Predicting Dementia With Routine Care EMR Data

Zina Ben Miled[a,b,*], Kyle Haas[a], Christopher M. Black[c], Rezaul Karim Khandker[c], Vasu Chandrasekaran[c], Richard Lipton[d], Malaz A. Boustani[b,e]

[a] *Department of Electrical and Computer Engineering, School of Engineering and Technology, Indiana University Purdue University at Indianapolis, 723 W. Michigan Street, Indianapolis, Indiana 46202 USA.*

[b] *Regenstrief Institute, Inc., 1101 W. 10th Street, Indianapolis, Indiana 46202 USA.*

[c] *Merck & Co., Inc., 2000 Galloping Hill Road, Kenilworth, New Jersey 07033 USA.*

[d] *Albert Einstein College of Medicine, 1225 Morris Park Avenue, Bronx, New York 10461.*

[e] *Indiana University School of Medicine, Center for Aging Research, 340 W. 10th Street, Suite 6200, Indianapolis, Indiana 46202 USA.*

Abstract

Our aim is to develop a machine learning (ML) model that can predict dementia in a general patient population from multiple health care institutions one year and three years prior to the onset of the disease without any additional monitoring or screening. The purpose of the model is to automate the cost-effective, non-invasive, digital pre-screening of patients at risk for dementia.

Towards this purpose, routine care data, which is widely available through Electronic Medical Record (EMR) systems is used as a data source. These data embody a rich knowledge and make related medical applications easy to deploy at scale in a cost-effective manner. Specifically, the model is trained by using structured and unstructured data from three EMR data sets: diagnosis,

[*] corresponding author
Email address: zmiled@iupui.edu (Zina Ben Miled)

Preprint submitted to Artificial Intelligence in Medicine *November 24, 2019*

prescriptions, and medical notes. Each of these three data sets is used to construct an individual model along with a combined model which is derived by using all three data sets. Human-interpretable data processing and ML techniques are selected in order to facilitate adoption of the proposed model by health care providers from multiple institutions.

The results show that the combined model is generalizable across multiple institutions and is able to predict dementia within one year of its onset with an accuracy of nearly 80% despite the fact that it was trained using routine care data. Moreover, the analysis of the models identified important predictors for dementia. Some of these predictors (e.g., age and hypertensive disorders) are already confirmed by the literature while others, especially the ones derived from the unstructured medical notes, require further clinical analysis.

Keywords: Dementia, Prediction, Random Forest, EMR, Machine Learning.

1. Introduction

The recent increase in life expectancy, although desirable, has also resulted in an increase in the number of persons affected by chronic diseases [1]. Indeed, according to the CDC, one half of the US adult population has one or more chronic disease [2]. Moreover, chronic diseases are not only the leading cause of deaths in the US, but also contribute to lower quality of life and carry a substantial burden on the family, the patient, and the health care system [3]. In particular, dementia affects a large number of the adult population. For instance, it is estimated that 5.7 million Americans are living with Alzeihmer's dementia [4] and this chronic disease is projected to cost in excess of 1 trillion US dollars annually by 2050 in the US [3]. Identifying persons likely to develop dementia in the future can help support the development of preventive interventions.

The proposed predictive model is based on a random forest (RF) classifier which is trained by EMR data of dementia and non-dementia patients from a large number of health institutions in Indiana. These data are obtained from the Indiana Network for Patient Care and Research through the Regenstrief Institute. Prescriptions (Rx), diagnosis (Dx) and medical notes (Nx) are extracted from the records of the patients. The Rx and Dx data are structured, while the Nx data are unstructured (i.e., free text). These data sets are augmented with demographic information (i.e., age, gender, race and institution affiliation). Different dimension reduction techniques are used for each data set and the resulting data sets are used to derive models with one and three years prediction horizons prior to the onset of dementia.

This prediction horizon may provide the opportunity for early screening and for delaying the onset of the disease through adequate health plan-based interventions. In fact, it was estimated that delaying the onset of the disease by one year can help reduce cost by as much as 14% [3].

2. Background

Extensive research work has been devoted to developing data-driven ML models that can automate disease diagnosis and prognosis in recent years [5, 6]. This section includes a brief review of related ML studies from the following aspects: 1) data sources, 2) feature engineering, 3) ML techniques, and 4) target applications. The review primarily focuses on dementia and discusses the scalability and interpretability of the sources and techniques as these represent key requirements for our proposed model. A summary of a selected set of most relevant previously proposed ML models is included in Table 1.

2.1. Data Sources

Data sources for ML models fall under two broad categories: structured (e.g., diagnosis, prescriptions, images, medical tests, etc...) and unstructured (e.g., medical notes).

Most previous dementia models are derived by using structured data from targeted medical tests (e.g., MRI [10, 8, 12, 13, 14] and cognitive tests [7, 15, 16, 17]). For instance,

- Cognitive tests for a cohort of 400 patients are used in [7],
- Neuropsychological tests for a cohort of 321 patients are used in [9],
- A combination of MRI images and cognitive tests are used in [10] for a cohort of 825 MCI patients, and
- MRI images, PET scans and cerebrospinal fluid (CSF) from 186 patients are used in [8].

The extensive use of structured data is not limited to dimentia studies but also extends to other healthcare areas with a large portion of structured data being images [6]. This data category is appealing because it requires relatively limited pre-processing compared to unstructured data. However, the underlying tests are expensive to administer which in turn restricts the

Table 1: Summary of target applications, data sources, machine learning techniques, prediction horizons and accuracies for a selected set of most relevant previous studies. X/Y is used to denote the classes of the model (e.g., AD/HC denotes AD patients versus HC). MCI-to-AD is used to denote the conversion from MCI patients to AD patients. When multiple techniques are used in a study, the technique with the highest performance is reported. HC = Healthy Controls, MCI = Mild Cognitive Impairment, AD = Alzheimer's Disease. (*) Dementia, AD, MCI not included.

