184 | 178 | 177 |
| Zip | 10468 | 10453 | 14621 | 10452 | 11542 | 10033 | 11717 | 10031 | 10469 | 14609 |
| Population | 176 | 172 | 168 | 164 | 163 | 160 | 154 | 152 | 152 | 146 |

Table 4.2. Number of hospitalized flu patients in the state of New York for the period 2003-2012 (HCUP data).

Simple inspection of these tables discovered that the highest populated zip codes didn't have the largest number of patients with flu complications. The highest populated zip code at the 2010 census was 11368 (Queens) with 53.6 % males and 46.4 % females and average age 31.8 (significantly lower than the state average). About 59% were singles. Comparing to NY state averages we found that average income was 3 times lower than the average state income. Hispanic population percentage (74%) was significantly above the state average, median age below the state average, foreign-born population percentage significantly above the state average. There are no hospitals in this zip code, and hospitals in the area are not well ranked. The percentage of the population without health insurance was extremely high 31%, which is much higher than the state average (8.7%). The second highest populated zip code was 11226 (Brooklyn) with 44.9 % males and 55.1 % females. About 71% of the population were black, and 14% Hispanic. Comparing to the state average: Black race population percentage (71%) significantly above the state average, followed by 14% Hispanics, median age below the state average. Foreign-born population percentage was above the state average. There are few hospitals in this zip code, but the percentage of the population without health insurance was 15% (higher than neighborhood areas). The third most populated zip code was 11373 (Elmhurst) with 50.4% males and 49.6 % females. 22% of the population didn't have health insurance, which was significantly higher than surrounding areas. Hospitals are available in this area. Hispanic (43%) and Asian (47%) population percentages were above the state average and foreign-born population percentage significantly above the state average. The average income was significantly below the state average.

Analyses of the top 20 highest populated zip codes in the NY state lead us to discoveries, that all of these zip codes are located in the NY City area, mostly in Brooklyn. Top 5 zip codes have significantly lower average income than the average income in the state. They also have a significantly higher percentage of the population without health insurance, and high percentage of foreign born, as well as Hispanic and black population. The average number of household members was higher than the state average. Common conclusion for this population could be that, due to lower income and a lower percentage of health insurance, residents of these zip codes haven't visited hospitals. One of zip codes among the top 20 highest populated zip codes was a zip code 10025 with a predominantly white population and higher average income and a higher percentage of the population with health insurance. This zip code is not on the list of the top 20 zip codes with the highest number of hospitalized patients, which could lead to the conclusion that patients got necessary health care treatment before the flu developed complications, or they had fewer flu cases due to preventive measures.

The top zip code with the most patients who were hospitalized due to Influenza, was 10457 (Bronx) with 241 patients. Census data for 2010, show that the zip code had the population of 70496 (65% Hispanic, 30% black, 1.5% white and Asian...). The median age was 29.8 with 52.5% women and 47.5% men. The median income was about \$24000 (3 times less than the state average income). The second zip code by the absolute number of hospitalized patients from Influenza was 10458 (Bronx) with 221 patients. This zip code had a population of 79492 (64% Hispanic, 20% black, 10% white, 4% Asian...). The median age was 29.3 with 52% women and 48% men. The median income was about \$25700 (3 times

less than the state average income). The third zip code by an absolute number of hospitalized patients from Influenza was 10467 (Bronx) with 208 patients. This zip code had a population of 97060 (48% Hispanic, 33% black, 10% white, 6% Asian...). The median age was 33.6 with 52% women and 48% men. The median income was about \$31500 (2.5 times less than the state average income). Further analysis revealed that, among the top five zip codes, 4 are in Bronx, with 870 patients hospitalized due to flu complications. We can also notice that among the top 20 zip codes, most of the patients were in Bronx.

Furthermore, we normalized the number of hospitalized patients per population in zip codes and calculated a percentage of the population affected by severe flu complications, that required hospitalizations. We show percentages of affected population per zip code in table 4.3. This time, the highest percentage of affected population was in the zip code 11509 (Atlantic Beach, NY). In this zip code, we found 29 hospitalized flu patients per 2,653 residents. Zip codes with small numbers of residents have a higher percentage of affected people than more populated zip codes. It's interesting that none of the zip codes from Bronx are in the top 20, despite the fact that they had the highest total numbers of patients.

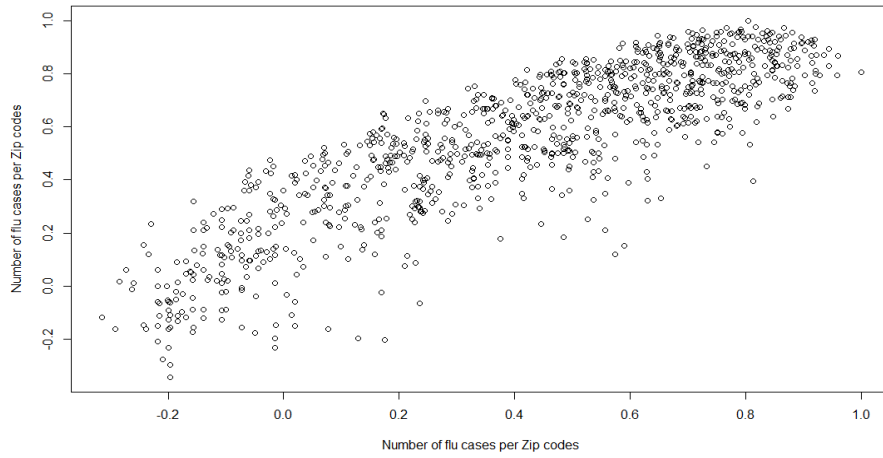
| | | | | | | | | | | |
|---------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| Zip | 11509 | 11542 | 13205 | 14895 | 13452 | 13367 | 13202 | 13904 | 14605 | 14621 |
| % of patients | 0.011 | 0.006 | 0.006 | 0.006 | 0.006 | 0.005 | 0.005 | 0.005 | 0.005 | 0.005 |
| Zip | 13204 | 13203 | 14514 | 13669 | 11798 | 13208 | 13905 | 14513 | 14482 | 14843 |
| % of patients | 0.004 | 0.004 | 0.004 | 0.004 | 0.004 | 0.004 | 0.004 | 0.004 | 0.004 | 0.004 |

Table 4.3. Percentages of affected population for 20 zip codes with the highest percentages of hospitalized flu patients.

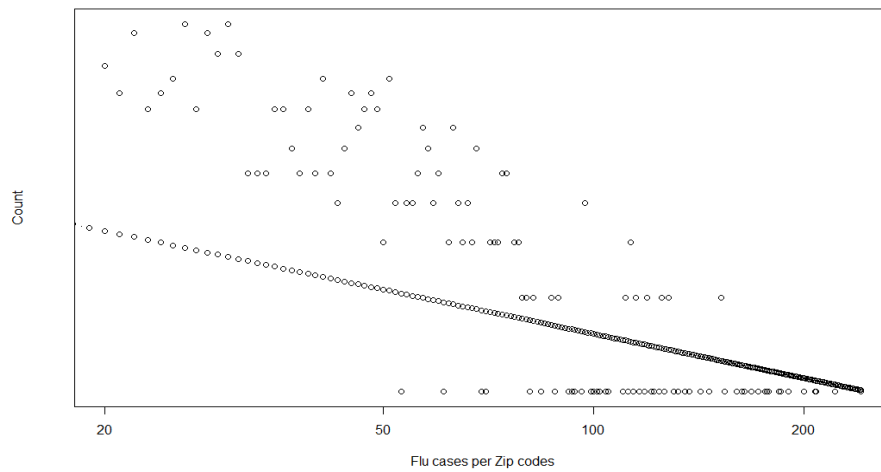
To further analyze the geographical distribution of hospitalized patients with Influenza infections a power-law type network was constructed. In order to create this network, we calculated correlations among zip codes, considering the distribution of hospitalized flu cases, over the 10 years period. A common property of power-law type networks is that the node degrees span several orders of magnitude. Outliers, or exceptionally high-degree nodes, are not only allowed but expected in these networks. The main reason to construct a power-law type network was to identify highly connected nodes (hubs). If we can locate hubs, that will help us to prepare an adequate public health strategy to eliminate them and decrease the magnitude of influenza infections in those spatial regions, which will significantly alleviate the cost that influenza infections impose on populations.

We performed statistical analysis to determine if the network follows power-laws. A Kolmogorov–Smirnov test was applied at the significance level of 0.05. Obtained results show that at this significance level, the network is of power-laws type ($p=0.01$). We created the function that helped us to estimate the exponent, plotted the log–log data and the fitted line (Figure.4). The calculated value of the degree exponent $\gamma = 2.5935$ (t-statistic value = 9.804 (p-value very small), SE = 0.2645, distance distribution = 3.618).

The network is first represented as a weighted correlation adjacency matrix. Initially, we calculated Pearson’s correlation among all 1,471 zip codes (plot of correlation is shown in Figure 4.4).



a)



b)

Figure 4.4. a) Plot of correlation between 1,471 zip codes with respect to number of hospitalized patients with flu complications in those zip codes between 2003 – 2012 b) Plotted the estimate of the power law exponent, the log–log data, fitted line at $\gamma = 2.5935$.

Detailed network analysis was performed on 100 zip codes with the largest number of patients (70 and more). We used the correlation matrix, after the power transformation (normalization) as adjacent matrices to plot the network. We selected the cutoff correlation

of 0.9 and higher to be plotted as an edge. The network is shown on Figure 4.5. We picked two different colors to make nodes and labels more visible, with no other meanings.

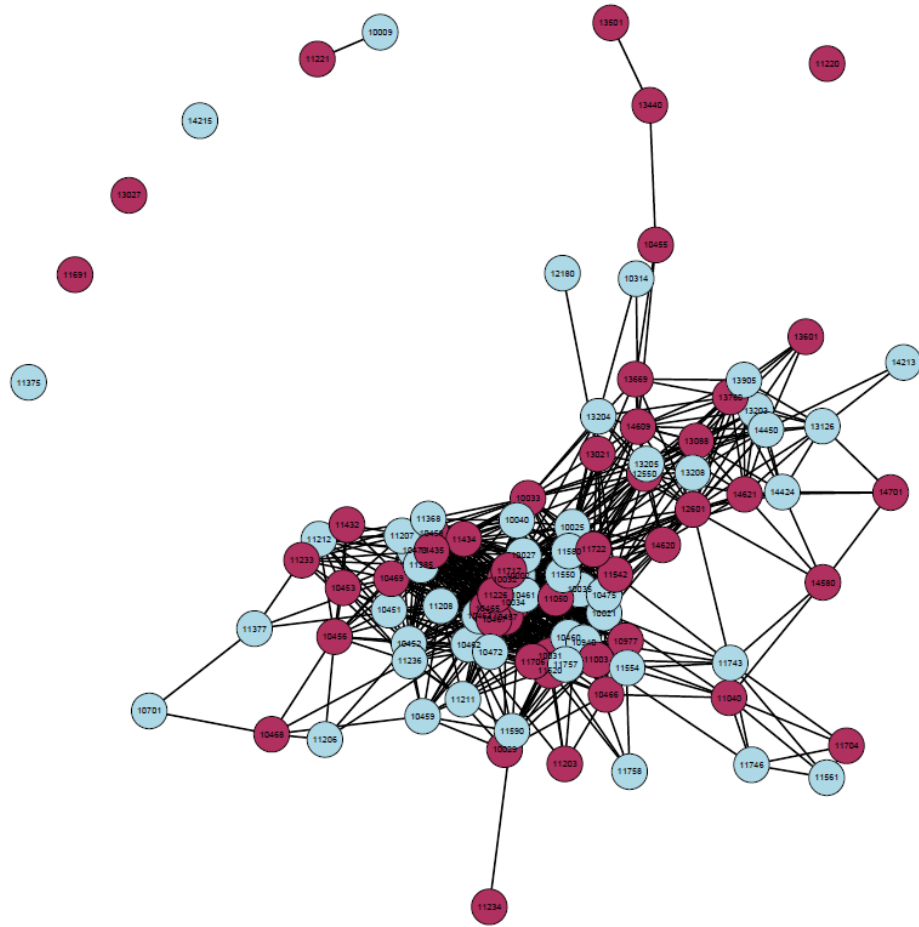


Figure 4.5. The power law type network representation of hospitalized flu cases in the State of New York between 2003 and 2012. based on correlation between zip codes. Nodes correspond to zip codes that are linked, based on strength of calculated correlations. Zip codes (nodes in the network) with the highest node degrees are: 10465 (Bronx) with the degree of 86, then 11226 (Brooklyn) with the degree of 84 and 10027 (NY City) – degree 80.

Analyzing the network, we can clearly see nodes with high degrees (hubs) as well as isolated nodes. Hubs correspond to flu cases in big urban areas (cities) in the State of NY. Disconnected nodes correspond to isolated cases in rural areas. By studying the network,

it's possible to identify critical clusters, hubs, and routes, that could be subjects of intervention, in order to minimize the spread of Influenza, decrease the numbers of complicated cases that require hospitalization, and decrease the cost.

In order to determine the significance of particular zip codes and which zip codes were hubs in the network, we calculated degrees for all 100 nodes. Zipcodes (that represent nodes in the network) with the highest node degrees are: 10465 (Bronx) with the degree of 86, then 11226 (Brooklyn) - degree of 84 and 10027 (NY City) - degree 80. The top 20 zip codes with the highest degrees are shown in Table 4.4.

| | Zip code | Township names | Node degree |
|----|----------|---|-------------|
| 1 | 10465 | Eastchester Bay, Bronx | 86 |
| 2 | 11226 | Flatbush, Brooklyn | 84 |
| 3 | 10027 | NY City | 80 |
| 4 | 10457 | Morningside Heights, Uptown, Manhattan, NY City | 80 |
| 5 | 10463 | Riverdale, Bronx | 80 |
| 6 | 11580 | Valley Stream, NY | 80 |
| 7 | 10035 | East Harlem, Harlem, NY City | 78 |
| 8 | 10467 | Van Cortlandt Pk, Bronx | 78 |
| 9 | 10032 | Washington Heights, Manhattan, NY City | 76 |
| 10 | 10460 | Bronx Park South, Bronx | 74 |
| 11 | 11550 | Hempstead, NY | 74 |
| 12 | 10034 | Inwood, Uptown, Manhattan, NY City | 72 |
| 13 | 11722 | Central Islip | 72 |
| 14 | 10461 | Westchester Square, Bronx | 70 |
| 15 | 11717 | Brentwood | 70 |
| 16 | 11003 | Elmont | 68 |
| 17 | 11520 | Freeport | 68 |
| 18 | 10025 | Upper West Side, West Side, NY City | 66 |
| 19 | 10462 | Van Nest, Bronx | 66 |
| 20 | 11208 | City Line, Brooklyn | 66 |

Table 4.4. Zip code with the highest node degrees

Census data for 2010, show that the zip code 10465 had a population of 42230 (51% white, 37% Hispanic, 7% black, 3% Asian...). The median age was 40.7 with 52% women and 48% men. The median income was about \$55400 (1.4 times less than the state average income). The number of people without health insurance was 7.3% which is better than the state average (8.7%). Few hospitals are available in the area. According to 2010 census data, zip code 11226 had a population of 101572 (71% black, 17% Hispanic, 6% white, 3% Asian...). The median age was 34.3 with 55% women and 45% men. The median income was about \$33400 (2.3 times less than the state average income). Uninsured population was 14.7%, which is higher than the state average and there are few hospitals in this area. Zip code 10027 had a population of 59707 (40% black, 26% Hispanic, 23% white, 8% Asian...). The median age was 30.8 with 53.5% women and 46.5% men. Median income was about \$50000 (1.5 times less than the state average income). The number of uninsured people was 10%, with few hospitals in the area. Zip code 10457 (Morningside Heights, Uptown, Manhattan, NY City) had a population with a higher number of uninsured people 13.6% and lower income than the state average. This zip code had the highest number of hospitalized patients (241) in NY state. Zip code 10463 (Bronx) had the 10th highest number of hospitalized patients (177). The number of people without health insurance was 9.3% and the average income was slightly lower than the state average. Further inspection of the list of zip codes with the highest node degrees in the constructed social network, shows that zip codes were located in the NY City area, with moderate to high population sizes. Most of them had higher than average percentages of the population without health insurance. The vast majority of zip codes had lower than average income.

Females were the majority of the population in most zip codes. And almost all zip codes had significantly higher Hispanic and black population than the national average. We constructed heatmaps to visualize geographic locations of the top 20 zip codes with the highest node degrees (Figure 4.6). We can clearly conclude that all 20 zip codes with the highest node degrees (hubs) are in the NY City area.

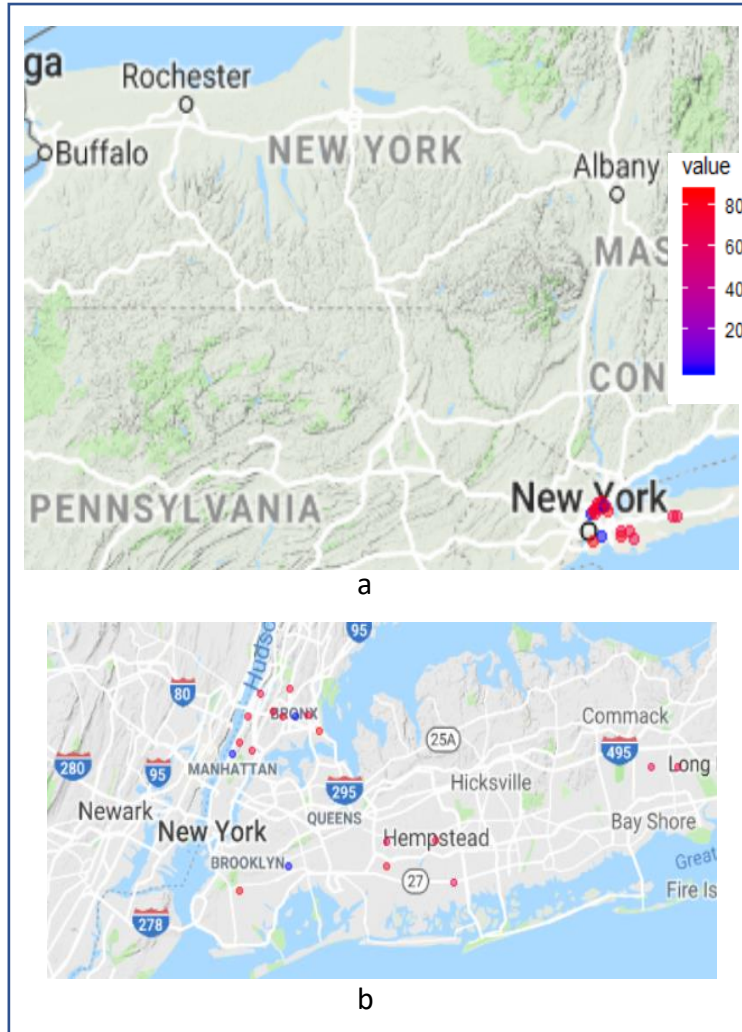


Figure 4.6. Geographic location of 20 zip codes with the highest nodes degrees (hubs) in the state of New York. a) the whole NY State, b) NY City area – 20 top hubs are in this area. Zip codes (dots on maps) with the highest node degrees are: 10465 (Bronx) with the degree of 86, 11226 (Brooklyn) with the degree of 84 and 10027 (NY City) – degree 80.

4.5. Discussion

The goal of this analysis of influenza is to contribute to a better understanding of the spatial spreading of hospitalized flu cases. If we convey this research into the lower numbers of infected people, the result will be more saved lives, a decrease in medical costs, and economic losses. We wanted to show that the ability to more accurately analyze and assess infection levels in geographic regions, that have higher infection risk, in the future, can suggest targeted planning and measures to deal with complications of Influenza. We provided a detailed analysis of hospitalized cases, caused by flu infection in the state of NY. We constructed heatmaps to visualize the findings of our research. Heatmaps show that the majority of cases are located in big cities. Other interesting finding, from heatmaps, was that the spreading of Influenza around the state was along highways and big roads, which means that influenza spreads along the routes that people use to travel. Out of 20 zip codes with the highest absolute number of flu infected patients, who required hospitalizations, most of them were in Bronx. Demographic data from the 2010 census show, that the top 3 zip codes (10457, 10458, 10467) by the highest absolute number of patients were located in Bronx. All 3 zip codes had a predominantly Hispanic population, followed by black, white, and Asian population. Further, all 3 zip codes had significantly lower median age (29.8, 29.3, 33.6) of the population than the median age of NY state, which is 38.4. The average income in all 3 zip codes was significantly lower than the average income in NY state (2.5-3 times). We also found that normalized data per number of residents in zip codes show that zip codes with the largest percentage of population affected by influenza were different than the zip codes with the largest absolute numbers

of patients. The largest percentage of patients has been found in Atlantic Beach, which has a relatively small population. The second largest percentage of patients per population was found in Glen Clove. Both of these zip codes are in proximity to NY City, but they have small populations.

Constructed power-law network reveals hubs (high degree nodes). Most zip codes among the top 20 by the highest node degree are located in Bronx, NY City, and Brooklyn area. This finding confirms that most hubs are in high populated areas in big cities. Detailed analysis of 3 zip codes with the highest node degrees show that all 3 of them have significantly lower average income than the state income. Zip code 10465, has a predominantly white and Hispanic population, but the average age was higher than the state average (40.7, vs 38.4). Other 2 zip codes (11226, 10027) had a predominantly black population, followed by Hispanic and white. Both areas had higher percentage of female population than the state average percentage of females. Although we presented demographic data for the top 3 zip codes (in cases of the highest numbers of hospitalized Influenza cases or highest node degrees) the trend in the top 20 zip codes in both cases was similar. Data reveal that all zip codes with a higher number of hospitalized Influenza cases, or higher node degrees in the correlation network, have significantly lower average income than the rest of the population in NY state. Also, most of these top zip codes show a younger population predominantly black or Hispanic, as well as higher than the state average of foreign born population. A common and important characteristic of zip codes with high node degrees is that almost all of them have a significantly higher percentage of the population without health insurance than the state average. In case, if the white population

was the majority, those zip codes had higher average age, than the average age in NY state. Findings of the social network analysis were in alignment with findings shown on heatmaps and in 4 tables. All zip codes with the highest node degrees were also among the top zip codes with the highest number of hospitalized influenza cases or the highest number of people living in those zip codes. Discoveries of our social network research could suggest some obvious measures that could lead to a lower number of complicated flu cases that needed hospitalizations. The Discovery that most of zip codes among the top 20 with high node degrees had lower income and a high percentage of foreign born population suggests that these issues should be addressed. Finding that most of the zip codes had a high percentage of uninsured population suggests that health insurance affordability is a very important factor and measures that will increase the number of insured people should be applied. The next issue that arises in areas with lower incomes and a higher number of uninsured people is the availability and affordability of primary care physicians, which also needs to be addressed. More affordable primary care medical offices, that accept patients without insurance or with not so good insurance coverage are needed in these zip codes. Further, a lot of areas have hospitals that are not well ranked, which can be managed by providing more resources to these hospitals. Additional problems are that hospitals in many cases are concentrated in medical centers and there are not enough hospitals in areas where people live. It's known that a lot of people without health insurance do not visit primary care offices and wait till they are very sick to go directly to Emergency rooms. During epidemics, a lot of medical personal are busy with other patients, so wait periods till patients are seen by doctors could be very long. During that time health conditions can

further deteriorate. These problems could be solved by employing more doctors, or trainees like residents, or Physician assistants, Nursing practitioners... An additional problem that people with significantly lower incomes than average and without healthcare insurance face is a lower percentage of preventive vaccinations and other preventive measures. Often, zip codes like this don't have enough pharmacy stores in the area which is the problem during epidemics, because people have to travel longer distances to purchase necessary medications.

Our social network study clearly identified hubs among zip codes, that need some of the suggested measures to improve prevention and treatment that will decrease the number of hospitalized cases. This study of social networks provides a wealth of information for understanding the influence of population sizes and demographics, in particular zip codes, on the spread of influenza viruses. We used a novel methodology to construct heatmaps and the network representation of hospitalized flu patients in the state of NY (2003-2012), based on the correlation among zip codes. Results of this study have important implications for predicting the geographical spread of hospitalized cases of influenza and prioritizing some of the suggested public health measures. Results can help adequate planning of resources for infectious disease outbreaks and their efficient control, as well as planning of hospital resources for more severe cases in the future.

4.6. Conclusion

Our research brings together medicine, biomedical informatics, computer science and social network science, in an attempt to explain the distribution of flu infections with

complications that required hospital admissions. The desire to have realistic networks based on the spatial distribution of complicated cases, for entire populations, provides important insights into how the size of the population and demographics, influence the distribution of influenza. The future research framework in this field would allow for different networks (from different times or different locations) to be compared. It will be important, for further development of the network science and its ability to analyze the spread of Influenza, to have effective data collecting protocols, and to use the statistical techniques to analyze collected data.

Our research was conducted on HCUP data, and we recommend further study of the geographic distribution of Influenza on more different datasets in order to improve understanding of the geographical distribution of hospitalized influenza patients and what demographic and socio-economic factors contribute to that distribution.

CHAPTER 5

COMORBIDITY NETWORK ANALYSIS AND GENETICS OF COLORECTAL CANCER

5.1. Introduction

There are more than 140,000 new cases and 51,000 deaths from colorectal cancer (CRC) in the U.S. each year.^{163,164} CRC is the third most common cancer in the U.S. and the second leading cause of cancer death. It affects men and women almost equally. Incidence, mortality and survival rates for CRC have regional variations and change over time.¹⁶⁵ CRC is mainly a disease of developed countries where it accounts for over 63% of all cases. It ranges from more than 40 per 100,000 people in the U.S., Australia, and Western Europe to less than 5 per 100,000 in Africa and parts of Asia.¹⁶⁵ The incidence of CRC is gradually decreasing in the U.S., mostly due to cancer screening and early detection of precancerous polyps. CRC survival is highly dependent on stage of cancer at diagnosis, and ranges from a 90% 5-year survival rate for early stages, to 10% for cancer with metastases.¹⁶⁵ The treatment of CRC has improved considerably in recent years. Better therapies have resulted in prolonged survival for patients with CRC.¹⁶⁶ Patients with CRC usually present in the older age group, with multiple comorbidities. About 59% of patients with CRC had one

comorbidity, and about 19% of patients had 4 or more comorbidity conditions.¹⁶⁷ Comorbidities are strong prognostic factors of survival in CRC patients in addition to sociodemographic and cancer characteristics. Early identification and management of comorbidities could help to optimize care for CRC patients.¹⁶⁸

Comorbidity network analyses can help to understand illness progression.¹⁶⁹⁻¹⁷¹ A social network analysis method is proposed to represent the progression of cardiovascular diseases (CVD) in patients with Diabetes Mellitus Type 2 (T2DM).¹⁷² A social network was developed to help understand which comorbidities have a higher influence on T2DM progression.¹⁷⁰ Many genetic networks are readily available. Hidalgo and Barabasi described the use of networks to integrate different genetic, proteomic, and metabolic datasets as a viable path toward elucidating the origins of specific diseases.¹⁶⁹ Numerous factors suggest a genetic contribution to CRC such as a family history of CRC or polyps.¹⁷³ A significant expansion of the genetic understanding of colonic carcinogenesis in the last 30 years occurred.¹⁷⁴⁻¹⁷⁶ The American College of Medical Genetics and Genomics has published guidelines for evaluating patients with CRC, to identify individuals whose clinical findings require referral for genetics consultations.¹⁷⁷

The objective of our research was to discover the most common comorbidities in different stages of CRC on the HCUP SID California database using network science. We used discovered comorbidities to identify genes associated with CRC, and comorbidity diseases. The text mining tool BeFree was employed to extract relationships between CRC and genes from PubMed.¹⁷⁸ BeFree consists of the Biomedical Named Entity Recognition (BioNER) module based on dictionaries using fuzzy and pattern matching methods to find and

uniquely identify entity mentioned in the literature, and a module for Relation Extraction (RE) based on Support Vector Machine (SVM).¹⁷⁸

The next objective of our study was to analyze associations between CRC and comorbidities to genes in abstracts indexed in the PubMed database. Identification of relationships between comorbidity diseases and shared genes can have important implications on early discovery and outcomes of cancer. We view genes and comorbidities as interconnected risk factors for CRC. Findings from the PubMed text mining were compared with results from expert curated sources. We used DisGeNET as the expert source to validate the findings of the text mining of PubMed. This is one of the largest collections of genes involved in human diseases. DisGeNET integrates data from expert curated repositories, the GWAS catalog, animal models, and scientific publications.¹⁷⁹

5.2. Materials and methods

We performed analyses of comorbidities associated with CRC on the HCUP, SID California inpatient database (ICD-9/10 codes format) which includes the largest collection of longitudinal hospital discharge data in the U.S.¹⁸⁰ We analyzed data for a period of 9 years (2003-2011). Diseases were represented with ICD-9 codes in this period. We used ICD-9 codes 153 and 154 to designate presence of CRC. SQL queries and Python code were created to extract comorbidities of CRC. We share our Python code on this link: <https://github.com/martinpavlovski/cancer-comorbidity-analysis>. We completed analyses separately for patients older than age 50, and patients younger than age 50. In the group of patients older than 50, the comorbidities associated with CRC were downloaded and

divided into 2 groups. We included all patients who had CRC without diagnosed metastases into the early stages of cancer and patients with already diagnosed metastases into advanced stages. We used ICD-9 codes 196, 197, and 198 to designate presence of metastases. The HCUP database contains only diagnoses of CRC without specifications of TNM stages, and it contains the information about the presence of metastases, which limited our comorbidity analysis to two groups: without and with diagnosed metastases. We could not analyze CRC and comorbidities by TNM staging system, because that information is not contained in the HCUP database. Regarding the group of patients age 50 and younger, the number of patients identified in the HCUP database, for the studied period, was relatively small and we couldn't perform separate analyses for patients with and without metastases. We calculated prevalence of comorbidities separately for males and females for this age group.

In the group of patients older than age 50, data were analyzed separately for the early and late stages of CRC and separately for males and females. Our comprehensive approach comprises the creation of ranked lists of comorbidities using frequencies of their occurrence, analysis of prevalence of comorbidities, construction of comorbidity networks, and calculation of centrality measures to estimate the significance of comorbidities. We used the following formula to calculate the prevalence of development of comorbidities of CRC:

$$P_i = \frac{n_i}{T} \quad (5)$$

P_i is the prevalence, n_i is the number of patients with comorbidity, and T is the total number of patients with CRC in the HCUP dataset.

In order to construct comorbidity networks, we had to determine the strength of comorbidity relations among diseases. Two types of comorbidity networks were constructed, one based on ϕ -correlation and the other based on Relative Risk (RR). These two comorbidity measures were used to quantify the strength of relations between the two comorbidity diseases.¹⁶⁹ The strength of relations was calculated between each pair of top comorbidities. CRC was not included in these calculations. The RR of observing a pair of diseases i and j affecting the same patient is given by the following formula:

$$RR_{ij} = \frac{C_{ij}N}{P_i P_j} \quad (6)$$

where C_{ij} is the number of patients affected by both diseases, N is the total number of patients in the population, and P_i and P_j are the prevalence of diseases i and j .

ϕ -correlation is computed as:

$$\phi_{ij} = \frac{C_{ij}N - P_i P_j}{\sqrt{P_i P_j (N - P_i)(N - P_j)}} \quad (7)$$

RR overestimates relationships involving rare diseases and underestimates highly prevalent comorbidities. ϕ -correlation overestimates comorbidities between diseases of similar prevalence but underestimates the comorbidity between rare diseases.¹⁶⁹ We constructed two types of networks, separately for each measure. In both models, the top 100 comorbidities from the rank lists were used to construct adjacency matrix $A[ij]$, which encodes whether and how a pair of nodes is connected. 86 % comorbidity conditions from the ranked list in females and 87% comorbidities in males in early stages, as well as 84%

comorbidity conditions in females, and 85% comorbidities in males in advanced stages of CRC were present in only 1 patient. It is difficult to consider any disease as a comorbidity of CRC if that disease appears in only 1 patient out of 30,000 patients. We opted to involve comorbidities that appear in about 5% of patients diagnosed with CRC in our network analysis, which turned out to be approximately 100 patients in each of the experimental settings. We created a power-law type of network. A Kolmogorov–Smirnov statistical test, at the significance level of 0.05, was conducted to determine if the networks follow power laws. We constructed a signed correlation network from the adjacency matrix and performed power transformation (normalization) of correlations.¹⁸¹ For normalization, we used a signed network normalization $|(\text{corMatrix}+1)/2|^\beta$ for a 100 x 100 adjacency matrix. By raising the absolute value of the correlation to a power $\beta \geq 1$ (soft threshold), signed correlation networks emphasize higher correlations versus lower correlations. We tested β -values between 1 and 10 and chose $\beta = 2$ for ϕ -correlation and $\beta = 5$ for RR. These β -values were selected because they provide the best visualization of networks. We used the correlation matrix after the transformation (normalization) as adjacent matrices to plot the network. Comorbidity networks were constructed in R program, using WGCNA, Statnet, and ggplot2 packages. Centrality measures (degrees of nodes and betweenness) were calculated to analyze the significance of each of the comorbidity. The significance of a comorbidity disease in the network is characterized by its node degree or its betweenness. The node degree represents the number of links the node has to other nodes in the network. The betweenness is the number of the shortest paths that pass through the node. We used the following formula to calculate the betweenness.¹⁸²

$$g(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}} \quad (8)$$

Where $g(v)$ is the betweenness of node v , σ_{st} is the total number of shortest paths from node s to node t , and $\sigma_{st}(v)$ is the number of those paths that pass through v .

Furthermore, we leveraged lists of ranked comorbidities of CRC for text mining of PubMed, to identify genes associated with CRC and comorbidities. The BeFree text mining system was utilized to extract associations between CRC, ranked comorbidity diseases, and genes by using morpho-syntactic information of the text. We also extracted relationships between genes and diseases from expert curated sources (DisGeNET).

We compared a proportion of gene-comorbidity disease associations discovered in PubMed with data collected in DisGeNET. Genes most often described in association with CRC were further matched with comorbidities from the ranking list. The number of abstracts in PubMed where these associations were described was counted. Lists of numbers of PubMed abstracts that show associations of the six most common genes and comorbidities of CRC were created to show the significance of these associations.

5.3. Results

5.3.1 Comorbidity analysis

In the group of patients older than age 50, the total number of females in the HCUP SID California database (2003-2011) with early stages of CRC was 31,503, and males 30,870. The total number of patients with the advanced stage of CRC was: females – 26,906 and

males – 27,454. We conducted analyses of comorbidities in patients with early stages of CRC and identified 5,786 different comorbidity conditions in females and 5,607 in males. In patients with the advanced stages of CRC, we discovered 5,609 different comorbidity conditions in females and 5,427 in males. Some of the top comorbidity conditions are specific to patients with CRC, but most conditions are similar to those in the general population.

Table 5.1. shows ranked comorbidities in females, older than age 50, in early stages (table 1a) and in advanced stages (table 1b) of CRC. The table also shows the prevalence of particular comorbidities according to the HCUP data. The top-ranked comorbidity in both stages is essential hypertension, with a very similar prevalence in both stages. Hyperlipidemia, history of malignant disease, anemia diabetes mellitus type 2, coronary artery sclerosis, paralytic ileus, malnutrition, acute kidney diseases, an encounter of palliative care, the absence of parts of large or small intestines have a few percentage points higher prevalence in advanced stages than in early stages of cancer, which can be attributed to aging.

| | 1a | | 1b | |
|----|---|------------|---|------------|
| | Females – Early stages | | Females – Advanced stages | |
| | Diagnosis | Prevalence | Diagnosis | Prevalence |
| 1 | Essential hypertension | 0.63 | Essential hypertension | 0.60 |
| 2 | Anemia | 0.33 | Hyperlipidemia | 0.41 |
| 3 | Hyperlipidemia | 0.32 | Personal history of tobacco use | 0.38 |
| 4 | History of malignant Neo of large bowel | 0.30 | History of malignant Neo of large bowel | 0.33 |
| 5 | Urinary tract infection | 0.26 | Anemia | 0.33 |
| 6 | Hypopotassemia | 0.24 | Diabetes mellitus type II | 0.27 |
| 7 | Diabetes mellitus type II | 0.23 | Coronary atherosclerosis | 0.26 |
| 8 | Esophageal reflux | 0.22 | Paralytic ileus | 0.23 |
| 9 | Personal history of tobacco use | 0.20 | Atrial fibrillation | 0.22 |
| 10 | Hypothyroidism | 0.20 | Congestive heart failure | 0.22 |
| 11 | Congestive heart failure | 0.19 | Esophageal reflux | 0.21 |
| 12 | Paralytic ileus | 0.19 | Hypothyroidism | 0.20 |

| | | | | |
|----|--|------|--|------|
| 13 | Atrial fibrillation | 0.18 | Acute kidney failure | 0.19 |
| 14 | Coronary atherosclerosis | 0.18 | Pure hypercholesterolemia | 0.18 |
| 15 | Dehydration | 0.17 | Digestive system complications | 0.18 |
| 16 | Diverticulosis of colon | 0.17 | Chronic airway obstruction | 0.17 |
| 17 | Pure hypercholesterolemia | 0.16 | Hypopotassemia | 0.17 |
| 18 | Osteoporosis without path. fracture | 0.16 | Dehydration | 0.16 |
| 19 | Osteoarthritis | 0.15 | Urinary tract infection | 0.16 |
| 20 | Pneumonia | 0.14 | Pneumonia | 0.15 |
| 21 | Hyposmolality and/or hyponatremia | 0.14 | Diverticulosis of colon | 0.15 |
| 22 | Acute posthemorrhagic anemia | 0.14 | Tobacco use disorder | 0.15 |
| 23 | Iron deficiency anemia | 0.14 | Hypertensive chronic kidney disease | 0.15 |
| 24 | Chronic airway obstruction | 0.14 | Benign neoplasm of colon | 0.15 |
| 25 | Acute kidney failure | 0.14 | Acute posthemorrhagic anemia | 0.14 |
| 26 | Overweight, obesity | 0.14 | Old myocardial infarction | 0.14 |
| 27 | Iron deficiency anemia due to blood loss | 0.14 | Overweight, obesity | 0.14 |
| 28 | Depressive disorder | 0.13 | Hyposmolality and/or hyponatremia | 0.14 |
| 29 | Digestive system complications | 0.13 | Cardiac dysrhythmias | 0.14 |
| 30 | Benign neoplasm of colon | 0.12 | Iron deficiency anemia due to blood loss | 0.14 |

Table 5.1. Comorbidities in early and advanced stages of CRC in females, older than age 50. 1a: Ranked comorbidities occurred in the early stages. 1b: Ranked comorbidities occurred in the advanced stages of CRC. The prevalence of the occurrence of comorbidities is shown.

In table 5.2. we present ranked comorbidities of CRC in males, in the early stages (table 2a) and in the advanced stages (table 2b). The top comorbidity in both stages of cancer is Essential hypertension, which is the same finding as in female patients. The prevalence of this disease is similar in both stages in females and males and varies between 0.58 and 0.63. Anemia, smoking history, hyperlipidemia, coronary atherosclerosis, congestive heart failure, atrial fibrillation and other cardiac dysrhythmias, esophageal reflux, history of malignant neoplasms of prostate, lungs, colon, and other cancers, acute kidney failure, dehydration, colostomy status, pneumonia, and acute respiratory failure happened more often in advanced stages of CRC. We can notice that majority of diseases have a few points higher prevalence in advanced stages than in early stages of CRC, which can be explained by aging and the progress of CRC.

| 2a | | | 2b | |
|----------------------|---|------------|---|------------|
| Males – Early stages | | | Males – Advanced stages | |
| | Diagnosis | Prevalence | Diagnosis | Prevalence |
| 1 | Essential hypertension | 0.60 | Essential hypertension | 0.58 |
| 2 | Hyperlipidemia | 0.34 | Personal history of tobacco use | 0.38 |
| 3 | Personal history of tobacco use | 0.31 | Anemia | 0.37 |
| 4 | History of malignant neo of large intestine | 0.30 | Hyperlipidemia | 0.33 |
| 5 | Anemia | 0.29 | Pneumonia | 0.28 |
| 6 | Diabetes mellitus type II | 0.26 | Diabetes mellitus type II | 0.25 |
| 7 | Paralytic ileus | 0.24 | Acute kidney failure | 0.25 |
| 8 | Coronary atherosclerosis | 0.24 | Dehydration | 0.25 |
| 9 | Atrial fibrillation | 0.20 | Chronic airway obstruction | 0.24 |
| 10 | Congestive heart failure | 0.19 | Coronary atherosclerosis | 0.24 |
| 11 | Esophageal reflux | 0.19 | Tobacco use disorder | 0.20 |
| 12 | Acute kidney failure | 0.18 | Hyposmolality and/or hyponatremia | 0.20 |
| 13 | Benign hypertrophy of prostate | 0.18 | Urinary tract infection | 0.20 |
| 14 | Pure hypercholesterolemia | 0.18 | Atrial fibrillation | 0.19 |
| 15 | Digestive system complications | 0.17 | Hypopotassemia | 0.18 |
| 16 | Chronic airway obstruction | 0.16 | Esophageal reflux | 0.18 |
| 17 | Dehydration | 0.16 | Congestive heart failure | 0.17 |
| 18 | Hypopotassemia | 0.16 | Encounter for palliative care | 0.17 |
| 19 | Pneumonia | 0.15 | Anemia in neoplastic disease | 0.17 |
| 20 | Urinary tract infection | 0.15 | Malignant neoplasm of prostate | 0.16 |
| 21 | Diverticulosis of colon | 0.14 | Unspecified protein-calorie malnutrition | 0.16 |
| 22 | Tobacco use disorder | 0.14 | Malignant neo of bronchus and lung | 0.16 |
| 23 | Chronic kidney disease | 0.13 | Acute respiratory failure | 0.16 |
| 24 | Benign Neo of colon | 0.13 | Pure hypercholesterolemia | 0.16 |
| 25 | Acute posthemorrhagic anemia | 0.13 | Constipation | 0.15 |
| 26 | Overweight / Obesity | 0.12 | Irradiation, presenting hazards to health | 0.15 |
| 27 | Hyposmolality and/or hyponatremia | 0.12 | Hypertrophy (benign) of prostate | 0.15 |
| 28 | Old myocardial infarction | 0.12 | Hearing loss | 0.15 |
| 29 | Cardia dysrhythmia | 0.11 | Malignant neopl without specified site | 0.14 |
| 30 | Iron deficiency anemia due to blood loss | 0.11 | History of malignant neopl of prostate | 0.14 |

Table 5.2. Comorbidities in early and advanced stages of CRC in males, older than age 50. 2a: Ranked comorbidities occurred in the early stages. 2b: Ranked comorbidities occurred in the advanced stages of CRC. The prevalence of the occurrence of comorbidities is shown.

In table 3, we present ranked comorbidities of CRC and their prevalence for patients age 50 and younger, separately for males (3a) and females (3b). In this age group, we identified 3794 male patients with CRC and 3640 female patients. HCUP contains very small numbers of patients without metastases in this age group, in the studied period, so we analyzed all patients together, regardless of metastases. Males in this age group had a higher prevalence of essential hypertension, tobacco use disorder, hyperlipidemia, diabetes

mellitus, and sepsis than females. On the other side, female patients had a higher prevalence of anemia, hypopotassemia, urinary tract infections, obesity, and depressive disorder than male patients.

| | 3a | | 3b | |
|----|---|------------|--|------------|
| | Males younger than 50 | | Females younger than 50 | |
| | Diagnosis | Prevalence | Diagnosis | Prevalence |
| 1 | Essential hypertension | 0.32 | Anemia | 0.32 |
| 2 | Anemia | 0.28 | History of malignant neo of large intestine | 0.30 |
| 3 | History of malignant neo of large intestine | 0.27 | Essential hypertension | 0.27 |
| 4 | Paralytic ileus | 0.22 | Hypopotassemia | 0.23 |
| 5 | Tobacco use disorder | 0.21 | Paralytic ileus | 0.21 |
| 6 | Hypopotassemia | 0.17 | Urinary tract infection | 0.20 |
| 7 | Dehydration | 0.16 | Dehydration | 0.18 |
| 8 | History of malignant neo of rectum | 0.15 | Anemia in neoplastic disease | 0.15 |
| 9 | Hyposmolality and/or hyponatremia | 0.15 | Tobacco use disorder | 0.15 |
| 10 | Esophageal reflux | 0.14 | Esophageal reflux | 0.14 |
| 11 | Colostomy status | 0.14 | Acquired absence of intestine (large, small) | 0.14 |
| 12 | Diabetes mellitus | 0.14 | Obesity | 0.14 |
| 13 | Anemia in neoplastic disease | 0.14 | Personal history of malignant neo of rectum | 0.14 |
| 14 | Acute kidney failure | 0.14 | Depressive disorder | 0.14 |
| 15 | Hyperlipidemia | 0.14 | Personal history of irradiation | 0.13 |
| 16 | Personal history of irradiation | 0.13 | Peritoneal adhesions | 0.13 |
| 17 | Acquired absence of intestine | 0.12 | Iron deficiency anemia | 0.13 |
| 18 | Peritoneal adhesions | 0.12 | Nausea with vomiting | 0.13 |
| 19 | Intestinal obstruction | 0.12 | Hyposmolality and/or hyponatremia | 0.12 |
| 20 | Urinary tract infection | 0.11 | Constipation | 0.12 |
| 21 | Protein-calorie malnutrition. | 0.11 | Intestinal obstruction | 0.12 |
| 22 | Postoperative wound infection | 0.11 | Diabetes mellitus | 0.11 |
| 23 | Iron deficiency anemia | 0.11 | Colostomy status | 0.11 |
| 24 | Obesity | 0.10 | Anxiety | 0.11 |
| 25 | Pneumonia | 0.10 | Protein-calorie malnutrition. | 0.10 |
| 26 | Hearing loss | 0.10 | Diarrhea | 0.10 |
| 27 | Constipation | 0.10 | Hyperlipidemia | 0.10 |
| 28 | Chronic blood loss anemia | 0.10 | Chronic blood loss anemia. | 0.10 |
| 29 | Diarrhea | 0.10 | Family history of malignant neo of GI tract | 0.10 |
| 30 | Sepsis | 0.10 | Asthma | 0.10 |

Table 5.3. Ranked comorbidities in CRC patients, age 50 and younger: Males (3a), and females (3b). The prevalence of the occurrence of comorbidities is shown.

5.3.2. Comorbidity networks

Comorbidity networks were constructed for patients older than age 50, separately for early and advanced stages of CRC, for males and females. Nodes represent diseases (comorbidities) and links connect comorbidities according to distance measures described above (RR, ϕ -correlation). Calculated node degrees and betweenness show the most connected (significant) comorbidities in each of eight networks. In Figure 5.1 we present two networks for the early stages of CRC in females.

