Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Influence of Medical Domain Knowledge on Deep Learning for Alzheimer's Disease Prediction

Branimir Ljubic, Shoumik Roychoudhury, Xi Hang Cao, Martin Pavlovski, Stefan Obradovic, Richard Nair, Lucas Glass, Zoran Obradovic

PII: S0169-2607(20)31598-4
DOI: https://doi.org/10.1016/j.cmpb.2020.105765
Reference: COMM 105765

To appear in: Computer Methods and Programs in Biomedicine

Received date: 23 April 2020
Accepted date: 16 September 2020

Please cite this article as: Branimir Ljubic, Shoumik Roychoudhury, Xi Hang Cao, Martin Pavlovski, Stefan Obradovic, Richard Nair, Lucas Glass, Zoran Obradovic, Influence of Medical Domain Knowledge on Deep Learning for Alzheimer's Disease Prediction, Computer Methods and Programs in Biomedicine (2020), doi: https://doi.org/10.1016/j.cmpb.2020.105765

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2020 Published by Elsevier B.V.
**Highlights**

- LSTM, RNN model for prediction of Alzheimer’s diseases (AD) is developed from EMR data.
- Information from 3 EMR domains were used: conditions, measurements and drugs.
- We created positive AD cohorts using relevant medical knowledge as model inputs.
- Selection of relevant input cohorts was crucial for overall RNN model prediction.
- We efficiently applied the drugs and the measurement domain in prediction of AD.
Influence of Medical Domain Knowledge on Deep Learning for Alzheimer’s Disease Prediction

Branimir Ljubic\textsuperscript{a}, Shoumik Roychoudhury\textsuperscript{a}, Xi Hang Cao\textsuperscript{a}, Martin Pavlovski\textsuperscript{a}, Stefan Obradovic\textsuperscript{a}, Richard Nair\textsuperscript{b}, Lucas Glass\textsuperscript{b}, Zoran Obradovic\textsuperscript{a}

\textsuperscript{a}. Temple University, Center for Data Analytics and Biomedical Informatics (DABI), Philadelphia, PA 19121, U.S.A.
\textsuperscript{b}. IQVIA, Plymouth Meeting, PA 19462, U.S.A.

Corresponding author:
Branimir Ljubic, MD
Email: branimir.ljubic@temple.edu
Phone: 215 433 3526
Mailing address:
Temple University
1925 N 12\textsuperscript{th} Street, SERC 334, 035-10
Philadelphia, PA 19121
ABSTRACT

BACKGROUND AND OBJECTIVE: Alzheimer’s disease (AD) is the most common type of dementia that can seriously affect a person’s ability to perform daily activities. Estimates indicate that AD may rank third as a cause of death for older people, after heart disease and cancer. Identification of individuals at risk for developing AD is imperative for testing therapeutic interventions. The objective of the study was to determine could diagnostics of AD from EMR data alone (without relying on diagnostic imaging) be significantly improved by applying clinical domain knowledge in data preprocessing and positive dataset selection rather than setting naïve filters.

METHODS: Data were extracted from the repository of heterogeneous ambulatory EMR data, collected from primary care medical offices all over the U.S. Medical domain knowledge was applied to build a positive dataset from data relevant to AD. Selected Clinically Relevant Positive (SCRP) 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.

RESULTS: Risk scores prediction of AD using the drugs domain information in an SCRP AD dataset of 2,324 patients achieved high out-of-sample score - 0.98-0.99 Area Under the Precision-Recall Curve (AUPRC) when using 90% of SCRP dataset for training. AUPRC dropped to 0.89 when training the model using less than 1,500 cases from the SCRP dataset. The model was still significantly better than when using naïve dataset selection.

CONCLUSION: The LSTM RNN method that used data relevant to AD performed significantly better when learning from the SCRP dataset than when datasets were selected naïvely. The
integration of qualitative medical knowledge for dataset selection and deep learning technology provided a mechanism for significant improvement of AD prediction.

Accurate and early prediction of AD is significant in the identification of patients for clinical trials, which can possibly result in the discovery of new drugs for treatments of AD. Also, the contribution of the proposed predictions of AD is a better selection of patients who need imaging diagnostics for differential diagnosis of AD from other degenerative brain disorders.

**KEY WORDS**: Electronic medical records, Alzheimer’s disease prediction, Cognitive impairment, Deep learning, Recurrent Neural Networks.
1. INTRODUCTION

According to the National Institute of Aging, more than 5.5 million Americans are diagnosed with AD.[1] Estimates indicate that AD may rank third as a cause of death for older people, only second to heart disease and cancer.[2,3] Identification of individuals at risk for developing AD is imperative for testing therapeutic interventions.[4] Many researchers have presented an overview of the classification of Mild Cognitive Impairment (MCI).[5] 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.[6] Diagnostic imaging can be very costly and the question is what is the cost-effectiveness of imaging in AD.[7]

Heterogeneous structures of Electronic Medical Records (EMR) data pose a challenge for machine learning (ML) algorithms.[8-10] For our 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.[11-13] An application of RNN models was designed to differentiate AD patients from healthy control individuals using neuroimaging data.[14] Longitudinal EMRs were used to study the progression of chronic diseases like AD.[15,16] 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).[17,18] 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.[19] 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. 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.[20] 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.[21]

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 (SCRP)** dataset. One SCRP 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 1.

| Dataset                                           | Number of patients |
|---------------------------------------------------|--------------------|
| Naive 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 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 1).
Figure 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 naive 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 naive 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 naive 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).

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. 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 3.

**Figure 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 3. The average precision/recall comparison of models using condition, measurement, and an ensemble of two domains (combined).