	Target Disease	Data Source	ML Technique	Pred. Horizon	Acc.
[7]	MCI-to-Dementia	Cognitive tests	SVM	0.5-4 yrs	76%
[8]-a	MCI/HC, AD/HC	MRI, PET, CSF	SVM	0	83%, 93%
[8]-b	MCI-to-AD	MRI, PET, CSF	SVM	2 yrs	74%
[9]	AD/MCI/HC	Neurophysilogical tests	BN	0	83%
[10]	MCI-to-AD	Cognitive tests, MRI	RF	0-3yrs	82%
[11]	multiple*	EMR	RF	0-1yr	> 85%
proposed	HC-to-dementia	EMR	RF	1 yr	77%
proposed	HC-to-dementia	EMR	RF	3 yrs	74%

inclusion criteria as exemplified by the above studies. Moreover, this aspect also limits the scalability of the models derived from these sources to a large population.

Wearable and home devices, an emerging source of structured data, may help overcome the above restriction. For example, accelerometor data from an ankle bracelet was used in [18] to model gait and movement activities for dementia patients and healthy controls (HC). A multisensor home device was used to monitor the daily activities of ten dementia patients and HC in [19]. A comprehensive discussion on the use of these devices and their potential contribution to dimentia detection and monitoring is provided in [20]. The authors mention that while these devices can become ubiquitous, their main disadvantage is privacy infringement. That said, as the sensor technology continues to mature, it is foreseeable that EMR records will start incorporating this source of patient-provided data.

Unstructured data in health applications is primarily extracted from medical notes. Using this source of data is a new area of research [5]. It was used to develop models for different disease conditions (e.g., Type-2 diabetes [21], heart failure [22], colorectal cancer [23] and early readmission for psychiatric patients [24]). A large number of patient health records from a single health care institution consisting of both structured and unstructured data was used in [11] for the prediction of the onset of multiple disease conditions.

The use of medical notes for the development of dementia models in the literature is scarce. One example study [25] considers text notes for visits and medical history from the ADNI data set. The model proposed in this paper also uses structured and unstructured data and confirms the findings of previous studies [11, 25] which concluded that combining these two data sources yields ML models with better performance than each source individually.

There are recent sources of unstructured data that are gaining importance in health care applications and may benefit dementia prediction. For example, transcripts of daily conversation are used in [26] to identify social behavior (e.g., giving advice, receiving advice, conversation). Another study [27] uses speech samples and MRI data from 32 semantic dementia patients and 10 HC.

2.2. Feature Engineering

Both structured and unstructured data must be transformed into a feature vector that can be used as an input to the ML model. Feature selection for structured data is relatively well studied. In the case of numerical data (e.g., age, blood pressure) and categorical data (e.g., gender, disease condition), the process is either guided by human experts [17] or relies on a well established nomenclature (e.g., disease codes or drug groups) [11]. A more difficult procedure is required to transform free text into structured features [5].

A preliminary step in processing free text consists of syntactic transformations such as removing punctuation and stop words [26]. This is followed by a process which identifies features that are representative of the text. Techniques for feature extraction from free text range from frequency-based to more advanced techniques that employ autoencoders and deep neural networks. A review of the latter two techniques in health care applications is provided in [5].

Selecting features based on the most frequent keywords is referred to as bag of words and was used in [25]. A mapping technique that can further reduce the feature space and enhance the entropy of each feature consists of grouping a set of related keywords under a single concept. For example, using latent Dirichlet allocation (LDA), keywords are grouped by concept in [11] and [26]. Word2vec [28] and autoencoders have also been used for keyword-to-concept mapping in several health care applications [5]. LDA, word2vec and autoencoders are computationally expensive [29] and may select features that are not interpretable [30]. For these reasons and because of the size of the corpus which is derived from a large number of different types of medical notes in this paper, we designed a variant of the bag of words. This approach is inspired by [28] and selects features based on the differential frequency of the keywords in dementia cases and healthy controls rather than an absolute frequency. The modified bag of words used in this study is scalable and interpretable.

2.3. ML Techniques

The choice of the ML technique is highly dependent on the application domain and the source of the data [6]. For instance, support vector machine (SVM) [31], RF [32] and aritificial neural networks (ANN) [33] are used in [7] to identify conversion from mild cognitive impairment (MCI) to dementia. The SVM model had the highest accuracy (76%) and the RF model ranked second best (73%). However, the SVM model had significantly lower sensitivity. RF is also used in [10] to predict conversion from MCI to Alzheimer's Disease (AD) with an accuracy of 82%. In [8], SVM was used to classify MCI and AD patients versus HC with an accuracy of 83% and 93%, respectively.

Naive Bayesian network (BN), SVM and ANN are used in [9] to develop a multi-class classifiers for MCI, AD and HC with accuracies of approximately 83%, 60% and 82%, respectively.

It is hard to adequately compare the performance of different ML techniques across multiple studies because of variances in the sources of data, features and number of records in the training and validation data sets. However, one common aspect of all of the above studies is that they use structured data with relatively small number of features.

Despite feature reduction, when a model uses unstructured data, the number of features still remains large and thus the feature space is highly dimensional. RF is often the ML technique of choice in these cases. For example, in [11] once feature reduction is performed on the medical notes by using LDA, a multi-class RF classifier is trained to identify future disease conditions for each patient. SVM and RF are used in [26] to classify features extracted from transcription of daily conversations by using LDA into different social behavior classes. In this case, RF had better accuracy and precision than SVM.

For the dementia prediction model being proposed in this paper, we opted to use RF. This choice was motivated by several factors that were derived from the literature and from our own preliminary investigation. Namely, RF is interpertable, computationally efficient and can handle a high dimensional space of noisy, continuous and categorical features [34, 35, 36]. That said, we also experimented with SVM and ANN. SVM had a comparable accuracy but was not as interpretable as RF [35]. ANN had a lower accuracy compared to both RF and SVM, primarily because the number of available patient records is much lower than what would be needed to adequately train a high dimensional deep neural network. Similar observations are made in [26]. In addition, ANNs are also not interpretable [5]. BN was not considered since an efficient BN model requires significant engineering effort and is computationally demanding especially for a model with a large number of features [32].

2.4. Applications

The accuracy of predictive ML applications depends on the prediction horizon and the ability to infer the outcome from the evidence. Applications with short-term horizons support diagnosis whereas applications with long-term horizons support prognosis. The latter are for early pre-screening and therefore are most useful when the horizon is long enough to allow adequate intervention.

Typically, short term horizon applications are based on evidence available up to the time of occurance of the target outcome. This temporal proximity allows the use of evidence with a higher entropy. For example, in [9], the evidence (e.g., age, gender, education, functional ability and visual memory) is used to classify patients into three classes: HC, MCI and dementia.

In [8], the authors first create short prediction horizons models that classifies MCI versus HC and AD versus HC. These models show that discriminating between AD and HC has a higher accuracy compared to MCI versus HC. This result is anticipated since AD indicators are typically stronger. The authors then create a model dedicated to the conversion of MCI to AD with a prediction horizon of 2 years. This model did not have access to evidence within the prediction horizon, and therefore had a lower accuracy than both short term models (Table 1). Models for HC to AD or HC to MCI conversion with a 2 year horizon were not available for comparison with the models proposed in this paper.

MCI to dementia and MCI to AD conversion models are introduced in [7] and [10], respectively. These models did not segregate patients according to the prediction horizon. Therefore, the outcome and the evidence can be from 0.4 to 4 years and 0 to 3 years apart, respectively.

A general model for the prediction of the onset of multiple diseases (e.g., cancer, diabetes, and schizophrenia) with an accuracy greater than 85% was introduced in [11]. As mentioned above, this model uses structured and unstructured EMR data from a single institution. However, dementia is not included as a target disease. Moreover, the prediction horizon is 0 to 1 year from the time of evidence and varies across the patients.

In the current study, we removed the records of MCI patients in order to ensure that they do not bias the proposed model which is intended for the prediction of dementia in a general patient population. Furthermore, no evidence is used within the prediction horizon of 1 year or 3 years for all patients.

3. Methods

In order to train and test the proposed models, a group of dementia (cases) and non-dementia (controls) patients are identified. The processing steps used to extract the training and testing data from the EMR database as well as the methodology used to develop the predictive models are discussed in the next two subsections.

3.1. Data Preprocessing

The dementia cases are identified by using their diagnosis code. Only cases of incident dementia are retained. MCI or prevalent dementia cases are excluded in order to avoid the bias they may introduce in the model. The diagnosis date (index date) for each case is then established and 3 to 4 matching controls are identified for each case. The matching criteria between the cases and controls are based on birth year, gender, race and index date (within 6 months).

For both cases and controls, a query was developed to retrieve all the medical records within 10 years prior to the index date. Only patients that had at least one encounter per year were retained. However, patients may or may not have complete records that span the entire 10 year period (e.g., patients that seek additional treatment outside the network, partially migrated patient records from a legacy EMR system). This query resulted in 2,159 cases and 11,558 controls from 15 and 25 different institutions, respectively. The distribution of both cases and controls across the institutions is not uniform. Moreover, patients may have records spanning multiple institutions. With respect to race and gender, the distribution of the cases and controls (Table 2) is similar, thereby limiting any gender or race bias among the two classes. However, Table 2 shows that within a class, there are more females and more patients of race white than the other gender and races. This reflects the demographic of incident dementia patients in Indiana.

Features are extracted from the prescription (Rx), diagnosis (Dx) and medical notes (Nx) of the EMR record of each patient. Age, gender, race and institution are also included as features in all models. In the case of insti-

Table 2: Demographics of the cases and controls.

	African American		White		Other	
	Male	Female	Male	Female	Male	Female
Cases	10%	21%	23%	41%	2%	4%
Controls	11%	19%	24%	42%	1%	3%

tution, if a patient is associated with multiple institutions, for each data set, the most recent institution on record during the model period was selected.

Feature reduction for each data set is performed by using a different approach. Each Rx feature corresponds to a drug group according to the Generic Product Identifier (GPI) classifier [37] for a total of 100 features. The value of the feature is the number of times a medication from the given drug group was prescribed for the patient within the model period. Age, gender, race and institution (InstRx) are added to the drug groups in order to construct a feature vector with 104 features. A similar approach was used in [35].

A total of 19 dementia related disorders (Table 3) are identified using expert opinion. These groups of disease conditions are represented by using ICD-10 or ICD-9 codes in the EMR database. Codes for the target disease groups were identified and the relevant records for each patient were extracted. The value for each feature in the Dx data sets corresponds to the count of the diseases in the corresponding disease group during the model period. As in the case of prescriptions, age, gender, race and institution (InstDx) are added to these features in order to form a Dx feature vector with 23 features.

Table 3: List of dementia related disorders or disease groups in Dx.

Angina	Chronic Ischemic Heart Disease
Anxiety	Transient Ischemic Attack
Abnormal Weight Loss	Transient Ischemic Attack Related Syndromes
Bipolar Disorder	Other Acute Ischemic Heart Disease
Depression	Stroke/Cerebral Infarction
Insomnia	Acute/ Subesequent MI
Hypercholesterolemia	Hemorrhagic Cerebrovascular accident
Hypertensive Disorders	other Cardiovascular diseases
Schizophrenia	Claudication/Atherosclerosis
Sleep Apnea	

Medical notes are a sequence of records starting with the patient's unique ID, the date of the record followed by a list of reports. Each report consists of a report type and a report content in the form of unstructured free text. There is a total of 2,146 different report types. This large number of report types is due, in part, to the fact that each institution may use its own set of reports. In addition, there is a large number of administrative reports such as "attending MD", "encounter ID", "signature", "enterer ID", "verified by" and "dictated by". These administrative reports are excluded. A total of 340 different report types out of the original 2,146 report types are retained. The retained report types consist of notes or findings recorded by the health care provider such as "radiology impression", "recommendation", "history of present illness", "md progress note" and "social history". An example report is shown below:

```
<p>EXAM: KUB SINGLE VIEW</p><p>HISTORY:  Recurrent left flank
pain. History of nephrolithiasis.</p><p>FINDINGS:</p><p>The
bowel gas pattern is nonspecific but does not appear grossly
<br/>obstructive.</p><p>The renal shadows are obscured by
stool/bowel gas. No definitive<br/>calculi are appreciated.
</p><p>IMPRESSION: Negative KUB.</p><p>Read By: Malaz Boustani
<br/>Reviewed and Electronically Signed By: Malaz Boustani<br/>
</p></text></text_report>
```

The processing of free text in the medical notes start with the preliminary syntactic transformation described in Section 2.2. This is a procedure which is commonly applied to free text [6]. When using the bag-of-word approach, the next step consists of selecting the keywords with the highest count as features [24]. However, some of the frequent keywords may be informative (e.g., abdomen) while others (e.g., patient, doctor, nurse) may not include significant information towards the target outcome. In fact, the sub-sampling of some frequent words should be considered as recommended in [28].

In the traditional bag of word, the keyword count is talied in all notes irrespective of whether the note is attributed to cases or controls. Inspired by [28], we calculate the count of each keyword with respect to the class (i.e., cases or controls). For example, the keyword *abdomen* occurs 16,988 times in cases reports and 65,635 times in controls reports. Based on the number of cases and controls in the study, the difference in relative count for this keyword is $16,988/2,159$ (cases) - $65,635/11,558$ (controls) $= 2.19$. This differential frequency is defined as follows:

$$DF(keyword) = \left| \frac{count(keyword|cases)}{number\ of\ cases} - \frac{count(keyword|controls)}{number\ of\ controls} \right| \quad (1)$$

The value of DF tends to be closer to zero for equally frequent keywords in cases and controls. However, when the keyword is more prevalent in cases than controls or vice-a-versa, the value of DF is high.

In this study, a total of 173,972 unique keywords are identified. Clusters are constructed from these keywords using Algorithm 1. First, each keyword with $DF > 1.0$ is selected as a seed of a cluster. Other keywords are assigned to these clusters if their distance to the keywords in the cluster is less than a given threshold. This assignment is human-validated at each iteration with the assistance of an online medical dictionary. Most of the clusters included a large number of keywords for various reasons. Spelling errors are the most common reason. For instance, there are 43 different incorrect spelling of the word abdomen (e.g., abbdomen, abdoomen, obdomin, etc.). Conceivably, spelling errors can be eliminated with a spell checker. However, this implies that medical notes from future patients need to also be corrected for spelling. In order to avoid this processing step, the incorrect spellings of each keyword were kept in the cluster.

In total, 110 clusters were constructed by using the above procedure. The resulting Nx feature vector includes these clusters of keywords in addition to age, gender, race and InstNx for a total of 114 features.

3.2. Model Development

The previous data processing steps are used to construct Rx, Dx, and Nx training and testing data sets for various dementia prediction models. The

```
Data: Keyword list L, number of cases, number of controls
Result: A set of clusters where each cluster is an Nx feature
for $k_i \in L$ do
    if $DF(k_i) > 1.0$ then
        assign $k_i$ to $c_i$
    end
end
$C = set\ of\ c_i$
while not done do
    foreach $k_j \notin C$ do
        foreach $k_i$ do
            if $distance(k_j, k_i) < threshold$ then
                assign $k_j$ to $c_i$
            end
        end
        validate
    end
end
```

Algorithm 1: Clustering algorithm for NX features.

$1Yr$ model is trained and tested by using the available records during the period (index date - 10 years) to (index date - 1 year). Similarly, the $3Yr$ model is trained and tested by using the available patient records during the period (index date - 10 years) to (index date - 3 years). The aim of these models is to predict dementia one year and three years prior to the onset of the disease.

The number of cases and controls in the training and testing data sets for each model is shown in Table 4. In order to avoid class imbalance a 50/50 split between cases and controls is maintained in the training data set. Moreover, for cases an 80/20 split is maintained between the training and testing data sets.

Table 4: Number of cases and controls in the training and testing data sets for each dementia model.

Model	Training		Testing	
	Cases	Controls	Cases	Controls
$1Yr$	1,728	1,728	431	9,830
$3Yr$	869	869	216	4,817
$1Yr^s$	1,225	1,225	299	2,167

Most of the patients in the data set are affiliated with multiple institutions, in which case the value of InstRx, InstDx, and InstNx represents the most recent institution on record for the patient during the model period. The $1Yr^s$ model (Table 4) has the same period as the $1Yr$ model. However, it only uses patients that are affiliated with a single institution. Patients that are affiliated with multiple institutions in a given data set are excluded. This model was developed in order to better understand the impact of institution affiliation.

RF is used to train all the models. The hyperparameter of the RF include the number of trees and the number of features which are presented to each node of the tree. In [38], a number of trees in the RF equal to the number of features in the model is recommended. In [34], it is recommended to increase the number of trees until the model stabalizes. In this study, we experimented with 100, 500, and 1000 trees and did not find a significant difference. The performance metrics reported in the result section are for 500 trees in each model. The second hyperparameter is the number of features presented to each node in the tree. In this study, it is set to $\sqrt{n}$, where n is the number of features in the model as recommended in [32]. Moreover, the best dementia vs non-dementia classification at each node is measured by using the Gini Impurity (GI) [39].

When the models are validated using the test data sets, the accuracy, sensitivity and specificity of the models are computed. In addition, features that are strong predictors for each model are identified. The approach used to identify these predictors is based on the selection of each node in the classification decision of the patient [40]. This approach is used in [40] and [41] to identify important features for RF models aimed at predicting cardiovascular risk and cancer mortality, respectively. As each patient in the test data set is classified, a count is maintained of the number of nodes associated with each feature that are traversed in the decision tree. This count represents the number of times each feature (f_i) participates in the classification decision. This metric is labeled $CD(f_i)$ in the remainder of the paper. As in [40] and [41], the features with the highest $CD(f_i)$ are reported as strong predictors.

4. Results

The performance metrics corresponding to the various models are shown in Table 5. These metrics are the mean and standard deviation across the 5 groups in a 5-fold cross validation. The first three sections of this table are dedicated to the models developed by using the Rx, Dx and Nx data sets, respectively. The last two sections report the results associated with models that are trained by using a data set that combines the Rx, Dx and Nx. This combined data set (RDNx) consists of all the features from the above three data sets where the duplicate age, gender and race features are removed for a total of 235 features. As previously mentioned, each data set includes a feature that represents the institution of the patient (e.g., InstRx for the Rx data set). The combined RDNx data set includes the three institutions from the three underlying data sets, namely, InstRx, InstDx, and InstNx. The last section of Table 5 reports the results associated with a model developed using the combined data set without these three institution features (RDNx w/o Inst).

Table 6 shows the top predictors for cases and controls for the models developed by using the individual data sets. Similarly, Table 7 shows the top predictors for the models developed by using the combined data sets. Only the top five ranked features are reported (i.e., with the highest five CD). The top predictors can be different for cases and controls as shown in tables 6 and 7.

An effort was made to train the models with the same patients. This was done in order to avoid variations due to different patients (e.g., complete vs. incomplete records) and number of patients. In general, the more records

Table 5: Accuracy, sensitivity and specificity for the 1Yr and 3Yr dementia prediction models trained by using different data sets. For each metric, the entry in the table corresponds to the mean value of all groups in a 5-fold cross validation and the number in parenthesis is the standard deviation across the groups.

	Model	Accuracy (%)	Sensitivity (%)	Specificity (%)
Rx	$1Yr$	70.39 (0.88)	68.94 (2.35)	70.46 (0.98)
	$3Yr$	65.63 (1.24)	65.00 (3.91)	65.65 (1.43)
	$1Yr^s$	67.93 (0.94)	72.24 (3.02)	67.33 (1.47)
Dx	$1Yr$	65.21 (0.74)	66.06 (2.44)	65.18 (0.80)
	$3Yr$	62.91 (1.71)	63.80 (4.26)	62.87 (1.95)
	$1Yr^s$	69.27 (1.69)	65.15 (3.37)	69.84 (2.22)
Nx	$1Yr$	74.07 (0.98)	72.01 (1.72)	74.16 (1.01)
	$3Yr$	70.13 (2.65)	67.31 (3.28)	70.25 (2.85)
	$1Yr^s$	78.47 (1.44)	73.04 (2.98)	79.22 (1.97)
RDNx	$1Yr$	77.43 (1.89)	76.01 (1.88)	77.49 (2.02)
	$3Yr$	73.50 (2.03)	70.93 (2.18)	73.61 (2.17)
	$1Yr^s$	79.68 (1.29)	76.35 (2.10)	80.15 (1.30)
RDNx w/o Inst	$1Yr$	72.64 (0.34)	73.97 (3.23)	72.58 (0.41)
	$3Yr$	67.88 (1.82)	68.52 (3.48)	67.85 (1.96)
	$1Yr^s$	78.60 (0.32)	75.05 (1.22)	79.09 (0.28)

Table 6: Strong predictors for models trained by using individual data sets. Features with the top five CD are reported for each model. When the strong predictors for the two classes (i.e., cases/controls) differ, they are reported separately.

Rx

$1Yr$	InstRx, Age, Antidepressants, Diuretics, Antihyperlipidemics
$3Yr$	InstRx, Age, Antihyperlipidemics, Antidepressants, Antihypertensives
$1Yr^s$	InstRx, Age, Antidepressants, Psychotherapeutic & Neurological Agents - Misc., Analgesics - Opiod

Dx

$1Yr$	InstDx, Age, Hypertensive Disorders, Hypercholesterolemia, Chronic Ischemic Heart Disease
$3Yr$	InstDx, Age, Hypertensive Disorders Hypercholesterolemia, Chronic Ischemic Heart Disease
$1Yr^s$	*cases*: InstDx, Age, Hypertensive Disorders, Depression, Hypercholesterolemia
	controls: InstDx, Age, Hypertensive Disorders, Hypercholesterolemia, Chronic Ischemic Heart Disease

Nx

$1Yr$	*cases*: InstNx, Age, accurate, independently, oral
	controls: InstNx, Age, accurate, independently, oldest
$3Yr$	*cases*: InstNx, Age, accurate, oral, independently
	controls: InstNx, Age, accurate, independently, oral
$1Yr^s$	*cases*: InstNx, accurate, oldest, independently, participate
	controls: InstNx, Age, radiology, accurate, independently

Table 7: Strong predictors for models trained by using the combined data sets. Features with the top five CD are reported for each model. When the strong predictors for the two classes (i.e., cases/controls) differ, they are reported separately.

RDNx

$1Yr$	*cases*: InstNx, InstRx, Age, InstDx, oldest
	controls: InstNx, InstDx, InstRx, oldest, Age
$3Yr$	*cases*: InstNx, InstDx, InstRx, oral, oldest
	controls: InstDx, InstRx, oral, oldest, Antihyperlidemics
$1Yr^s$	*cases*: InstNx, oldest, participate, independently, InstRx
	controls: InstNx, Age, InstRx, oldest, participate

RDNx w/o Inst

$1Yr$	*cases*: Age, accurate, oldest, independently, oral
	controls: Age, accurate, oldest, independently, participate
$3Yr$	Age, oldest, accurate, participate, oral
$1Yr^s$	*cases*: Age, participate, accurate, oldest, independently
	controls: Age, participate, accurate, oldest, oral

available to train an ML model, the higher the accuracy of the model. The quality of the data is also important. Missing and incomplete data may lead to poor predictive performance. In this study, the number of patients was not only limited by the number of cases but also the number of cases that had Nx data. Moreover, fewer patients are available for the training of the $3Yr$ compared to the $1Yr$ models (Table 4). This indicates that the data set includes a large number of patients that had less than three years of complete medical records prior to the index date. Similarly, there are fewer patients available for the training of the $1Yr^s$ compared to the $1Yr$ models (Table 4). The difference in this case is due to patients that tend to seek treatment from one institution versus multiple institutions.

5. Discussion

Several observations can be drawn from the results shown in Table 5. Some of these observations are expected while others are worth discussing. The $1Yr$ models have higher accuracy, sensitivity and specificity than the $3Yr$ models. This is expected since the former models are trained with more patients and have access to more recent records of the patients prior to the onset of the disease. For example, the $1Yr$ and $3Yr$ Rx models have an accuracy of 70.39% and 65.63%, respectively. They were trained with 3,456 and 1,738 patients, respectively.

Among the three data sets (i.e., Rx, Dx and Nx), Nx generated models with the highest accuracy, sensitivity and specificity. Moreover, the accuracy of the model derived by using the combined RDNx data set is higher than the models derived by using the individual data sets. This is an indication that, despite the fact that Nx models have a higher prediction accuracy, some of the Rx and Dx features (e.g., antihyperlidemics) make a significant contribution to the overall accuracy of the combined model.

The most interesting aspect of the analysis is, undoubtedly, associated with the features that contribute the most to the predictive models. As shown in tables 6 and 7, cases and controls nearly have the same top predictive features across all models. Moreover, with respect to patient's demographics, age is consistently among the top features of all the models for both cases and controls. This is particularly interesting since the match between cases and controls was based on age, gender and race. Gender and race do not, however, appear as top features in any of the models and are unlikely, according to these models, to be significant predictors of dementia.

The institution feature is also present in most models as a top feature. In most cases, the $1Yr^s$ models have higher accuracy than the $1Yr$ models despite the fact that they were trained using fewer patients. Moreover, the combined model based on the RDNx data set includes all three institutions from the Rx, Dx, and Nx data sets as top features. The importance of the institution as a feature in the predictive models is intriguing and we hypothesize that it may due to various reasons including the following:

- An artifact of the way the medical records of the patients are stored or extracted from the information systems of the various institutions.

- An indication of the completeness of the health record of the patients.

- An indication of the socio-economic demographics of the patients.

- A representative feature of the health institution's unique processes, culture and areas of expertise.

The first reason is unlikely since multiple institutions are covered by cases and controls in all the data sets. For instance, the training set of the $1Yr$ RDNx model included patients from more than 20 institutions. The higher accuracy of the $1Yr^s$ models compared to the $1Yr$ models indicates the possibility that patients that are affiliated with one institution tend to have more complete records whereas those that are affiliated with multiple institutions may have gaps in their records. There may also be key demographic differences between the different health institutions and the availability of information from certain institutions may bias the results. For instance, individuals may chose to seek care at different institutions for certain comorbidities. If this is the case, the $1Yr^s$ models are excluding patients with certain comorbidities and are more likely to retain healthier patients. These aspects will be investigated as part of future work.

In addition to the demographic features, several clinical features are important. For the $1Yr$ and $3Yr$ Rx models, "antidepressants", "diuretics", "antihyperlipedimics" and "antihypertensives" are top features. For the $1Yr^s$ Rx model, "diuretics" and "antihyperlipedimics" are less important than "psychotherapeutic & neurological agents" and "analgesics - opiod".

For the Dx data set, "hypertensive disorders", "hypercholesterolemia", "chronic ischemic heart disease", "depression" and "other cardiovascular diseases" have a high CD. These results validate the experts' opinion that these disease groups are important predictors. However, the results also show that the disease groups "other acute ischemic heart diseases" and "transient ischemic attack related syndromes" are not necessarily strong predictors. This finding is counterintuitive since "chronic ischemic heart disease" is a strong predictor. We believe that one potential explanation for the low importance of the disease groups "other acute ischemic heart diseases" and "transient ischemic attack related syndromes" is due to the low incidence rates associated with these features in the data set used in this study. Indeed, only 32 and 40 patients out of the 3,456 used for the training of the $1Yr$ Dx model had "other acute ischemic heart diseases" and "transient ischemic attack related syndromes", respectively.

As discussed in Section 3, the Dx models are developed using a top-down approach. In this case, most of the disease groups identified by the domain experts are confirmed by the model as strong predictors for dementia with the exception of "other acute ischemic heart diseases" and "transient ischemic attack related syndromes". Another validation of these features as strong predictors can be derived by comparison with the Rx models which were developed using a bottom-up approach (i.e., with no pre-conditioning). For instance the top feature "depression" in the Dx models aligns with the drug group "antidepressants" in the Rx models. Similarly, "hypertensive disorders" aligns with the "antihypertensives" drug group and "hypercholesterolemia" aligns with the "antihyperlipidemics" drug group. Moreover, "diuretics" are most commonly prescribed for "hypertensive disorders". This correlation between the top features of the Rx and Dx models reinforces the validity of the disease groups identified by domain experts as strong predictors.

In addition to the age and institution, the top features for the Nx model are “accurate”, “independently”, “oral”, “oldest”, “participate”, and “radiology”. These labels are representative of a cluster of keywords. For instance, the clusters “accurate” and “participate” have a $DF > 4.3$.

The top clinical features of the RDNx models include “oldest”, “oral”, “participate”, “independently” and “antihyperlidemics”. The first four are also top features of the Nx models and the fifth is a top feature of the Rx model. The three institution features (i.e., InstNx, InstDx and InstRx) also appear as top features in the combined RDNx models. They were omitted in the RDNx w/o Inst model and the top features of the resulting model are primarily from the Nx data set.

The results also show that the predictive accuracy of the $1Yr^s$ RDNx w/o Inst model is impacted the least by the removal of the institution feature and the $3Yr$ RDNx w/o Inst model is impacted the most. This fact seem to reinforce that the institution feature may be an indication of the completeness of the patient’s records.

As a final step in the analysis of the proposed models, we wanted to understand the distribution of the cases and controls that are accurately predicted by the proposed models. For this purpose, we focused on the combined RDNx models since they had the highest predictive accuracy. The goal of this analysis is to identify salient characteristics of the cases or controls that may have contributed towards a higher prediction accuracy. The cases and controls of the $1Yr$ and $3Yr$ RDNx models are affiliated with multiple institutions regardless of whether the prediction is correct or not. The cases and controls of the $1Yr^s$ RDNx models were selected based on their affiliation with a single institution. Patients from multiple institutions are included in the test data sets and there are no bias toward higher predictive accuracy with one institution versus another.

6. Conclusion

Models based on routine care EMR data collected prior to dementia can help identify high risk individuals from the general patient population and support the development of a customized health plan based interventions at various points in the progression trajectory of the disease.

This paper presents a methodology for constructing ML models for dementia prediction by using structured and unstructured routine care EMR data. Several RF models are developed from three EMR data sets, namely, drug prescription (Rx), diagnosis (Dx) and medical notes (Nx). The highest accuracy was obtained with a model that combines these three data sets and uses a total number of 235 features. This combined model is generalizable across multiple institutions and can predict dementia within one year of its onset with an accuracy of 77.43% , a sensitivity of 76.01% and a specificity of 74.16%. Moreover, the model was analyzed and was found not to be affected by biases related to institution affiliation, race or gender.

Using routine care EMR data to develop the proposed models makes them accessible to a wide range of patients at a reduced cost. Despite this fact, the proposed models have a performance comparable to dementia models that are based on specialized medical tests such as MRI and cognitive tests (Table 1). One potential use of the proposed models is for the pre-screening for dementia in the general patient population. A step that can then be followed by targeted medical tests for patients that are at-risk.

Top demographics and clinical features were identified by the models. In the demographic feature set, institution affiliation and age are found to be important predictors while race and gender are not. The clinical features are extracted from prescriptions, diagnosis and medical notes. The diagnosis features suggested by domain experts were confirmed by the exploratory models developed from the prescriptions data set. The study also shows that medical notes are the best source of predictive features. These features require further analysis in order to understand their clinical relationship to cognitive and executive functioning.

Future work include improving the proposed clustering approach and potentially combining it with word embedding techniques while maintaining interpretability. We also would like to pursue an aggressive feature reduction technique as a minimalist model is more cost-effective to deploy in production across multiple institutions.

7. Acknowledgments

This research was supported in part by Merck Sharp & Dohme Corp., a subsidiary of Merck & Co., Inc., Kenilworth, NJ, USA. The authors would also like to thank Paul Richard Dexter, Stephen Duong, Jarod Baker, and Jane Wang of the Regenstrief Institute for their support and feedback.

References

[1] R. A. Goodman, S. F. Posner, E. S. Huang, A. K. Parekh, H. K. Koh, Peer reviewed: Defining and measuring chronic conditions: Imperatives

for research, policy, program, and practice, Preventing chronic disease 10 (2013).

[2] B. Ward, J. Shiller, A. Goodman, Multiple chronic conditions among us adults: A 2012 update.(2014), Prevention of Chronic Disease, April (11) (2015) 130389–91.

[3] 2018 alzheimer's disease facts and figures 14 (2018) 367–429.

[4] Alzheimer's disease, 2018 [internet], 2018. URL: https://www.cdc.gov/dotw/alzheimers/.

[5] B. Shickel, P. J. Tighe, A. Bihorac, P. Rashidi, Deep ehr: a survey of recent advances in deep learning techniques for electronic health record (ehr) analysis, IEEE journal of biomedical and health informatics 22 (2017) 1589–1604.

[6] F. Jiang, Y. Jiang, H. Zhi, Y. Dong, H. Li, S. Ma, Y. Wang, Q. Dong, H. Shen, Y. Wang, Artificial intelligence in healthcare: past, present and future, Stroke and vascular neurology 2 (2017) 230–243.

[7] J. Maroco, D. Silva, A. Rodrigues, M. Guerreiro, I. Santana, A. de Mendonça, Data mining methods in the prediction of dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests, BMC research notes 4 (2011) 299.

[8] D. Zhang, D. Shen, A. D. N. Initiative, et al., Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in alzheimer's disease, NeuroImage 59 (2012) 895–907.

[9] J. A. Williams, A. Weakley, D. J. Cook, M. Schmitter-Edgecombe, Machine learning techniques for diagnostic differentiation of mild cognitive impairment and dementia, in: Workshops at the twenty-seventh AAAI conference on artificial intelligence, 2013, pp. 71–76.

[10] E. Moradi, A. Pepe, C. Gaser, H. Huttunen, J. Tohka, A. D. N. Initiative, et al., Machine learning framework for early mri-based alzheimer's conversion prediction in mci subjects, Neuroimage 104 (2015) 398–412.

[11] R. Miotto, L. Li, B. A. Kidd, J. T. Dudley, Deep patient: an unsupervised representation to predict the future of patients from the electronic health records, Scientific reports 6 (2016) 26094.

[12] R. Chen, E. H. Herskovits, Machine-learning techniques for building a diagnostic model for very mild dementia, Neuroimage 52 (2010) 234–244.

[13] Y. Zhang, Z. Dong, P. Phillips, S. Wang, G. Ji, J. Yang, T.-F. Yuan, Detection of subjects and brain regions related to alzheimer's disease using 3d mri scans based on eigenbrain and machine learning, Frontiers in Computational Neuroscience 9 (2015) 66.

[14] L. Khedher, J. Ramírez, J. M. Górriz, A. Brahim, F. Segovia, A. s Disease Neuroimaging Initiative, et al., Early diagnosis of alzheimer s disease based on partial least squares, principal component analysis and

support vector machine using segmented mri images, Neurocomputing 151 (2015) 139–150.

[15] W. R. Shankle, S. Mani, M. J. Pazzani, P. Smyth, Detecting very early stages of dementia from normal aging with machine learning methods, in: Conference on Artificial Intelligence in Medicine in Europe, Springer, 1997, pp. 71–85.

[16] S. P. Woods, A. I. Tröster, Prodromal frontal/executive dysfunction predicts incident dementia in parkinson's disease, Journal of the International Neuropsychological Society 9 (2003) 17–24.

[17] F. L. Seixas, B. Zadrozny, J. Laks, A. Conci, D. C. M. Saade, A bayesian network decision model for supporting the diagnosis of dementia, alzheimer s disease and mild cognitive impairment, Computers in biology and medicine 51 (2014) 140–158.

[18] T. Kirste, A. Hoffmeyer, P. Koldrack, A. Bauer, S. Schubert, S. Schröder, S. Teipel, Detecting the effect of alzheimer's disease on everyday motion behavior, Journal of Alzheimer's Disease 38 (2014) 121–132.

[19] P. Urwyler, R. Stucki, L. Rampa, R. Müri, U. P. Mosimann, T. Nef, Cognitive impairment categorized in community-dwelling older adults with and without dementia using in-home sensors that recognise activities of daily living, Scientific reports 7 (2017) 42084.

[20] S. Teipel, A. König, J. Hoey, J. Kaye, F. Krüger, J. M. Robillard, T. Kirste, C. Babiloni, Use of nonintrusive sensor-based information

and communication technology for real-world evidence for clinical trials in dementia, Alzheimer's & Dementia 14 (2018) 1216–1231.

[21] S. Mani, Y. Chen, T. Elasy, W. Clayton, J. Denny, Type 2 diabetes risk forecasting from emr data using machine learning, in: AMIA annual symposium proceedings, volume 2012, American Medical Informatics Association, 2012, p. 606.

[22] P. C. Austin, J. V. Tu, J. E. Ho, D. Levy, D. S. Lee, Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes, Journal of clinical epidemiology 66 (2013) 398–407.

[23] R. Kop, M. Hoogendoorn, A. Ten Teije, F. L. Büchner, P. Slottje, L. M. Moons, M. E. Numans, Predictive modeling of colorectal cancer using a dedicated pre-processing pipeline on routine electronic medical records, Computers in biology and medicine 76 (2016) 30–38.

[24] A. Rumshisky, M. Ghassemi, T. Naumann, P. Szolovits, V. Castro, T. McCoy, R. Perlis, Predicting early psychiatric readmission with natural language processing of narrative discharge summaries, Translational psychiatry 6 (2016) e921.

[25] J. Bullard, R. Murde, Q. Yu, C. O. Alm, R. Proano, Inference from structured and unstructured electronic medical data for dementia detection, in: INFORMS Computing Society Conference, 2015, pp. 236–244.

[26] K. Y. Yordanova, B. Demiray, M. R. Mehl, M. Martin, Automatic detection of everyday social behaviours and environments from verbatim

transcripts of daily conversations, in: 2019 IEEE International Conference on Pervasive Computing and Communications (PerCom, IEEE, 2019, pp. 1–10.

[27] P. Garrard, V. Rentoumi, B. Gesierich, B. Miller, M. L. Gorno-Tempini, Machine learning approaches to diagnosis and laterality effects in semantic dementia discourse, Cortex 55 (2014) 122–129.

[28] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Advances in neural information processing systems, 2013, pp. 3111–3119.

[29] Y. Goldberg, O. Levy, word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method, arXiv preprint arXiv:1402.3722 (2014).

[30] A. Subramanian, D. Pruthi, H. Jhamtani, T. Berg-Kirkpatrick, E. Hovy, Spine: Sparse interpretable neural embeddings, in: Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[31] C. Cortes, V. Vapnik, Support-vector networks, Machine learning 20 (1995) 273–297.

[32] L. Breiman, Random forests, Machine learning 45 (2001) 5–32.

[33] D. E. Rumelhart, G. E. Hinton, R. J. Williams, Learning representations by back-propagating errors, nature 323 (1986) 533.

[34] A.-L. Boulesteix, S. Janitza, J. Kruppa, I. R. König, Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 2 (2012) 493–507.

[35] R. J. Kate, R. M. Perez, D. Mazumdar, K. S. Pasupathy, V. Nilakantan, Prediction and detection models for acute kidney injury in hospitalized older adults, BMC medical informatics and decision making 16 (2016) 39.

[36] Y. Qi, J. Klein-Seetharaman, Z. Bar-Joseph, Random forest similarity for protein-protein interaction prediction from multiple sources, in: Biocomputing 2005, World Scientific, 2005, pp. 531–542.

[37] S. C. Miller, P. Gozalo, V. Mor, et al., Outcomes and Utlization for Hospice and Non-hospice Nursing Facility Decedents, Office of Disability, Aging and Long-Term Care Policy, US, 2000.

[38] H. Bonab, F. Can, Less is more: a comprehensive framework for the number of components of ensemble classifiers, IEEE Transactions on neural networks and learning systems (2019).

[39] P. Mather, B. Tso, Classification methods for remotely sensed data, CRC press, 2016.

[40] S. F. Weng, J. Reps, J. Kai, J. M. Garibaldi, N. Qureshi, Can machine-learning improve cardiovascular risk prediction using routine clinical data?, PloS one 12 (2017) e0174944.

[41] R. B. Parikh, C. Manz, C. Chivers, S. H. Regli, J. Braun, M. E. Draugelis, L. M. Schuchter, L. N. Shulman, A. S. Navathe, M. S. Patel, et al., Machine learning approaches to predict 6-month mortality among patients with cancer, JAMA network open 2 (2019) e1915997–e1915997.