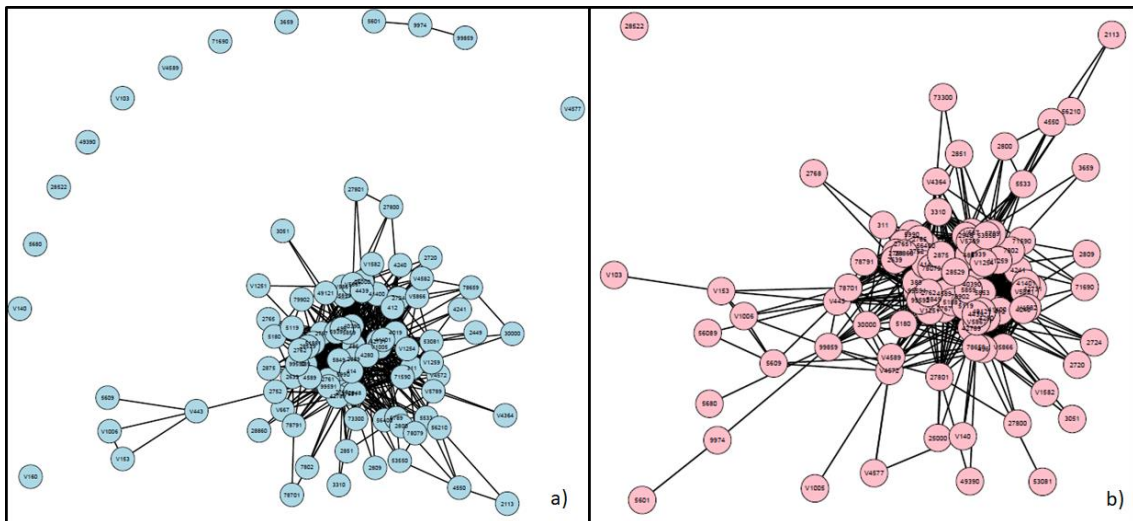


Figure 5.1. Networks of comorbidities in the early stages of CRC in females older than age 50. Comorbidity networks, based on: a) ϕ -correlation and b) RR in females. Nodes represent the top 100 comorbidities (ICD9 codes). Correlations greater than 0.32 ($\beta = 2$) and RR greater than 5.99 ($\beta = 5$) were applied for construction of edges.

Considering ϕ -correlation, the following comorbidity conditions had the highest node degrees in the network 1a: urinary tract infection (UTI)-65, congestive heart failure (CHF)-63, coronary artery disease (CAD)-52, anemia-50, and acute kidney failure (AKF)-49. Comorbidities that had the highest betweenness in the network 1a were: UTI-329, CHF-275, CAD-254, colostomy status-248, anemia-153, and dehydration-140. Based on RR

measures, comorbidities with the highest node degrees in the network 1b were: anemia-61, obstructive chronic bronchitis (OCB)-57, hypoxemia-56, hyperkalemia-55, and peripheral vascular disease (PVD)-53. The following conditions had the highest betweenness in the network 1b: postoperative infection-301, colostomy status-263, gastro-duodenitis-256, chest pain-242, and OCB-221.

Figure 5.2 shows two networks (ϕ -correlations and RR) for the advanced stages of CRC in females.

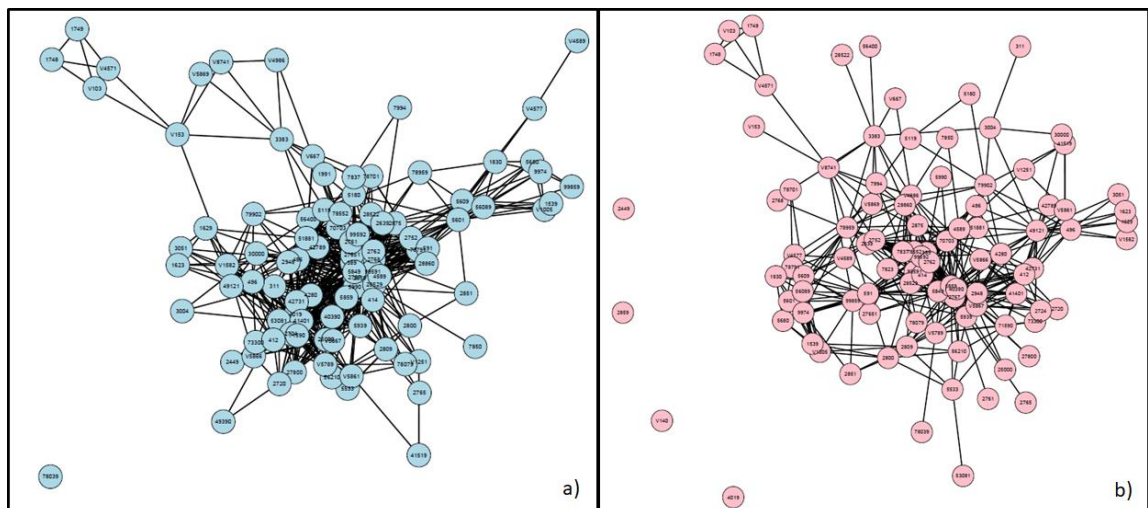


Figure 5.2. Networks of comorbidities in the advanced stages of CRC in females older than age 50. Comorbidity networks, based on: a) ϕ -correlation and b) RR in females. Nodes represent the top 100 comorbidities (ICD9 codes). Correlations greater than 0.30 ($\beta = 2$) and RR greater than 8.99 ($\beta = 5$) were applied for construction of edges.

Comorbidities with the highest node degrees in the network 2a (ϕ -correlation) were: UTI-57, anemia-52, CHF-47, and hypokalemia-40. The highest betweenness were: UTI-467, anemia-437, dehydration-420, history of irradiation-405, and paralytic ileus-329. Considering RR measure, the highest node degrees in the network 2b were: chronic kidney

disease-56, pressure ulcer lower back-50, hypertensive chronic kidney disease-48, hyperkalemia-42, sepsis-42, and long-term use of insulin-41. The highest betweenness in the network 2b were: chronic kidney disease-494, OCB-421, history of chemotherapy-368, and hypoxemia-307.

Next, we present two networks for the early stages of CRC in males (figure 5.3). We used the same metrics as in females, ϕ -correlations Fig 5.3a. and RR – Fig 5.3b.

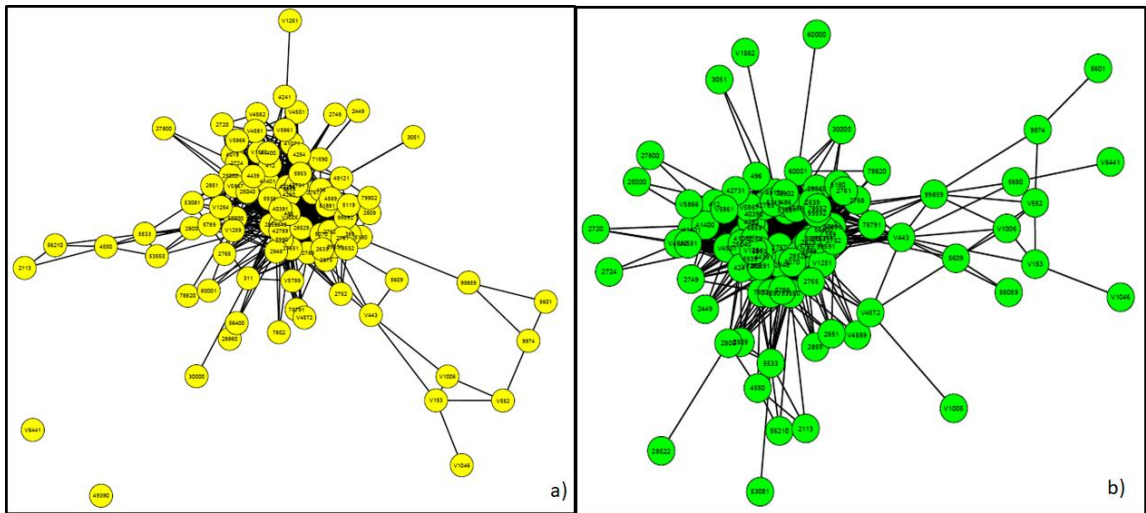


Figure 5.3. Networks of comorbidities in the early stages of CRC in males older than age 50. Comorbidity networks, based on: a) ϕ -correlation and b) RR. Nodes represent the top 100 comorbidities (ICD9 codes). Correlations greater than 0.32 ($\beta = 2$) and RR greater than 8.99 ($\beta = 5$) were applied for edges.

Comorbidities with the highest node degrees in the network 3a (ϕ -correlation) were: CHF-64, AKF-59, anemia-57, UTI-55, and CAD-50. The highest betweenness in the network 3a were: UTI-492, anemia-431, CHF-394, AKF-350, and hearing loss-277. The highest node degrees in the network 3b. (RR) were: anemia-63, hyperkalemia-60, OCB-58, chronic kidney disease-55, primary cardiomyopathies-55, PVD-53, and sepsis-50. The highest

betweenness in the network 3b. were: colostomy status-418, anemia-294, and postoperative infection-267.

We created two networks for advanced stages of CRC in males (figure 5.4). We used the same metrics as in females, ϕ -correlation (Fig 5.4a), and RR (Fig 5.4b).

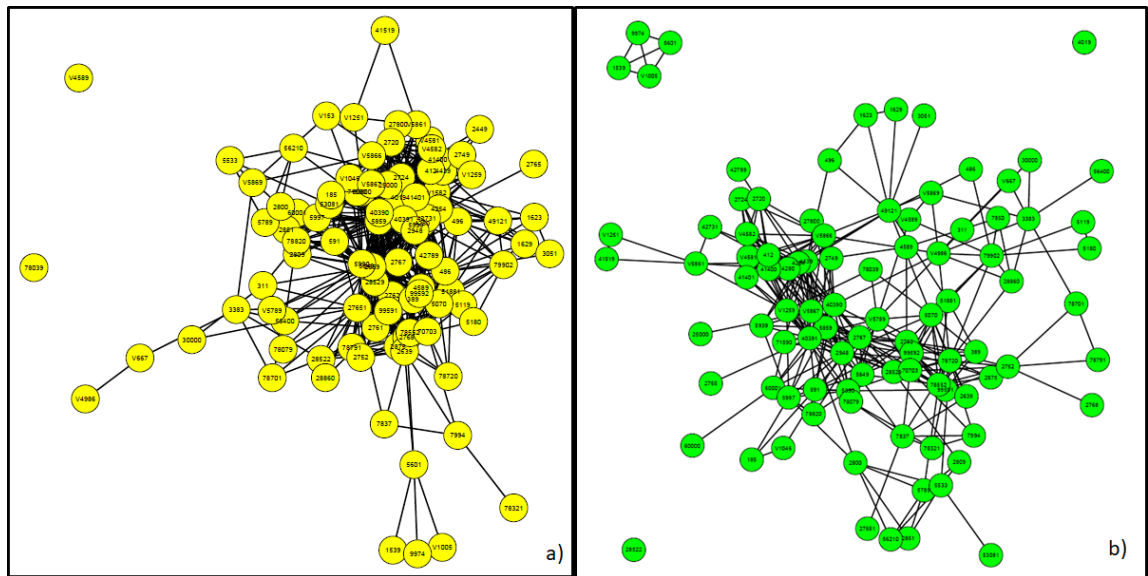


Figure 5.4. Networks of comorbidities in the advanced stages of CRC in males older than age 50. Comorbidity networks, based on: a) ϕ -correlation and b) RR. Nodes represent the top 100 comorbidities (ICD9 codes). Correlations greater than 0.30 ($\beta = 2$) and RR greater than 8.99 ($\beta = 5$) were used for edges.

Conditions with the highest node degrees in the network 4a (ϕ -correlation) were: anemia-49, CHF-48, AKF-44, and CAD-37. The highest betweenness in the network 4a: anemia-414, dehydration-401, CHF-386, AKF-310, UTI-301, and protein-calorie malnutrition-287. The highest node degrees in the network 4b (RR): Primary cardiomyopathies-42, chronic kidney disease-40, PVD-39, and persistent mental disorders-39. The highest

betweenness in the network 4b: Dysphagia-480, primary cardiomyopathies-410, OCB-384, hypoxemia-294, and do not resuscitate status-293.

5.3.3. Genes associated with CRC and comorbidities

Ranked lists of comorbidities of CRC were used for text mining of PubMed and expert curated sources, specifically DisGeNET, which is a comprehensive platform integrating information on human disease-associated genes. The BeFree text mining system was applied for the extraction of associations between genes, comorbidities, and CRC from PubMed.

PubMed indexes more than 170,000 publications on CRC (MeSH Major topic search). The Befree data mining system extracted 1,937 different genes associated with CRC from PubMed. (The list of 1,937 provided as Suppl.1) We found 150 different genes associated with CRC in DisGeNET. Ninety-six genes are present in both sources. Most gene–CRC associations in PubMed were described in only one abstract (1160). The most often mentioned genes associated with CRC were: TP53 (241 abstracts in PubMed), APC (115), and KRAS (106). All 3 genes had DisGeNET scores of 0.5. The DisGeNET score for gene-disease associations takes into account the number and type of sources (level of curation), and the number of publications supporting the association. The scores range from 0 to 1. Two more genes had a DisGeNET score of 0.5: MLH1 (98 abstracts) and TGFBR2 (18). One gene (PPARG) had a DisGeNET score of 0.6, and it's mentioned 43 times in PubMed abstracts.

In table 5.4, we present associations of the APC gene with comorbidities of CRC. The APC gene is located on the long arm of chromosome 5. This gene signals the production of the APC protein, a tumor suppressor, which slows down the growth and division of cells, and controls how cells attach and move.¹⁸³ Mutations in the APC gene have been associated with several cancers including familial adenomatous polyposis and CRC. APC gene has also been described in people with a benign desmoid tumor, primary macronodular adrenal hyperplasia, Turcot syndrome, brain neoplasm (medulloblastoma), stomach (gastric) cancers, prostatic cancer, etc. Our text mining of PubMed extracted these tumors, as well as a few types of leukemia associated with the APC gene. BeFree text mining system also extracted a few comorbidity conditions such as Amyloidosis, AD, Systemic Lupus erythematosus, Diabetes Mellitus, and few more conditions as potentially associated with the APC gene.

| No | Diseases | PubMed abstracts |
|----|--|------------------|
| 1 | Malignant neoplasm of stomach | 20 |
| 2 | Malignant neoplasm of prostate | 18 |
| 3 | Amyloidosis | 15 |
| 4 | Barrett's esophagus | 8 |
| 5 | Malignant neoplasm of pancreas | 8 |
| 6 | AD | 6 |
| 7 | Malignant neoplasm of ovary | 6 |
| 8 | Malignant neoplasm of esophagus | 5 |
| 9 | Malignant neoplasm of bladder | 5 |
| 10 | Brain neoplasm | 4 |
| 11 | Malignant neoplasm of intestinal tract, part unspecified | 4 |
| 12 | Systemic lupus erythematosus | 4 |
| 13 | Ulcerative colitis | 4 |
| 14 | Rheumatoid arthritis | 3 |
| 15 | Burkitt's tumor or lymphoma | 3 |
| 16 | Megakaryocytic leukemia | 3 |
| 17 | Septicemia | 3 |
| 18 | Hypogammaglobulinemia | 3 |
| 19 | Diabetes mellitus | 3 |

| | | |
|----|---|---|
| 20 | Sebaceous cyst | 3 |
| 21 | Graft-versus-host disease | 3 |
| 22 | Secondary and unspecified malignant neoplasm of lymph nodes | 3 |
| 23 | Candidiasis of mouth | 3 |

Table 5.4. APC genes associated with CRC and comorbidities. Presented are numbers of abstracts in PubMed where these associations were described in 3 or more abstracts.

The TP53 gene is located on the short arm of chromosome 17. The TP53 gene regulates the production of a protein p53, a tumor suppressor, which prevents cells from growing and dividing without control.¹⁸⁴ Mutations in the TP53 gene are the most common genetic changes in human cancer. Mutations in the TP53 gene are associated with the risk of developing breast cancer, bladder cancer, cholangiocarcinoma, squamous cell carcinomas, bone cancer, sarcomas, lung cancer, melanoma, CRC, etc.

Our text mining findings contain tumors mentioned above, as well as few types of leukemia. Several comorbidity diseases are also described in associations with this gene, such as: Hodgkin's disease, ulcerative colitis, rheumatoid arthritis, Alzheimer's disease, etc.

Associations between CRC and comorbidities with gene TP53 are shown in table 5.5.

| No | Disease | PubMed abstracts |
|----|---|------------------|
| 1 | Malignant neoplasm of prostate | 263 |
| 2 | Lymphoid leukemia, chronic | 248 |
| 3 | Malignant neoplasm of ovary | 239 |
| 4 | Myeloid leukemia, acute | 192 |
| 5 | Malignant neoplasm of stomach | 167 |
| 6 | Malignant neoplasm of bladder | 159 |
| 7 | Malignant neoplasm of pancreas | 134 |
| 8 | Malignant neoplasm of cervix uteri | 130 |
| 9 | Leukemia of unspecified cell type | 112 |
| 10 | Brain neoplasm | 87 |
| 11 | Secondary and unspecified malignant neoplasm of lymph nodes | 82 |

| | | |
|----|--|----|
| 12 | Myelodysplastic syndrome, unspecified | 76 |
| 13 | Malignant neoplasm of esophagus | 76 |
| 14 | Multiple myeloma | 68 |
| 15 | Myeloid leukemia, chronic | 62 |
| 16 | Burkitt's tumor or lymphoma | 53 |
| 17 | Lymphoid leukemia, acute | 51 |
| 18 | Malignant neoplasm of mouth, unspecified | 51 |
| 19 | Hodgkin's disease | 50 |
| 20 | Barrett's esophagus | 38 |
| 21 | Ulcerative colitis | 36 |
| 22 | Malignant neoplasm of liver, secondary | 31 |
| 23 | Mantle cell lymphoma | 29 |
| 24 | Rheumatoid arthritis | 29 |
| 25 | Malignant neoplasm of liver | 27 |
| 26 | Malignant neoplasm of thyroid gland | 27 |
| 27 | Actinic keratosis | 25 |
| 28 | Unspecified viral hepatitis C | 23 |
| 29 | Leukemia of unspecified cell type, acute | 22 |
| 30 | Neurofibromatosis, type 1 [von Recklinghausen's disease] | 21 |
| 31 | Alzheimer's disease | 20 |

Table 5.5. TP53 genes associated with CRC and comorbidities. Presented are numbers of abstracts in PubMed where these associations were described in 20 or more abstracts.

The KRAS gene is located on the short arm of chromosome 12. It regulates synthesis of a K-Ras protein.¹⁸⁵ Mutations in the KRAS gene are associated with autoimmune lymphoproliferative syndrome, cholangiocarcinoma, acute myeloid leukemia, epidermal nevus, lung cancer, pancreatic cancer, CRC, as well as conditions such as intellectual disability, distinctive facial features, short stature, macrocephaly, heart defects, skin abnormalities, Noonan syndrome, etc. In table 5.6, associations between CRC, KRAS genes, and comorbidities are presented. Text mining of PubMed extracted associations of the KRAS gene with multiple tumors (predominantly pancreatic cancer) and several non-tumorous diseases. The KRAS gene was also extracted in associations with unspecified viral hepatitis C, chronic pancreatitis, ulcerative colitis, dengue, atrial premature beats, endometriosis, etc.

| No | Disease | PubMed abstracts |
|----|---|------------------|
| 1 | Malignant neoplasm of pancreas | 123 |
| 2 | Malignant neoplasm of liver, secondary | 27 |
| 3 | Malignant neoplasm of stomach | 26 |
| 4 | Unspecified viral hepatitis C | 25 |
| 5 | Malignant neoplasm of ovary | 17 |
| 6 | Secondary malignant neoplasm of lung | 14 |
| 7 | Secondary and unspecified malignant neoplasm of lymph nodes | 14 |
| 8 | Myeloid leukemia, acute | 10 |
| 9 | Malignant neoplasm of prostate | 8 |
| 10 | Chronic pancreatitis | 7 |
| 11 | Patent ductus arteriosus | 7 |
| 12 | Ulcerative colitis | 7 |
| 13 | Dengue | 6 |
| 14 | Atrial premature beats | 5 |
| 15 | Endometriosis | 5 |
| 16 | Leukemia of unspecified cell type | 5 |
| 17 | Malignant neoplasm of thyroid gland | 5 |
| 18 | Malignant neoplasm of biliary tract, part unspecified site | 5 |
| 19 | Multiple myeloma | 5 |
| 20 | Pleural effusion | 4 |
| 21 | Uterine leiomyoma | 4 |
| 22 | Lymphoid leukemia, acute | 4 |

Table 5.6. KRAS genes associated with CRC and comorbidities. Presented are numbers of abstracts in PubMed where these associations were described in 4 or more abstracts.

The MLH1 gene is located on the short arm of chromosome 3. The MLH1 gene encodes a protein that has a crucial role in repairing DNA errors.¹⁸⁶ Mutations in the MLH1 gene have been associated with constitutional mismatch repair deficiency syndrome, CRC, brain cancer, leukemia, lymphoma, Lynch syndrome, neurofibromatosis type 1, cancers of the endometrium, ovaries, stomach, etc. In table 5.7, associations between CRC, MLH1 genes, and comorbidities found in PubMed abstracts are presented.

| No | Disease | PubMed abstracts |
|----|---|------------------|
| 1 | Malignant neoplasm of stomach | 38 |
| 2 | Malignant neoplasm of ovary | 24 |
| 3 | Leukemia of unspecified cell type | 10 |
| 4 | Malignant neoplasm of bladder | 7 |
| 5 | Malignant neoplasm of prostate | 7 |
| 6 | Ulcerative colitis | 6 |
| 7 | Secondary and unspecified malignant neoplasm of lymph nodes | 5 |
| 8 | von Recklinghausen's disease | 5 |
| 9 | Huntington's chorea | 3 |
| 10 | Nodular lymphoma | 3 |
| 11 | Malignant neoplasm of uterus, part unspecified | 3 |
| 12 | Malignant neoplasm of pancreas | 3 |
| 13 | Endometrial hyperplasia with atypia | 3 |

Table 5.7. MLH1 genes associated with CRC and comorbidities. Presented are numbers of abstracts in PubMed where these associations were described in 3 or more abstracts.

The TGFBR2 gene is positioned on the short arm of chromosome 3. It encodes transforming growth factor-beta (TGF- β) receptor type 2, which transduces a signal that prevents cells from growing and dividing too rapidly.¹⁸⁷ Mutations in the TGFBR2 gene have been identified in people with Marfan's syndrome (familial thoracic aortic aneurysm and dissection), with Loey's-Dietz syndrome type II, various cancers (CRC, and others). Our text mining of PubMed discovered these conditions in addition to Alzheimer's disease, asthma, ulcerative colitis, etc. In table 5.8 we present associations between CRC, TGFBR2 genes, and comorbidities.

| No | Disease | PubMed abstracts |
|----|---|------------------|
| 1 | Marfan's syndrome | 30 |
| 2 | Malignant neoplasm of stomach | 10 |
| 3 | Malignant neoplasm of prostate | 8 |
| 4 | Malignant neoplasm of pancreas | 7 |
| 5 | Aneurysm | 4 |
| 6 | Alzheimer's disease | 2 |
| 7 | Asthma | 2 |
| 8 | Ulcerative colitis | 2 |
| 9 | Patent ductus arteriosus | 2 |
| 10 | Ehlers-Danlos syndrome | 2 |
| 11 | Systemic sclerosis | 2 |
| 12 | Migraine | 2 |
| 13 | Malignant neoplasm of mouth, unspecified | 2 |
| 14 | Malignant neoplasm of gallbladder | 2 |
| 15 | Dissection of aorta, unspecified site | 2 |
| 16 | Secondary and unspecified malignant neoplasm of lymph nodes | 2 |

Table 5.8. TGFBR2 genes associated with CRC and comorbidities. Presented are numbers of abstracts in PubMed where these associations were described in 2 or more abstracts.

The PPARG gene is located on the short arm of chromosome 3. This gene encodes the peroxisome proliferator-activated receptor (PPAR) subfamily of nuclear receptors.¹⁸⁸ PPARs have been involved in the pathology of numerous diseases including obesity, atherosclerosis, prostatic cancer, lipodystrophy, glioma, obesity, insulin resistance, diabetes mellitus, dyslipidemia, hypertension and CRC. Our text mining findings correspond to this description. In table 5.9. we show associations between CRC, PPARG genes, and comorbidities.

| No | Disease | PubMed abstracts |
|----|--------------------------------|------------------|
| 1 | Obesity | 208 |
| 2 | Diabetes mellitus | 100 |
| 3 | Atherosclerosis | 37 |
| 4 | Dysmetabolic syndrome x | 35 |
| 5 | Alzheimer's disease | 34 |
| 6 | Malignant neoplasm of prostate | 33 |
| 7 | Hypertensive disease | 31 |
| 8 | Heart failure | 27 |

| | | |
|----|--|----|
| 9 | Huntington's chorea | 21 |
| 10 | Polycystic ovaries | 17 |
| 11 | Coronary atherosclerosis | 14 |
| 12 | Lipodystrophy | 14 |
| 13 | Malignant neoplasm of thyroid gland | 13 |
| 14 | Cardiovascular Diseases | 12 |
| 15 | Diabetic retinopathy | 12 |
| 16 | Malignant neoplasm of stomach | 12 |
| 17 | Diabetes with renal manifestations | 11 |
| 18 | Osteoarthritis, unspecified whether generalized or localized | 11 |
| 19 | Parkinson's disease | 11 |
| 20 | Amyotrophic lateral sclerosis | 10 |
| 21 | Malignant neoplasm of pancreas | 10 |

Table 5.9. PPARG genes associated with CRC and comorbidities on PubMed. Comorbidities that are associated with CRC and PPARG in 10 or more abstracts are given.

In table 5.10 we present 96 genes identified in both sources: PubMed and DisGeNET.

| | | | | | |
|---------|--------|-------|--------|----------|-----------|
| ALDH1B1 | CHEK2 | IGF2 | MTHFR | PTP4A3 | SULT1A1 |
| ALOX5 | CTNNB1 | IL1B | MUTYH | PTPN12 | TAGLN |
| APC | CXCL10 | IL32 | MYC | PTPRJ | TCF7L2 |
| ASCL2 | DMBT1 | JUN | NDRG2 | RAD54B | TGFBR2 |
| ATP7A | DNMT1 | KCNH2 | NFKB1 | RCOR1 | TNF |
| AXIN2 | DPYD | KDM1A | NOS2 | RECK | TNFRSF10A |
| BAX | EGFR | KRAS | NOTCH1 | S100A4 | TNFSF10 |
| BCL2 | ERBB2 | LEF1 | NOX1 | SERPINB5 | TP53 |
| BECN1 | FBP1 | LEP | NOX4 | SERTAD1 | TP63 |
| BIRC5 | GAST | LGR5 | ODC1 | SLC2A1 | TPM3 |
| BRAF | GCG | MCM2 | PCNA | SLC5A8 | TYMP |
| BRD4 | HES1 | MECOM | POLB | SLCO1B3 | TYMS |
| CBR1 | HMGCS2 | MIR98 | PPARG | SOD2 | VEGFA |
| CCAT1 | HSPB1 | MLH1 | PRKN | SPARC | WT1 |
| CCND1 | ICAM1 | MMP7 | PROM1 | SRC | XAF1 |
| CDKN1A | IFNG | MMP9 | PTGS2 | STAT3 | YBX1 |

Table 5.10. Ninety-six genes associated with CRC and present in both sources: PubMed and DisGeNET.

5.4. Discussion

Chronic diseases such as CRC are associated with many comorbidities and complications, which affect the quality of life and the prognosis of CRC. Certain comorbidities (dementia, depressive disorder, alcoholism) are associated with a delayed CRC diagnosis, while some chronic comorbidities (T2DM, CVD, Hypertension) that require frequent medical care are associated with earlier CRC detection.^{168,189} Patients with comorbidity often receive different CRC treatments (surgery, chemotherapy, radiation therapy) than patients without comorbidity.^{168,189}

Most published studies presented social network methods that include multiple diseases (phenotypes) not focusing on any particular disease as the main subject.^{169,190,191} In our research we implemented a social network analysis of comorbidities of one disease (CRC). This cancer is mostly diagnosed at older age, when comorbidities are commonly present and when they can be an important risk and prognostic factor of CRC. Our prevalence and network analyses of comorbidities of CRC show which specific comorbidities are more prevalent and significant. The results of our study could have a practical implementation in medicine, by providing the information on which comorbidity conditions should be looked for as highly expected, and consequently prevented or treated. The results of our study show that the prevalence of some characteristic comorbidities, such as diverticulosis of colon, history of malignant neoplasms of intestines, benign neoplasms of intestines, paralytic ileus, and few others are higher in CRC than in the general population. Comparing findings between groups of patients younger or older than age 50, we can notice that patients younger than age 50 have a significantly lower prevalence of essential

hypertension than patients older than age 50. Younger groups of patients don't have cardiovascular conditions such as coronary atherosclerosis, congestive heart failure, atrial fibrillation and other cardiac dysrhythmias, in the top 30 ranked comorbidities. They also don't have history of malignant neoplasms of prostate, lungs, colon, and other cancers, in the top 30 the most prevalent comorbidities. These results show that many comorbidities are influenced by aging of patients. Certain comorbidities such as history of malignant neoplasms of large bowel and rectum, paralytic ileus, intestinal obstruction, colostomy status, and peritoneal adhesions are present in both groups, younger and older than age 50, which confirms that some of comorbidities are influenced by development of CRC.

Constructed scale-free networks and calculated centrality measures show that comorbidities are interlinked beyond simple coincidence.^{169,190} Centrality measures (node degree, betweenness) discovered the presence of highly connected comorbidities (hubs). Highly connected comorbidities, such as CHF, CAD, OCB, CKD, AKF, anemia, hypoxemia, and a few others should be followed and treated. Some of the conditions revealed as important by RR metric were: anemia, hypoxemia, colostomy status, gastro-duodenitis, chest pain, pressure ulcer-lower back, hyperkalemia, sepsis, etc. ϕ -correlation revealed comorbidities that are highly prevalent and expected. Evaluation of centrality measures discovered that both betweenness and node degrees come as valuable yet different measures considering findings of the most significant comorbidities associated with CRC.

Results of the comorbidity study were used to carry out analysis of associations between genes, CRC, and comorbidities. We extracted 627 more genes (described in 2 or more

abstracts) on PubMed than on DisGeNET. Genetic findings could be used to recruit more individuals who would benefit from genetic testing and consultations. Recent studies incorporated comorbidities in addition to the genomic data to identify new disease-genes associations.^{171,172,190,191} If the same gene is linked to two different diseases, this is often an indication that the two diseases have a common genetic origin. Comorbidity networks and text mining of big data such as PubMed could help to find previously unknown genes associated with CRC.¹⁹¹ CRC arises as the cumulative effect of multiple genetic mutations. In table 5.4, we presented associations of the APC genes with CRC and comorbidities. The APC genes mutations are responsible for Familial adenomatous polyposis and hereditary non-polyposis CRC.¹⁹² Mutations in the APC gene play a pivotal role in CRC pathogenesis. The APC became one of the most frequently mutated, known driver genes in CRC.¹⁹³ The APC tumor suppressor is mutated or hyper-methylated in some breast cancers,¹⁹⁴ and may also be associated with the development and progression of bladder cancer and prostate cancer.^{195,196}

In table 5.5, the most often cancers and comorbidities associated with TP53 gene are listed. The TP53 has an important role in several fundamental processes such as cancer, aging, senescence, and DNA repair. Mutations in the TP53 gene are common in CRC,^{197,198} brain tumors, leukemia, and lymphomas.¹⁹⁸ BeFree text-mining discovered associations between the TP53 gene and different cancers as well as Hodgkin's disease, ulcerative colitis, rheumatoid arthritis, Alzheimer's disease, etc.

Table 5.6, presents the most often conditions associated with KRAS genes. Ucar and colleagues found that the presence of multiple mutations in KRAS indicates better overall

survival than a single mutation.¹⁹⁹ Perdyan and colleagues studied the presence of KRAS mutation as a prognostic factor of CRC.²⁰⁰ KRAS, NRAS, and BRAF mutations are found in half of myeloma patients and contribute to proteasome inhibitor (PI) resistance.²⁰¹ RAS genes (HRAS, KRAS, and NRAS) are seen as some of the top causes of cancer deaths in the U.S. (lung, colorectal, and pancreatic cancer) which influenced that anti-RAS therapies became a major field for cancer research.²⁰² Our text mining identified associations between the KRAS gene and viral hepatitis C, chronic pancreatitis, ulcerative colitis, dengue, atrial premature beats, endometriosis, leiomyoma, etc.

The most often conditions associated with MLH1 genes are shown in table 5.7. Researchers demonstrated that MLH1 deficiency and tumor progression and metastasis are in close relation.²⁰³ A defective DNA mismatch repair (MMR) of genes especially MLH1 and MSH2 is frequently involved in the carcinogenesis of various tumors including gastric cancer.²⁰⁴

Associations of CRC, comorbidities, and TGFBR2 gene are presented in table 5.8. In CRC with microsatellite instability (MSI), the majority of cases are affected by inactivating mutations of TGFBR2.²⁰⁵ Genetic variants in TGFBR1 and TGFBR2 genes have been associated with hereditary connective tissue disorders including thoracic aortic aneurysm and dissection, Marfan syndrome, and Loeys-Dietz syndrome.²⁰⁶ BeFree system discovered associations between the TGFBR2 gene and Alzheimer's disease, asthma, ulcerative colitis, migraine, etc.

In table 5.9, PPARG genes in associations with CRC comorbidities are listed. Studies described how epigenetic modifications influence PPARG gene expression in CRC.²⁰⁷

PPARG has been analyzed in the regulation of metabolism, inflammation, atherosclerosis, cell differentiation, and proliferation, linking PPARG with conditions such as obesity and diabetes, cardiovascular disease, and cancer.²⁰⁸

Network data analysis used in our research provides insight into comorbidities and genes associated with CRC, which could help medical experts to formulate appropriate preventive health measures to address genes and high-risk comorbidities associated with CRC. Our study has several limitations. The social network approach may underestimate rare diseases, that don't show as hubs, but they could also affect outcomes. Sometimes testing for comorbidity diseases could be expensive, which could limit confirmation of those conditions.

5.5. Conclusion

Numerous studies point out that the presence of comorbidities weighs as a negative factor that contributes to faster deterioration of one's health. Our findings suggest which comorbidities should be highly expected along the course of the cancer disease. The results of this study contribute to a better understanding of risk factors (genes, comorbidities) to the development of CRC. Genes associated with CRC and comorbidities, found only in PubMed abstracts should be evaluated as a potential risk or predictive factor for the development of comorbidity conditions. Subsequent genetic research is necessary to evaluate genes-CRC-comorbidities associations and incorporate findings into expert domain-specific curated databases.

CHAPTER 6

CONCLUSIONS

Computer science and machine learning could play a significant role in the development of evidence-based precision medicine. ML models, combined with big medical data stored in large EHR databases, create a foundation for new approaches in medicine, new diagnostic methods, especially in the fields of imaging and genetics, and improved therapy. In the prediction of AD research project, we developed a new method that combines deep learning ML models and the implementation of domain medical knowledge in the selection of inputs for the model. The inclusion of relevant medical information significantly improved the prediction accuracy results of the LSTM RNN model, which could not be achieved by further improvements of the algorithms only.

The method developed in the research of the application of advanced ML models in the prediction of complications of DM2 proved that deep learning approaches, especially the RNN GRU model, were superior to traditional ML models on temporal EMR medical data. In this research, we evaluated the amount of EMR data and the domains of data required for the most optimal performance of the model.

The application of computer science network methods explained the distribution of flu infections that required hospital admissions. The computer science network methods and heatmap visualization of the distribution of severe flu cases facilitate the development of ML methods for the prediction of this distribution in the future as well as the prediction of necessary resources to hospitalize and treat patients when they need it. The developed models are generalizable to any disease and any geographic location.

Our ML methods applications research in colorectal cancer, thyroid cancer, and several other common cancers created new approaches to extract risk factors for these diseases in the form of comorbidities and genes. The combination of network science and advanced text mining methods enabled the discovery of risk factors from big medical databases and PubMed as the largest collection of indexed medical papers. These findings provide the source for further development of predictive ML methods that would help in the prevention and better control of risk factors and cancer disease overall.

Cumulative conclusions of the entire research discovered that deep learning approaches, especially the RNN GRU model, were superior to traditional ML models when applied to temporal EMR medical data. Traditional ML methods achieve lower accuracy but they have the advantage of simplicity and interpretability of models. Deep learning models produce higher accuracy but they suffer from issues of model interpretability. Computational methods that use optimization function to automatically select useful features, contribute to good prediction results of medical conditions.^{76,77,209} In addition to automatically selected features, medical domain knowledge significantly helps in the identification of features that could further improve prediction results.^{67,71,210-212}

Our research shows that complex medical issues cannot be addressed only by algorithms, they require collaboration between computer scientists and medical experts to create meaningful ML solutions applicable in medicine.

BIBLIOGRAPHY

1. Hoyt RE, Yoshihashi AK. Health Informatics: Practical Guide for Healthcare and Information Technology Professionals, Sixth Edition, Lulu Press. Morrisville 2014.
2. Gligorijevic D, Stojanovic J, Djuric N, et al. Large-scale discovery of disease-disease and disease-gene associations. *Scientific reports* 2016;6(1):1-2.
3. Gligorijevic D, Stojanovic J, Satz W, et al. Deep attention model for triage of emergency department patients. In *Proceedings of the 2018 SIAM International Conference on Data Mining*, pp 297-305, 2018.
4. Stojanovic J, Gligorijevic D, Radosavljevic V, et al. Modeling healthcare quality via compact representations of electronic health records. *IEEE/ACM Trans Comput Biol Bioinform* 2016;14(3):545-54.
5. Waringa J, Lindvall C, Umeton R. Automated machine learning: Review of the state-of-the-art and opportunities for healthcare. <https://doi.org/10.1016/j.artmed.2020.101822>.
6. Hamet P, Tremblay J. Artificial intelligence in medicine. *Metabolism* 2017;69:36-40. <https://doi.org/10.1016/j.metabol.2017.01.011>.
7. Daume H. A Course in Machine Learning. Second edition. http://ciml.info/dl/v0_99/ciml-v0_99-all.pdf (2017).
8. Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and regression trees. CRC press. Boca Raton, 1984.
9. Breiman L. Random forests. *Machine learning* 2001;45(1):5-32.
10. Breiman L. Bagging predictors. *Machine learning* 1996;24(2):123-40.
11. Schapire RE, Freund Y. Boosting: Foundations and algorithms. The MIT Press. Cambridge, 2014.

12. Arsov N, Pavlovski M, Basnarkov L, Kocarev L. Generating highly accurate prediction hypotheses through collaborative ensemble learning. *Scientific reports* 2017;7:44649.
13. Pavlovski M, Zhou F, Stojkovic I, Kocarev L, Obradovic Z. Adaptive skip-train structured regression for temporal networks. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases 2017 Sep 18*, pp. 305-21. Springer, Cham.
14. Pavlovski M, Zhou F, Arsov N, Kocarev L, Obradovic Z. Generalization-Aware Structured Regression towards Balancing Bias and Variance. In *IJCAI 2018 Jul 13*, pp. 2616-22.
15. http://ciml.info/dl/v0_99/ciml-v0_99-ch04.pdf
16. Rumelhart DE, Geoffrey EH, Williams RJ. "Learning Internal Representations by Error Propagation". David E. Rumelhart, James L. McClelland, and the PDP research group. (editors), *Parallel distributed processing: Explorations in the microstructure of cognition, Volume 1: Foundation*. MIT Press. Cambridge, 1986.
17. Barber D. *Bayesian reasoning and machine learning*. Cambridge University Press. Cambridge, 2012.
18. Cortes C, Vapnik V. Support-vector networks. *Machine learning* 1995;20(3):273-97.
19. Cover T, Hart P. Nearest neighbor pattern classification[J]. *Information Theory, IEEE Trans Inf Theory* 1967;13(1):21-7.
20. Weisberg S. *Applied linear regression*. John Wiley & Sons. Hoboken, 2005.
21. Murphy K. Logistic regression. *Machine Learning: A Probabilistic Perspective*, Chapter 8, pp. 245 –279. MIT Press. Cambridge, 2012.
22. Goodfellow I, Bengio Y, Courville A. *Deep learning*. MIT press. Cambridge, 2016.
23. Williams RJ, Zipser D. A learning algorithm for continually running fully recurrent neural networks. *Neural Comput* 1989;1(2):270–80.
24. Soguero-Ruiz C, Mora-Jiménez I, Mohedano-Munoz MA, Rubio-Sanchez M, Miguel-Bohoyo P, Sanchez A. Visually guided classification trees for analyzing chronic patients. *BMC Bioinformatics* 2020;21(Suppl 2):92. doi:10.1186/s12859-020-3359-3.

25. Joloudari JH, Joloudari EH, Saadatfar H, et al. Coronary Artery Disease Diagnosis; Ranking the Significant Features Using a Random Trees Model. *Int J Environ Res Public Health* 2020;17(3):731. doi: 10.3390/ijerph17030731.
26. Kuo KM, Talley PC, Huang CH, Cheng LC. Predicting hospital-acquired pneumonia among schizophrenic patients: a machine learning approach. *BMC Med Inform Decis Mak* 2019;19(1):42. doi: 10.1186/s12911-019-0792-1.
27. De Felice F, Crocetti D, Parisi M, et al. Decision tree algorithm in locally advanced rectal cancer: an example of over-interpretation and misuse of a machine learning approach. *J Cancer Res Clin Oncol* 2020;146(3):761-5. doi:10.1007/s00432-019-03102-y
28. Chavan S, Scherbak N, Engwall M, Repsilber D. Predicting Chemical-Induced Liver Toxicity Using High-Content Imaging Phenotypes and Chemical Descriptors: A Random Forest Approach. *Chem Res Toxicol* 2020;33(9):2261-75. doi:10.1021/acs.chemrestox.9b00459.
29. Tracy BM, Finnegan TM, Smith RN, Senkowski CK. Random forest modeling using socioeconomic distress predicts hernia repair approach. *Surg Endosc* 2020. doi:10.1007/s00464-020-07860-6.
30. Mar J, Gorostiza A, Ibarrodo O, et al. Validation of Random Forest Machine Learning Models to Predict Dementia-Related Neuropsychiatric Symptoms in Real-World Data. *J Alzheimers Dis* 2020;77(2):855-64. doi:10.3233/JAD-200345.
31. Iwendi C, Bashir AK, Peshkar A, et al. COVID-19 Patient Health Prediction Using Boosted Random Forest Algorithm. *Front Public Health* 2020;8:357. doi:10.3389/fpubh.2020.00357.
32. Yaman E, Subasi A. Comparison of Bagging and Boosting Ensemble Machine Learning Methods for Automated EMG Signal Classification. *Biomed Res Int*. doi:10.1155/2019/9152506.
33. Nikolaidis A, Solon Heinsfeld A, Xu T, Bellec P, Vogelstein J, Milham M. Bagging improves reproducibility of functional parcellation of the human brain. *Neuroimage* 2020;214:116678. doi:10.1016/j.neuroimage.2020.116678.
34. Saravanakumar S, Thangaraj P. A Computer Aided Diagnosis System for Identifying Alzheimer's from MRI Scan using Improved Adaboost. *J Med Syst* 2019;43(3):76. doi:10.1007/s10916-018-1147-7.
35. Li S, Zeng Y, Chapman WC Jr, et al. Adaptive Boosting (AdaBoost)-based multiwavelength spatial frequency domain imaging and characterization for ex vivo human

colorectal tissue assessment. *J Biophotonics* 2020;13(6):e201960241. doi:10.1002/jbio.201960241.

36. Car Z, Baressi Šegota S, Anđelić N, Lorencin I, Mrzljak V. Modeling the Spread of COVID-19 Infection Using a Multilayer Perceptron. *Comput Math Methods Med* 2020;2020:5714714. doi:10.1155/2020/5714714.

37. Leithner D, Mayerhoefer ME, Martinez DF, et al. Non-Invasive Assessment of Breast Cancer Molecular Subtypes with Multiparametric Magnetic Resonance Imaging Radiomics. *J Clin Med* 2020;9(6):1853. doi:10.3390/jcm9061853.

38. Madhukar NS, Khade PK, Huang L, et al. A Bayesian machine learning approach for drug target identification using diverse data types. *Nat Commun* 2019;10(1):5221. doi:10.1038/s41467-019-12928-6.

39. Gupta A, Slater JJ, Boyne D, et al. Probabilistic Graphical Modeling for Estimating Risk of Coronary Artery Disease: Applications of a Flexible Machine-Learning Method. *Med Decis Making* 2019;39(8):1032-44. doi:10.1177/0272989X19879095.

40. Varghese J, Fujarski M, Hahn T, Dugas M, Warnecke T. The Smart Device System for Movement Disorders: Preliminary Evaluation of Diagnostic Accuracy in a Prospective Study. *Stud Health Technol Inform* 2020;270:889-93. doi:10.3233/SHTI200289.

41. Satoh Y, Tamada D, Omiya Y, Onishi H, Motosugi U. Diagnostic Performance of the Support Vector Machine Model for Breast Cancer on Ring-Shaped Dedicated Breast Positron Emission Tomography Images. *J Comput Assist Tomogr* 2020;44(3):413-8. doi:10.1097/RCT.0000000000001020.

42. Lown M, Brown M, Brown C, et al. Machine learning detection of Atrial Fibrillation using wearable technology. *PLoS One* 2020;15(1):e0227401. doi:10.1371/journal.pone.0227401.

43. Jensen C, Carl J, Boesen L, Langkilde NC, Østergaard LR. Assessment of prostate cancer prognostic Gleason grade group using zonal-specific features extracted from biparametric MRI using a KNN classifier. *J Appl Clin Med Phys* 2019;20(2):146-53. doi:10.1002/acm2.12542.

44. Saçlı B, Aydınalp C, Cansız G, et al. Microwave dielectric property based classification of renal calculi: Application of a kNN algorithm. *Comput Biol Med* 2019;112:103366. doi:10.1016/j.compbiomed.2019.103366.

45. Verma V, Vishwakarma RK, Verma A, Nath DC, Khan HTA. Time-to-Death approach in revealing Chronicity and Severity of COVID-19 across the World. *PLoS One* 2020;15(5):e0233074. doi:10.1371/journal.pone.0233074.
46. Levy JF, Rosenberg MA. A Latent Class Approach to Modeling Trajectories of Health Care Cost in Pediatric Cystic Fibrosis. *Med Decis Making* 2019;39(5):593-04. doi:10.1177/0272989X19859875.
47. Ghosal S, Sengupta S, Majumder M, Sinha B. Linear Regression Analysis to predict the number of deaths in India due to SARS-CoV-2 at 6 weeks from day 0 (100 cases - March 14th 2020). *Diabetes Metab Syndr* 2020;14(4):311-5. doi:10.1016/j.dsx.2020.03.017
48. Ma X, Ng M, Xu S, et al. Development and validation of prognosis model of mortality risk in patients with COVID-19. *Epidemiol Infect* 2020;148:e168. doi:10.1017/S0950268820001727.
49. Tabaie A, Nemati S, Allen JW, et al. Assessing Contribution of Higher Order Clinical Risk Factors to Prediction of Outcome in Aneurysmal Subarachnoid Hemorrhage Patients. *AMIA Annu Symp Proc* 2020;2019:848-56.
50. Long C, Lv G, Fu X. Development of a general logistic model for disease risk prediction using multiple SNPs. *FEBS Open Bio* 2019;9(11):2006-2012. doi:10.1002/2211-5463.12722.
51. Deng Q. Dynamics and Development of the COVID-19 Epidemic in the United States: A Compartmental Model Enhanced with Deep Learning Techniques. *J Med Internet Res* 2020;22(8):e21173. doi: 10.2196/21173.
52. Jung SY, Jo H, Son H, Hwang HJ. Real-World Implications of a Rapidly Responsive COVID-19 Spread Model with Time-Dependent Parameters via Deep Learning: Model Development and Validation. *J Med Internet Res* 2020;22(9):e19907. doi: 10.2196/19907
53. Angenent-Mari NM, Garruss AS, Soenksen LR, Church G, Collins JJ. A deep learning approach to programmable RNA switches. *Nat Commun* 2020;11(1):5057. doi:10.1038/s41467-020-18677-1.
54. Ribeiro AH, Ribeiro MH, Paixão GMM, et al. Automatic diagnosis of the 12-lead ECG using a deep neural network. *Nat Commun* 2020;11(1):1760. doi:10.1038/s41467-020-15432-4.

55. Amarbayasgalan T, Park KH, Lee JY, Ryu KH. Reconstruction error based deep neural networks for coronary heart disease risk prediction. *PLoS One* 2019;14(12):e0225991. doi: 10.1371/journal.pone.0225991.
56. Banerjee A, Ray S, Vorselaars B, et al. Use of Machine Learning and Artificial Intelligence to predict SARS-CoV-2 infection from Full Blood Counts in a population. *Int Immunopharmacol* 2020;86:106705. doi:10.1016/j.intimp.2020.106705.
57. Winkler JK, Sies K, Fink C, et al. Melanoma recognition by a deep learning convolutional neural network-Performance in different melanoma subtypes and localisations. *Eur J Cancer* 2020;127:21-9. doi:10.1016/j.ejca.2019.11.020.
58. Sedik A, Iliyasu AM, Abd El-Rahiem B, et al. Deploying Machine and Deep Learning Models for Efficient Data-Augmented Detection of COVID-19 Infections. *Viruses* 2020;12(7):769. doi: 10.3390/v12070769
59. Ko H, Chung H, Kang WS, et al. COVID-19 Pneumonia Diagnosis Using a Simple 2D Deep Learning Framework with a Single Chest CT Image: Model Development and Validation. *J Med Internet Res* 2020;22(6):e19569. doi:10.2196/19569.
60. Oberai A, Varghese B, Cen S, et al. Deep learning based classification of solid lipid-poor contrast enhancing renal masses using contrast enhanced CT. *Br J Radiol* 2020;93(1111):20200002. doi:10.1259/bjr.20200002.
61. Wetstein SC, Onken AM, Luffman C, et al. Deep learning assessment of breast terminal duct lobular unit involution: Towards automated prediction of breast cancer risk. *PLoS One* 2020;15(4):e0231653. doi:10.1371/journal.pone.0231653.
62. Al-Ajlan A, El Allali A. CNN-MGP: Convolutional Neural Networks for Metagenomics Gene Prediction. *Interdiscip Sci* 2019;11(4):628-35. doi:10.1007/s12539-018-0313-4.
63. Mammone N, Ieracitano C, Morabito FC. A deep CNN approach to decode motor preparation of upper limbs from time-frequency maps of EEG signals at source level. *Neural Netw* 2020;124:357-72. doi:10.1016/j.neunet.2020.01.027.
64. Muramatsu C. Diagnosis of Glaucoma on Retinal Fundus Images Using Deep Learning: Detection of Nerve Fiber Layer Defect and Optic Disc Analysis. *Adv Exp Med Biol* 2020;1213:121-32. doi:10.1007/978-3-030-33128-3_8.
65. Hatanaka Y. Retinopathy Analysis Based on Deep Convolution Neural Network. *Adv Exp Med Biol* 2020;1213:107-20. doi: 10.1007/978-3-030-33128-3_7.

66. Golestani N, Moghaddam M. Human activity recognition using magnetic induction-based motion signals and deep recurrent neural networks. *Nat Commun* 2020;11(1):1551. doi: 10.1038/s41467-020-15086-2.
67. Ljubic B, Abdel Hai A, Stanojevic M, et al. Predicting Complications of Diabetes Mellitus Using Advanced Machine Learning Algorithms. *JAMIA* 2020;27(9):1343-51. doi:10.1093/jamia/ocaa120.
68. Lobo B, Abdel-Rahman E, Brown D, Dunn L, Bowman B. A recurrent neural network approach to predicting hemoglobin trajectories in patients with End-Stage Renal Disease. *Artif Intell Med* 2020;104:101823. doi:10.1016/j.artmed.2020.101823.
69. Li R, Wu Q, Liu J, Wu Q, Li C, Zhao Q. Monitoring Depth of Anesthesia Based on Hybrid Features and Recurrent Neural Network. *Front Neurosci* 2020;14:26. doi:10.3389/fnins.2020.00026
70. Scherpf M, Gräßer F, Malberg H, Zaunseder S. Predicting sepsis with a recurrent neural network using the MIMIC III database. *Comput Biol Med* 2019;113:103395. doi: 10.1016/j.combiomed.2019.103395.
71. Ljubic B, Roychoudhury S, Cao XH, et al. Influence of medical domain knowledge on deep learning for Alzheimer's disease prediction. *Comput Methods Programs Biomed* 2020;197:105765. doi:10.1016/j.cmpb.2020.105765.
72. Mosquera-Lopez C, Dodier R, Tyler N, Resalat N, Jacobs P. Leveraging a Big Dataset to Develop a Recurrent Neural Network to Predict Adverse Glycemic Events in Type 1 Diabetes. *IEEE J Biomed Health Inform* 2019. doi: 10.1109/JBHI.2019.2911701.
73. Maragatham G, Devi S. LSTM Model for Prediction of Heart Failure in Big Data. *J Med Syst* 2019;43(5):111. doi: 10.1007/s10916-019-1243-3.
74. <https://www.hcup-us.ahrq.gov/databases.jsp>.
75. Antun V, Renna F, Poon C, Adcock B, Hansen AC. On instability of deep learning in image reconstruction and the potential costs of AI. *Proceedings of the National Academy of Sciences of the United States of America* 2020;117(48):30088-95.
76. Ghalwash M, Cao XH, Stojkovic I, Obradovic Z. Structured feature selection using coordinate descent optimization. *BMC Bioinformatics* 17, 158 2016; doi: <https://doi.org/10.1186/s12859-016-0954-4>.
77. Stojkovic I, Ghalwash M, Obradovic Z. Ranking based Multitask Learning of Scoring Functions. In *ECML-PKDD 2017*;721-736.

78. <https://www.nia.nih.gov/health/alzheimers-disease-fact-sheet>.
79. <https://www.alz.org/media/Documents/alzheimers-facts-and-figures-2019-r.pdf>.
80. National Institutes of Health. National Institute on Aging. What Happens to the Brain in Alzheimer's Disease? Available at: <https://www.nia.nih.gov/health/what-happens-brain-alzheimers-disease>. September 14, 2018.
81. Wang T, Qiu RG, Yu M. Predictive Modeling of the Progression of Alzheimer's Disease with Recurrent Neural Networks. *Sci Rep* 2018;15;8(1):9161. doi:10.1038/s41598-018-27337-w.
82. Roberts R, Knopman DS. Classification and epidemiology of MCI. *Clin Geriatr Med* 2013;29(4):753-72. doi: 10.1016/j.cger.2013.07.003.
83. Mosconi L, Brys M, Glodzik-Sobanska L, et al. Early detection of Alzheimer's disease using neuroimaging. *Exp Gerontol* 2007;42(1-2):129-38.
84. McMahon PM, Araki SS, Sandberg EA, et al. Cost-effectiveness of PET in the diagnosis of Alzheimer's disease. *Radiology* 2003;228(2):515-22.
85. Zhao J, Papapetrou P, Asker L, et al. Learning from heterogeneous temporal data in electronic health records. <http://dx.doi.org/10.1016/j.jbi.2016.11.006>.
86. Tang F, Xiao C, Wang F, et al. Predictive modeling in urgent care: a comparative study of machine learning approaches. *JAMIA Open* 2018;1(1):87-98.
87. Qiu RG, Qiu JL, Badr Y. Predictive modeling of the severity/progression of Alzheimer's diseases. 2017 International Conference on Grey Systems and Intelligent Services (GSIS). 400-3. DOI: 10.1109/GSIS.2017.8077739.
88. Moradi E, Pepe A, Gaser C, et al. Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects. *Neuroimage* 2015;104:398-412. doi: 10.1016/j.neuroimage.2014.10.002.
89. Esmailzadeh S, Belivanis DI, Pohl KM, et al. End-To-End Alzheimer's Disease Diagnosis and Biomarker Identification. In: Shi Y., Suk HI., Liu M. (eds) *Machine Learning in Medical Imaging. MLMI 2018. Lecture Notes in Computer Science*, vol 11046. Springer, Cham.
90. Albright J. Forecasting the progression of Alzheimer's disease using neural networks and a novel preprocessing algorithm. <https://doi.org/10.1016/j.trci.2019.07.001>.

91. Aghili M, Tabarestani S, Adjouadi M, et al. Predictive Modeling of Longitudinal Data for Alzheimer's Disease Diagnosis Using RNNs. International Workshop on Predictive Intelligence In Medicine. Springer, Cham, 2018. https://doi.org/10.1007/978-3-030-00320-3_14.
92. Goldstein BA, Navar AM, Pencina MJ, et al. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *J Am Med Inform Assoc* 2017;24(1):198–208.
93. Yang J, McAuley J, Leskovec J, et al. Finding progression stages in time-evolving event sequences. In: Proceedings of the 23rd international conference on World wide web. (2014). <https://doi.org/10.1145/2566486.2568044>.
94. Hochreiter S, Schmidhuber J. Long Short Term Memory. *Neural Computation* 1997;9(8):1735-80.
95. Sherstinsky A. Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network. arXiv:1808.03314v4.
96. <https://www.cdc.gov/coronavirus/2019-ncov/index.html>.
97. <https://www.ohdsi.org/data-standardization/the-common-data-model>.
98. Marzban EN, Eldeib AM, Yassine IA, et al. Alzheimer's Disease Diagnosis From Diffusion Tensor Images Using Convolutional Neural Networks. *PLoS One* 2020;15(3):e0230409.
99. Arbabshirani MR, Plis S, Sui J, et al. Single Subject Prediction of Brain Disorders in Neuroimaging: Promises and Pitfalls. *Neuroimage*. 2017;145(Pt B):137-165.
100. Moore PJ, Lyons TJ, Gallacher J. Using Path Signatures to Predict a Diagnosis of Alzheimer's Disease. *PLoS One* 2019;14(9):e0222212.
101. Martinez-Murcia FJ, Ortiz A, Gorriz JM, et al. Studying the Manifold Structure of Alzheimer's Disease: A Deep Learning Approach Using Convolutional Autoencoders. *IEEE J Biomed Health Inform* 2020;24(1):17-26. doi: 10.1109/JBHI.2019.2914970.
102. Kam TE, Zhang H, Shen D. A Novel Deep Learning Framework on Brain Functional Networks for Early MCI Diagnosis. *Med Image Comput Comput Assist Interv* 2018;11072:293-301. doi: 10.1007/978-3-030-00931-1_34.

103. Moscoso A, Silva-Rodríguez J, Aldrey JM, et al. Prediction of Alzheimer's Disease Dementia With MRI Beyond the Short-Term: Implications for the Design of Predictive Models. *Neuroimage Clin* 2019;23:101837. doi: 10.1016/j.nicl.2019.101837.
- 104 Moore PJ, Lyons TJ, Gallacher J. Random Forest Prediction of Alzheimer's Disease Using Pairwise Selection From Time Series Data. *PLoS One* 2019;14(2):e0211558. doi: 10.1371/journal.pone.0211558.
- 105 Spasov S, Passamonti L, Duggento A, et al. A Parameter-Efficient Deep Learning Approach to Predict Conversion From Mild Cognitive Impairment to Alzheimer's Disease. *Neuroimage* 2019;189:276-287. doi: 10.1016/j.neuroimage.2019.01.031.
- 106 Carpenter KA, Huang X. Machine Learning-based Virtual Screening and Its Applications to Alzheimer's Drug Discovery: A Review. *Curr Pharm Des* 2018;24(28):3347-3358. doi: 10.2174/1381612824666180607124038.
107. Huang L, Jin Y, Gao Y, et al. Longitudinal Clinical Score Prediction in Alzheimer's Disease With Soft-Split Sparse Regression-Based Random Forest. doi:10.1016/j.neurobiolaging.2016.07.005.
108. Ford E, Rooney P, Oliver S, et al. Identifying Undetected Dementia in UK Primary Care Patients: A Retrospective Case-Control Study Comparing Machine-Learning and Standard Epidemiological Approaches. *BMC Med Inform Decis Mak* 2019;19(1):248. doi: 10.1186/s12911-019-0991-9.
109. Orasanu G, Plutzky J. The pathologic continuum of diabetic vascular disease. *J Am Coll Cardiol* 2009;53(5 Suppl):S35-42. doi: 10.1016/j.jacc.2008.09.055.
110. Deedwania PC. Management of Patients With Stable Angina and Type 2 Diabetes. *Rev Cardiovasc Med* 2015;16(2):105-13.
111. Duh EJ, Sun JK, Stitt AW. Diabetic retinopathy: current understanding, mechanisms, and treatment strategies. *JCI Insight* 2017;2(14):e93751. doi: 10.1172/jci.insight.93751.
112. <https://www.cdc.gov/diabetes/basics/diabetes.html>.
113. <https://www.who.int/health-topics/diabetes>.
114. <https://www.niddk.nih.gov/health-information/diabetes>.
115. Rodriguez-Gutierrez R, Gonzalez-Gonzalez JG, Zuñiga-Hernandez JA, et al. Benefits and harms of intensive glycemic control in patients with type 2 diabetes. *BMJ* 2019;367:l5887. doi: 10.1136/bmj.l5887.

116. Garber AJ, Abrahamson MJ, Barzilay JI, et al. Consensus statement by the American Association of Clinical Endocrinologists and American College of Endocrinology on the comprehensive type 2 Diabetes management Algorithm - 2018 Executive summary. *Endocr Pract* 2018;24(1):91-120.
117. Qureshi M, Gammoh E, Shakil J, et al. Update on Management of Type 2 Diabetes for Cardiologists. *Methodist Debakey Cardiovasc J* 2018;14(4):273-280.
118. Bailey CJ, Day C. Treatment of type 2 diabetes: future approaches. <https://doi.org/10.1093/brimed/ldy013>.
119. Cahn A, Shoshan A, Sagiv T, et al. Use of a Machine Learning Algorithm Improves Prediction of Progression to Diabetes. <https://doi.org/10.2337/db18-1286-P>.
120. Contreras I, Vehi J. Artificial Intelligence for Diabetes Management and Decision Support: Literature Review. *J Med Internet Res* 2018;20(5):e10775.
121. Bodman MA, Varacallo M. *Diabetic Neuropathy*. StatPearls Publishing, 2019.
122. Callaghan BC, Cheng H, Stables CL, et al. Diabetic neuropathy: Clinical manifestations and current treatments. *Lancet Neurol* 2012;11(6):521–34. doi: 10.1016/S1474-4422(12)70065-0.
123. Bouhairie VE. Diabetic Kidney Disease. *Mo Med* 2016;113(5):390–4.
124. Holt RIG, de Groot M, Golden SH. Diabetes and Depression. *Curr Diab Rep* 2014;14(6):491. doi: 10.1007/s11892-014-0491-3.
125. Konrad-Martin D, Reavis KM, Austin D, et al. Hearing Impairment in Relation to Severity of Diabetes in a Veteran Cohort. doi: 10.1097/AUD.000000000000137.
126. Leon BM, Maddox TM. Diabetes and cardiovascular disease: Epidemiology, biological mechanisms, treatment recommendations and future research. *World J Diabetes* 2015;6(13): 1246–58. doi: 10.4239/wjd.v6.i13.1246.
127. Dagliati A, Simone Marini S, Sacchi L, et al. Machine Learning Methods to Predict Diabetes Complications. <https://doi.org/10.1177/1932296817706375>.
128. Hayeri A. Predicting Future Glucose Fluctuations Using Machine Learning and Wearable Sensor Data. <https://doi.org/10.2337/db18-738-P>.
129. Massaro A, Maritati V, Giannone D, et al. LSTM DSS Automatism and Dataset Optimization for Diabetes Prediction. *Appl Sci* 2019;9:3532. doi:10.3390/app9173532.

130. Perveen S, Shahbaz M, Keshavjee K, et al. Prognostic Modeling and Prevention of Diabetes Using Machine Learning Technique. doi.org/10.1038/s41598-019-49563-6.
131. Apoorva S, Aditya SK, Snigdha P, et al. Prediction of Diabetes Mellitus Type-2 Using Machine Learning. In: Smys S., Tavares J., Balas V., Ilyasu A. (eds) Computational Vision and Bio-Inspired Computing. ICCVBIC 2019. Advances in Intelligent Systems and Computing, vol 1108. Springer, Cham. https://doi.org/10.1007/978-3-030-37218-7_42.
132. Sarwar MA, Kamal N, Hamid W, et al. Prediction of Diabetes Using Machine Learning Algorithms in Healthcare. 2018 24th International Conference on Automation and Computing (ICAC). doi.org/10.23919/ICAC.2018.8748992.
133. Zhang XS, Tang F, Dodge H, et al. MetaPred: Meta-Learning for Clinical Risk Prediction with Limited Patient Electronic Health Records. Proceedings of the 25th ACM SIG KDD International Conference on Knowledge Discovery and Data Mining (KDD). 2487- 95. doi.org/10.1145/3292500.3330779.
134. Wang F, Casalino LP, Khullar D. Deep Learning in Medicine—Promise, Progress, and Challenges. JAMA Intern Med 2019;179(3):293-4. doi:10.1001/jamainternmed.2018.7117
135. Man CJ. Observational research methods. Research design II: cohort, cross sectional, and case-control studies. Emerg Med J 2003;20:54–60.
136. Klema V, Laub A. The singular value decomposition: Its computation and some applications, IEEE Transactions on Automatic Control 1980;25(2):164-76.
137. Williams RJ, Zipser D. A learning algorithm for continually running fully recurrent neural networks. Neural Comput 1989;1(2):270–80.
138. Cho K, van Merriënboer B, Gulcehre C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In Conference on Empirical Methods in Natural Language Processing (EMNLP 2014). 2014.
139. Rumelhart DE, Geoffrey EH, Williams RJ. "Learning Internal Representations by Error Propagation". David E. Rumelhart, James L. McClelland, and the PDP research group. (editors), Parallel distributed processing: Explorations in the microstructure of cognition, Volume 1: Foundation. MIT Press. 1986.
140. Ngufor C, Houten HV, Caffo BS, et al. Mixed Effect Machine Learning: a framework for predicting longitudinal change in hemoglobin A1c. J Biomed Inform 2019;89:56–67. doi:10.1016/j.jbi.2018.09.001.

141. Choi E, Schuetz A, Stewart WF, et al. Using recurrent neural network models for early detection of heart failure onset. *J Am Med Inform Assoc.* 2017; 24(2): 361–370. doi: 10.1093/jamia/ocw112.
142. Centers for Disease Control and Prevention (CDC). Estimated influenza illnesses and hospitalizations averted by influenza vaccination-United States, 2012-13 influenza season. *MMWR. Morbidity and mortality weekly report* 2013;62(49):997.
143. Center for Disease Control. Disease burden of Influenza. <https://www.cdc.gov/flu/resource-center/freeresources/graphics/flu-burden.htm>.
144. Cauchemez S, Bhattarai A, Marchbanks TL, et al. Role of social networks in shaping disease transmission during a community outbreak of 2009 H1N1 pandemic influenza. *Proceedings of the National Academy of Sciences* 2011;108(7):2825-30.
145. Read JM, Ken T. D, Eames TD, et al. Dynamic social networks and the implications for the spread of infectious disease. *J. R. Soc. Interface* 2008;5:1001–7.
146. Viboud C, Bjørnstad O, Smith DL, et al. Synchrony, Waves, and Spatial Hierarchies in the Spread of Influenza. *Science*,2006;5772:447-51.
147. Gog JR, Ballesteros S, Viboud C, et al. Spatial Transmission of 2009 Pandemic Influenza in the US. *PLOS Computational Biology*, 2014;10(6):e1003635.
148. Kissler SM, Gog JR, Viboud C, et al. Geographic transmission hubs of the 2009 influenza pandemic in the United States. <https://doi.org/10.1016/j.epidem.2018.10.002>.
149. Malosh RE, Martin ET, Ortiz JR, et al. The risk of lower respiratory tract infection following influenza virus infection: A systematic and narrative review. *Vaccine*, 2018;36(1):141-147.
150. Pearce DC, McCaw JM, McVernon J, et al. Influenza as a trigger for cardiovascular disease: An investigation of serotype, subtype and geographic location. *Environ Res.* 2017;156:688-696.
151. Ortiz JR, Neuzil KM, Cooke CR, et al. Influenza Pneumonia Surveillance among Hospitalized Adults May Underestimate the Burden of Severe Influenza Disease. DOI:10.1371/journal.pone.0113903.
152. Ginsberg, J. et al. Detecting influenza epidemics using search engine query data. *Nature*, 2009;457:1012–5.

153. Davidson MW, Haim DA, Radin JM. Using networks to combine “big data” and traditional surveillance to improve influenza predictions. *Scientific reports* 2015;5:8154.
154. Miller RR, Markewitz BA, Rolfs RT, et al. Clinical findings and demographic factors associated with ICU admission in Utah due to novel 2009 influenza A(H1N1) infection. *Chest*. 2010;137(4):752-8.
155. Jung JJ, Pinto R, Zarychanski R, et al. 2009-2010 Influenza A(H1N1)-related critical illness among Aboriginal and non-Aboriginal Canadians. *PLoS One*, 2017;12(10):e0184013.
156. Danon L, Ford A, House T, et al. Networks and the Epidemiology of Infectious Disease. *Hindawi. Interdisciplinary Perspectives on Infectious Diseases* 2011; Article ID 284909: 1-28.
157. Gligorijevic Dj, Stojanovic J, Obradovic Z. Uncertainty Propagation in Long-Term Structured Regression on Evolving Networks. *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16) Phoenix, AZ, February 2016*, 1603-10.
158. Barabasi A. *Network Science* (book). 2016.
159. Easley D, Kleinberg J. *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*. 2010.
160. Meyer S, Held L. "Power-law models for infectious disease spread." *The Annals of Applied Statistics*, (2014);8(3):1612-1639.
161. Poolkhet, C, Chairatanayuth P, Thongratsakulet S et al. Social network analysis for assessment of avian influenza spread and trading patterns of backyard chickens in Nakhon Pathom, Suphan Buri and Ratchaburi, Thailand. *Zoonoses and public health* 2013;60(6):448-455.
162. Song R, Hall HI, McDavid-Harrison K, et al. Identifying the impact of social determinants of health on disease rates using correlation analysis of area-based summary information. *Public health reports*, 2011;126(supple.3): 70-80.
163. Kumar V, Abbas A, Fausto N, et al. *Robbins and Cotran. Pathologic basis of diseases*. 9th edition. 2014.
164. Survival Rates for Colorectal Cancer, by Stage. <https://www.cancer.org/cancer/colon-rectal-cancer/detection-diagnosis-staging/survival-rates.html>.

165. Hagggar FA, Boushey RP. Colorectal cancer epidemiology: incidence, mortality, survival, and risk factors. *Clin Colon Rectal Surg* 2009;22(4):191-197. doi:10.1055/s-0029-1242458.
166. Gallagher DJ, Kemeny N. Metastatic colorectal cancer: from improved survival to potential cure. *Oncology* 2010;78(3-4):237-48.
167. Hahn EE, Gould MK, Munoz-Plaza CE, et al. Understanding Comorbidity Profiles and Their Effect on Treatment and Survival in Patients With Colorectal Cancer. *J Natl Compr Canc Netw* 2018;16(1):23-34. doi: 10.6004/jnccn.2017.7026.
168. Boakye D, Rillmann B, Walter V, et al. Impact of comorbidity and frailty on prognosis in colorectal cancer patients: A systematic review and meta-analysis. *Cancer Treat Rev* 2018;64:30-39. doi: 10.1016/j.ctrv.2018.02.003.
169. Hidalgo CA, Blumm N, Barabasi AL, et al. A Dynamic Network Approach for the Study of Human Phenotypes. *PLoS Computational Biology* 2009;5(4):e1000353.
170. Khan A, Uddin S, Srinivasan U. Comorbidity network for chronic disease: A novel approach to understand type 2 diabetes progression. *Int J Med Inform.* 2018;115:1-9. doi: 10.1016/j.ijmedinf.2018.04.001.
171. Chen Y, Zhang X, Zhang G, et al. Comparative analysis of a novel disease phenotype network based on clinical manifestations. *J Biomed Inform.* 2015;53:113-20. <https://doi.org/10.1016/j.jbi.2014.09.007>.
172. Hossain ME, Uddin S, Khan A, et al. A Framework to Understand the Progression of Cardiovascular Disease for Type 2 Diabetes Mellitus Patients Using a Network Approach. *Int J Environ Res Public Health* 2020;17(2). pii: E596. doi: 10.3390/ijerph17020596.
173. Genetics of colorectal cancer. https://www.cancer.gov/types/colorectal/hp/colorectal-genetics-pdq#link/_89.
174. Kanth P, Grimmer J, Champine M, et al.: Hereditary Colorectal Polyposis and Cancer Syndromes: A Primer on Diagnosis and Management. *Am J Gastroenterol* 2017;112 (10): 1509-1525.
175. Rustgi AK. The genetics of hereditary colon cancer. *Genes Dev* 2007;21(20): 2525-38.
176. Barnetson RA, Tenesa A, Farrington SM, et al. Identification and survival of carriers of mutations in DNA mismatch-repair genes in colon cancer. *N Engl J Med* 2006;354 (26):2751-63.

177. Hampel H, Bennett RL, Buchanan A, et al. A practice guideline from the American College of Medical Genetics and Genomics and the National Society of Genetic Counselors: referral indications for cancer predisposition assessment. *Genet Med* 2015;17(1): 70-87.
178. Bravo A, Piñero J, Queralt-Rosinach N, et al. Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research. *BMC Bioinformatics* 2015;16:55. DOI 10.1186/s12859-015-0472-9.
179. Pinero J, Bravo A, Queralt-Rosinach N, et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Research*, 2017;45:833–9. doi: 10.1093/nar/gkw943.
180. <https://www.hcup-us.ahrq.gov/databases.jsp>.
181. Ljubic B, Gligorijevic Dj, Gligorijevic J, et al. Social network analysis for better understanding of influenza. *J Biomed Inform*, 2019;93:103161. <https://doi.org/10.1016/j.jbi.2019.103161>.
182. https://en.wikipedia.org/wiki/Betweenness_centrality.
183. <https://ghr.nlm.nih.gov/gene/APC>.
184. <https://ghr.nlm.nih.gov/gene/TP53>.
185. <https://ghr.nlm.nih.gov/gene/KRAS>.
186. <https://ghr.nlm.nih.gov/gene/MLH1>.
187. <https://ghr.nlm.nih.gov/gene/TGFBR2>.
188. <https://ghr.nlm.nih.gov/gene/PPARG>.
189. Sogaard M, Thomsen RW, Bossen KS, et al. The impact of comorbidity on cancer survival: a review. *Clinical epidemiology*. 2013; 5(Suppl 1):3–29.
190. Barabási AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. *Nat Rev Genet*. 2011;12(1):56-68. doi:10.1038/nrg2918.
191. Wang X, Gulbahce N, Yu H, et al. Network-based methods for human disease gene prediction, *Briefings in Functional Genomics*, <https://doi.org/10.1093/bfgp/elr024>.

192. Fearnhead NS, Wilding JL, Bodmer WF. Genetics of colorectal cancer: hereditary aspects and overview of colorectal tumorigenesis. *Br Med Bull* 2002;64:27-43. doi: 10.1093/bmb/64.1.27.
193. Wachsmannova L, Mego M, Stevurkova V, et al. Novel strategies for comprehensive mutation screening of the APC gene. *Neoplasia* 2017;64(3):338-343. doi: 10.4149/neo_2017_303.
194. Stefanski CD, Keffler K, McClintock S, et al. APC loss affects DNA damage repair causing doxorubicin resistance in breast cancer cells. *Neoplasia* 2019;21(12):1143-1150. doi: 10.1016/j.neo.2019.09.002. Epub 2019 Nov 20. PMID: 31759252; PMCID: PMC6872841.
195. Bai ZJ, Liu Q, Wang XS, et al. APC promoter methylation is correlated with development and progression of bladder cancer, but not linked to overall survival: a meta-analysis. *Neoplasia* 2019;66(3):470-480. doi: 10.4149/neo_2018_181009N753.
196. Richiardi L, Fiano V, Vizzini L, et al Promoter methylation in APC, RUNX3, and GSTP1 and mortality in prostate cancer patients. *J Clin Oncol* 2009;27(19):3161-8. doi: 10.1200/JCO.2008.18.2485.
197. Porcelli B, Frosi B, Terzuoli L, et al. Expression of p185 and p53 in benign and malignant colorectal lesions. *Histochem J* 2001;33(1):51-7. doi: 10.1023/a:1017543930661.
198. Tolomeo D, L'Abbate A, Lonoce A, et al. Concurrent chromothripsis events in a case of TP53 depleted acute myeloid leukemia with myelodysplasia-related changes. *Cancer Genet* 2019;237:63-68. doi: 10.1016/j.cancergen.2019.06.009.
199. Ucar G, Ergun Y, Aktürk Esen S, et al. Prognostic and predictive value of KRAS mutation number in metastatic colorectal cancer. *Medicine (Baltimore)* 2020;99(39):e22407. doi: 10.1097/MD.00000000000022407.
200. Perdyan A, Spychalski P, Kacperczyk J, et al. Circulating Tumor DNA in KRAS positive colorectal cancer patients as a prognostic factor - a systematic review and meta-analysis. *Crit Rev Oncol Hematol* 2020;154:103065. doi: 10.1016/j.critrevonc.2020.103065.
201. Shirazi F, Jones RJ, Singh RK, et al. Activating KRAS, NRAS, and BRAF mutants enhance proteasome capacity and reduce endoplasmic reticulum stress in multiple myeloma. *Proc Natl Acad Sci U S A* 2020;117(33):20004-20014. doi: 10.1073/pnas.2005052117.

202. Waters AM, Der CJ. KRAS: The Critical Driver and Therapeutic Target for Pancreatic Cancer. *Cold Spring Harb Perspect Med* 2018;8(9):a031435. doi: 10.1101/cshperspect.a031435.
203. Ackermann A, Schrecker C, Bon D, et al. Downregulation of SPTAN1 is related to MLH1 deficiency and metastasis in colorectal cancer. *PLoS One* 2019;14(3):e0213411. doi: 10.1371/journal.pone.0213411.
204. Haron NH, Mohamad Hanif EA, Abdul Manaf MR, et al. Microsatellite Instability and Altered Expressions of MLH1 and MSH2 in Gastric Cancer. *Asian Pac J Cancer Prev* 2019;20(2):509-517. doi: 10.31557/APJCP.2019.20.2.509.
205. Fricke F, Mussack V, Buschmann D, et al. TGFBR2-dependent alterations of microRNA profiles in extracellular vesicles and parental colorectal cancer cells. *Int J Oncol* 2019;55(4):925-937. doi: 10.3892/ijo.2019.4859.
206. De Cario R, Sticchi E, Lucarini L, et al. Role of TGFBR1 and TGFBR2 genetic variants in Marfan syndrome. *J Vasc Surg* 2018;68(1):225-233.e5. doi: 10.1016/j.jvs.2017.04.071.
207. Sabatino L, Fucci A, Pancione M, et al. PPARG Epigenetic Deregulation and Its Role in Colorectal Tumorigenesis. *PPAR Res* 2012;2012:687492. doi: 10.1155/2012/687492.
208. Bandera Merchan B, Tinahones FJ, Macías-González M. Commonalities in the Association between PPARG and Vitamin D Related with Obesity and Carcinogenesis. *PPAR Res* 2016;2016:2308249. doi: 10.1155/2016/2308249.
209. Stojkovic I, Obradovic Z. Sparse learning of the disease severity score for high-dimensional data. *Complexity* 2017; doi: <https://doi.org/10.1155/2017/7120691>.
210. Stojkovic I, Ghalwash M, Cao XH, Obradovic Z. Effectiveness of multiple blood-cleansing interventions in sepsis, characterized in rats. *Scientific Reports* 6, 24719 2016; doi: <https://doi.org/10.1038/srep24719>.
211. Cao XH, Stojkovic I, Obradovic Z. Predicting sepsis severity from limited temporal observations. In *Discovery Science* 2014; 37-48.
212. Stojkovic I, Obradovic Z. Predicting Sepsis Biomarker Progression under Therapy. In *IEEE CBMS* 2017; 19-24.