The final SCRP 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 SCRP 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 SCRP dataset experiments, in Figure 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 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).
Figure 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 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. 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. 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

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 SCRP 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 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 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.

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 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 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.)

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.

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.[14] 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.[13] 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).[22]
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.[23] 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.[24] 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.[25] 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%.[26-29] Carpenter and Huang analyzed ML methods (naïve Bayes, kNN, SVM, random forest, and neural networks) for virtual screening (VS).[30] 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.[31] 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%.[32]

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 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.[9,32] EMR patient records provide a valuable resource for creating diagnostic support algorithms for the detection of dementia and AD.[9,32] Tang and collaborators attempted to predict differential diagnoses of the top 25 most common conditions in the MIMC-III dataset.[9] 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.[9] 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.[9] 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.[9] 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.[21]

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.

DECLARATION OF COMPETING INTERESTS

The authors declare no conflict of interest.
AUTHORS CONTRIBUTION

B.L., S.R. and X.H.C designed and implemented the method and conducted all the experiments. All authors were involved in writing the paper.

ACKNOWLEDGMENTS

This study was supported in part by the IQVIA grant Disease Detection and Disease Progression Modeling and by Pennsylvania Department of Health CURE Health Data Science Research Project.

REFERENCES

[1] https://www.nia.nih.gov/health/alzheimers-disease-fact-sheet.
[2] https://www.alz.org/media/Documents/alzheimers-facts-and-figures-2019-r.pdf
[3] 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.
[4] T. Wang, R.G. Qiu, M. Yu. 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.
[5] R. Roberts, D.S. Knopman. Classification and epidemiology of MCI. Clin Geriatr Med 2013;29(4):753-72. doi: 10.1016/j.cger.2013.07.003.
[6] L. Mosconi, M. Brys, L. Glodzik-Sobanska, S. De Santi, H. Rusinek, M.J. de Leon. Early detection of Alzheimer's disease using neuroimaging. Exp Gerontol 2007;42(1-2):129-38.
[7] P.M. McMahon, S.S. Araki, E.A. Sandberg, P.J. Neumann, G.S. Gazelle. Cost-effectiveness of PET in the diagnosis of Alzheimer's disease. Radiology 2003;228(2):515-22.

[8] J. Zhao, P. Papapetrou, L. Asker, H. Bostrom. Learning from heterogeneous temporal data in electronic health records. http://dx.doi.org/10.1016/j.jbi.2016.11.006.

[9] F. Tang, C. Xiao, F. Wang, J. Zhou. Predictive modeling in urgent care: a comparative study of machine learning approaches. JAMIA Open 2018;1(1):87–98.

[10] R.G. Qiu, J.L. Qiu, Y. Badr. 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

[11] E. Moradi, A. Pepe, C. Gaser, H. Huttunen, J. Tohka. 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.

[12] S. Esmaeilzadeh, D.I. Belivanis, K.M. Pohl, E. Adeli. (2018) 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.

[13] J. Albright. 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

[14] M Aghili, S Tabarestani, M Adjouadi, E. Adeli. 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.
[15] B.A. Goldstein, A.M. Navar, M.J. Pencina, J.P.A. Ioannidis. 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.

[16] J. Yang, J McAuley, J. Leskovec, P. LePendu, N. Shah. 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.

[17] S. Hochreiter, J. Schmidhuber. Long Short Term Memory. Neural Computation 1997;9(8):1735-80.

[18] A. Sherstinsky. Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network. arXiv:1808.03314v4.

[19] https://www.cdc.gov/coronavirus/2019-ncov/index.html.

[20] https://www.ohdsi.org/data-standardization/the-common-data-model.

[21] B. Ljubic, A. Abdel Hai, M. Stanojevic, W. Diaz, D. Polimac, M. Pavlovski, Z. Obradovic. Predicting Complications of Diabetes Mellitus Using Advanced Machine Learning Algorithms. JAMIA, 2020. (in press).

[22] E.N. Marzban, A.M. Eldeib, I.A. Yassin, Y.M. Kadah. Alzheimer's Disease Diagnosis From Diffusion Tensor Images Using Convolutional Neural Networks. PLoS One 2020;15(3):e0230409.

[23] M. R. Arbabshirani, S. Plis, J. Sui, V.D. Calhoun. Single Subject Prediction of Brain Disorders in Neuroimaging: Promises and Pitfalls. Neuroimage. 2017;145(Pt B):137-165.

[24] P.J. Moore, T.J. Lyons, J. Gallacher. Using Path Signatures to Predict a Diagnosis of Alzheimer's Disease. PLoS One 2019;14(9):e0222212.
[25] F.J. Martinez-Murcia, A. Ortiz, JM. Gorriz, J. Ramirez, D. Castillo-Barnes. 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.

[26] TE. Kam, H. Zhang, D. Shen. 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

[27] A. Moscoso, J. Silva-Rodriguez, J. M. Aldrey, J. Cortés, A. Fernández-Ferreiro, N. Gómez-Lado, A Ruibal, P Aguiar. 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.

[28] P J Moore, T J Lyons, J Gallacher. 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.

[29] S. Spasov, L. Passamonti, A Duggento, P. Liò, N. Toschi. 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.

[30] K.A. Carpenter, X. Huang. 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.

[31] L. Huang, Y. Jin, Y. Gao, KH. Thung, D Shen. Longitudinal Clinical Score Prediction in Alzheimer's Disease With Soft-Split Sparse Regression-Based Random Forest. doi:10.1016/j.neurobiolaging.2016.07.005.
[32]. E. Ford, P. Rooney, S. Oliver, R. Hoile, P. Hurley, S. Banerjee, H. van Marwijk, J. Cassell. 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.
Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: