SDM-DMMH 2016

5th Workshop on Data Mining for Medicine and Healthcare

May 7, 2016, Miami, FL

16th SIAM International Conference on Data Mining (SDM 2016)

Honorary Chair

Zoran Obradovic, Temple University

Workshop Chairs

Fei Wang, University of Connecticut

Gregor Stiglic, University of Maribor

Nitesh Chawla, University of Notre Dame

Overview

In virtually every country, the cost of healthcare is increasing more rapidly than the willingness and the ability to pay for it. At the same time, more and more data is being captured around healthcare processes in the form of Electronic Health Records (EHR), health insurance claims, medical imaging databases, disease registries, spontaneous reporting sites, and clinical trials. As a result, data mining has become critical to the healthcare world. On the one hand, EHR offers the data that gets data miners excited, however on the other hand, is accompanied with challenges such as 1) the unavailability of large sources of data to academic researchers, and 2) limited access to data-mining experts. Healthcare entities are reluctant to release their internal data to academic researchers and in most cases there is limited interaction between industry practitioners and academic researchers working on related problems.

The objectives of this workshop are:

1. Bring together researchers (from both academia and industry) as well as practitioners to present their latest problems and ideas.

2. Attract healthcare providers who have access to interesting sources of data and problems but lack the expertise in data mining to use the data effectively.

3. Enhance interactions between data mining, text mining and visual analytics communities working on problems from medicine and healthcare.

Program Committee

Mohamed Ghalwash, Temple University

Andreas Holzinger, Medical University Graz

Robert Moskovitch, Columbia University

Mykola Pechenizkiy, Eindhoven University of Technology

Niels Peek, University of Manchester

Igor Pernek, Research Studios Austria

Chandan K. Reddy, Wayne State University

Stein Olav Skrøvseth, University Hospital of North Norway

Cristina Soguero Ruiz, Rey Juan Carlos University

Suzanne Tamang, Stanford University

Ping Zhang, IBM T.J. Watson Research

Jiayu Zhou, Michigan State University

Workshop Schedule

May 7, Saturday	
8:30 – 8:40	Workshop Opening
8:40 – 9:30	Invited talk I (Mykola Pechenizkiy, Eindhoven University of Technology)
9:30 – 10:00	Coffee Break
10:00 – 12:00	*Paula Lauren, Guangzhi Qu and Feng Zhang* Discriminant Word Embeddings on Clinical Narratives *Flavio Bertini, Giacomo Bergami, Danilo Montesi and Paolo Pandolfi* Predicting frailty in elderly people using socio-clinical databases *Xiaoli Liu, Peng Cao, Dazhe Zhao and Arindam Banerjee* Multi-task Spare Group Lasso for Characterizing Alzheimer's Disease *Wei Ye, Bianca Wackersreuther, Christian Boehm, Michael Ewers and Claudia Plant* IDEA: Integrative Detection of Early-stage Alzheimer's disease **Giulia Toti, Ricardo Vilalta, Peggy Lindner and Daniel Price* Effect of the Definition of Non-Exposed Population in Risk Pattern Mining
12:00 – 13:30	Lunch Break (on your own)
13:30 – 14:20	Invited talk II (Mitsunori Ogihara, University of Miami)
14:20 – 15:00	**Stephanie L. Hyland, Theofanis Karaletsos and Gunnar Rätsch* Knowledge Transfer with Medical Language Embeddings **Arman Cohan, Luca Soldaini and Nazli Goharian* Identifying Significance of Discrepancies in Radiology Reports
15:00 – 15:30	Coffee Break
15:30 – 16:40	*Jialiang Jiang, Sharon Hewner and Varun Chandola* Exploiting Hierarchy in Disease Codes - A Healthcare Application of Tree Structured Sparsity-Inducing Norms *Milan Vukicevic, Sandro Radovanović, Gregor Stiglic, Boris Delibašić, Sven Van Poucke and Zoran Obradovic* A Data and Knowledge Driven Randomization Technique for Privacy-Preserving Data Enrichment in Hospital Readmission Prediction **Thomas Quisel, Luca Foschini and Alessio Signorini* Behavioral Phenotyping of Digital Health Tracker Data
16:40 – 16:50	Closing

*Short papers will have 15 minutes for presentation and 5 min for questions (long papers 20 + 5).

Table of Contents

Exploiting Hierarchy in Disease Codes - A Healthcare Application of Tree Structured Sparsity-Inducing Norms

Jialiang Jiang* Sharon Hewner* Varun Chandola*

Abstract

Hospital readmissions have become one of the key measures of healthcare quality. Preventable readmissions have been identified as one of the primary targets for reducing costs and improving healthcare delivery. However, most data driven studies for understanding readmissions have produced black box classification and predictive models with moderate performance, which precludes them from being used effectively within the decision support systems in the hospitals. In this paper we present an application of structured sparsity-inducing norms for predicting readmission risk for patients based on their disease history and demographics. Most existing studies have focused on hospital utilization, test results, etc., to assign a readmission label to each episode of hospitalization. However, we focus on assigning a readmission risk label to a patient based on her disease history. Our emphasis is on interpretability of the learnt models. To achieve this, we exploit the domain induced hierarchical structure available for the disease codes which are the features for the classification algorithm. We use a tree based sparsity-inducing regularization strategy that explicitly uses the domain hierarchy. The resulting model not only outperforms standard regularization procedures but is also highly sparse and interpretable. We analyze the model and identify several significant factors that have an effect on readmission risk. Some of these factors conform to existing beliefs, e.g., impact of surgical complications and infections during hospital stay. Other factors, such as the impact of mental disorder and substance abuse on readmission, provide empirical evidence for several pre-existing but unverified hypotheses. The analysis also reveals previously undiscovered connections such as the influence of socioeconomic factors like lack of housing and malnutrition. The findings of this study will be instrumental in designing the next generation decision support systems for preventing readmissions.

1 Introduction

Hospital readmissions are prevalent in the healthcare system and contribute significantly to avoidable costs. In United States, recent studies have shown that the 30-day readmission rate among the Medicare[1] is over 17%, with close to 75% of these being avoidable [1], with an estimated cost of $15 Billion in Medicare spending. Similar alarming statistics are reported for other private and public insurance systems in US and other parts of the world. In fact, management of care transitions to avoid readmissions has become a priority for many acute care facilities as readmission rates are increasingly being used as a measure of quality [5].

Given that the rate of avoidable readmission has now become a key measure of the quality of care provided in a hospital, there have been increasingly large number of studies that use healthcare data for understanding readmissions. Most existing studies have focused on building models for predicting readmissions using a variety of available data, including patient demographic and social characteristics, hospital utilization, medications, procedures, existing conditions, and lab tests [8, 4, 7]. Other methods use less detailed information such as insurance claim records [19, 9]. Many of these methods use machine learning methods, primarily Logistic Regression, to build classifiers and report consistent performance. In fact, most papers about readmission prediction report AUC scores in the range of 0.65-0.75.

While the predictive models show promise, their moderate performance means that they are still not at a stage where hospitals could use them as "black-box" decision support tools. However, given that the focus of these predictive models has been on performance, the models themselves are not easily interpretable to provide actionable insights to the decision makers. In this paper, we present a methodology to infer such insights from healthcare data in the context of readmissions. We build a logistic regression based classifier to predict if a

*{jjiang6,hewner,chandola}@buffalo.edu, State University of New York at Buffalo

[1] A federally funded insurace program representing 47.2 % ($182.7 billion) of total aggregate inpatient hospital costs in the United States[16].

patient is likely to be readmitted based on their disease history available from insurance records. We use sparsity inducing regularizers in our predictive model to promote interpretability. In particular, we show that by exploiting the hierarchical relationship between disease codes using the *tree-structured hierarchical group regularization* [20], we are able to learn a predictive model that outperforms all other types of sparsity inducing norms. Moreover, the tree-structured norm allows us to incorporate the rich semantic information present in the disease code taxonomy into the model learning, yielding highly interpretable models.

By analyzing the model trained on claims data obtained from the New York State Medicaid Warehouse (MDW), we infer several important insights to improve the understanding of readmissions. Some of our findings conform to existing beliefs, for example, the importance of bacterial infections during hospital stay. Other findings provide empirical evidence to support existing hypotheses amongst healthcare practitioners, for example, the effect of the type of insurance on readmissions [10]. Most interesting findings from our study reveal surprising connections between a patient's non-disease background and the risk of readmission. These include behavioral patterns (mental disorders, substance abuse) and socio-economic background.

We believe that such findings can have a significant impact on how healthcare providers develop effective strategies to reduce readmissions. At present, the healthcare efforts in this context have been two fold. First is the effort to improve the quality of care within the hospital and the second is to develop effective post-discharge strategies such as telephone outreach, community-based interventions, etc. The results from this study inform the domain experts on both fronts.

Organization The rest of the paper is organized as follows. We review existing literature on readmission prediction in Section 2. We describe the data used for our experiments in Section 3 and formulate the machine learning problem in Section 4. We discuss the classification methodology in Section 5. The results are presented in Section 6. The analysis of the resulting model in the context of readmissions is provided in Section 7. We present analysis of experiments done on subpopulations corresponding to major chronic diseases in Section 8.

2 Related Work

Coincident with the rising importance of readmissions in reducing healthcare costs, there have been several papers that use clinical and insurance claims information to build predictive models for readmissions. We refer the readers to a recent survey on the topic [12] for a comprehensive review. Most of these models use machine learning models such as Logistic Regression [8, 4, 7, 15] and Support Vector Machines [19]. Futoma et. al [8] provide a comparison of several machine learning methods including Logistic Regression, SVM, Random Forests for readmission prediction using features such as the diagnosis codes, procedure codes, demographics, patient's hospitalization history. However, the focus of most of these papers has been on improving accuracy of the classifier and not on interpreting the models to improve the understanding of the readmission problem. Moreover, many of these studies have either focused on a specific patient cohort or patient data from one or few hospitals [19]. For example, there have been several studies that focus on patients with acute heart conditions [2]. Recently, healthcare community has begun to study the impact of behavioral and socioeconomic factors on readmissions [11]. However, none of the data driven predictive models exploit this aspect, primarily because such data is challenging to obtain. However, in this paper we show that the diagnosis codes in the claims data contains valuable non-disease information about the patient which can be leveraged to better understand the readmission issue. Finally, the hierarchical relationship has never been exploited for building predictive models for readmission. Singh, et. al, [17] have presented a similar approach in the context of predicting disease progression, however, the authors focus on using the disease hierarchy to come up new features that are fed into the classifier.

3 Data

The data is obtained from the New York State Medicaid Warehouse (MDW). Medicaid is a social health care program for families and individuals with low income and limited resources. We analyzed four years (2009–2012) of claims data from MDW. The claims correspond to multiple types of health utilizations including hospitalizations, outpatient visits, etc. While the raw data consisted of 4,073,189 claims for 352,716 patients, we only included the patients in the age range 18–65 with no obstetrics related hospitalizations. The number of patients with at least one hospitalization who satisfied these conditions were 11,774 and had 34,949 claims.

For each patient we have two types of information. First type of information includes demographic attributes (age, gender) and the type of insurance. The second type of information is a patient history extracted from four years of claims data represented as a binary vector that indicates if the patient was diagnosed with a certain disease in the last four years.

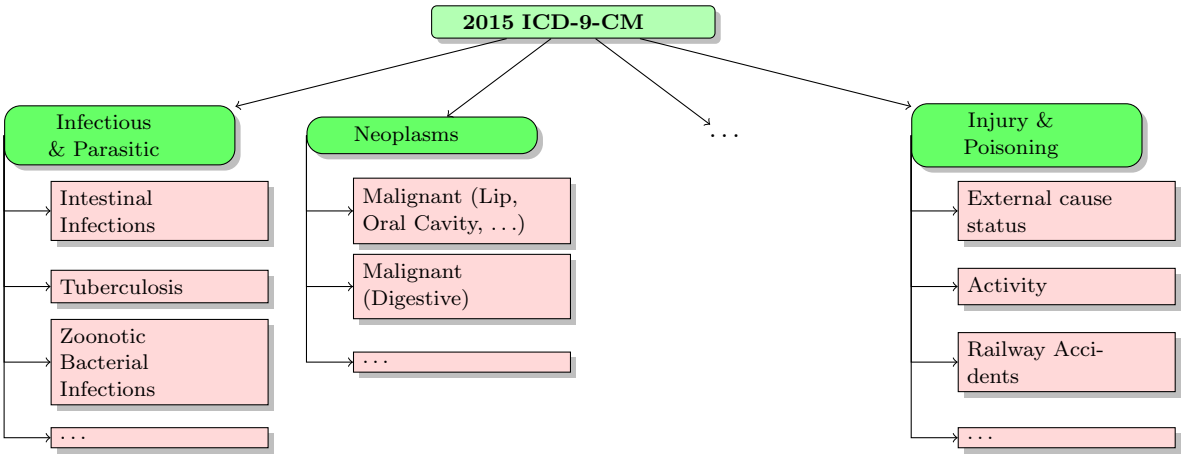


Figure 1: A sample portion of ICD9-CM classification. See `http://www.icd9data.com/2015/Volume1/default.htm` for complete hierarchy.

3.1 Insurance Plan Information Patients covered through Medicaid insurance can enroll into one of the two plan options. First option is to enroll in a *Managed Care Organization* (MCO). The MCO takes care of the delivery of the medical care to the patient and gets paid per person. The other option is called *Fee-for-service* (FFS) in which the healthcare provider gets paid for each service performed for the patient. While there has been a gradual transition from FFS to MCO style of insurance, the quality of care and costs under each plan has always been an important issue. In the context of readmissions it is important to understand how the readmission rate is impacted by the type of plan.

3.2 Diagnosis Codes Disease information is encoded in insurance claims using *diagnosis codes*. The *International Classification of Diseases* (ICD) is an international standard for classification of disease codes. The data used in this paper followed the ICD-9-CM classification which is a US adaptation of the ICD-9 classification. Conceptually, the ICD-9-CM codes are structured as a tree (See Figure 1 for a sample) with 19 broad disease categories at level 1. The entire tree has 5 levels and has total of 14,567 diagnosis codes. While the primary purpose of ICD taxonomy has been to support the insurance billing process, it contains a wealth of domain knowledge about the difference diseases.

3.3 Readmission Risk Flag For each patient in the above described cohort, we assign a binary flag for readmission risk. The readmission risk flag is set to 1 if the patient had *at least* one pair of consecutive hospitalizations within 30 days of each other in a single calendar year, otherwise it is set to 0.

4 Problem Statement

Given a patient's demographic information and disease history, we are interested in predicting the *readmission risk* (binary flag) for the patient. The problem formulation is different from many existing studies [8], where the focus is on assigning a readmission risk to a single hospitalization event. Our focus is on understanding the impact of socio-economic and behavioral factors on a readmission.

We denote each patient i as a vector $\mathbf{x}_i$ consisting of 11,884 elements corresponding to 11,881 disease codes and three elements for age, gender, and plan type. Note that while ICD-9-CM classification contains 14,567 codes, only 11,881 codes are observed in the data set used in this paper. All elements in the vector are binary except for age which takes 10 possible values corresponding to 10 equal width partitions between 18 and 65. The readmission risk flag is denoted using $y_i \in \{0, 1\}$ where 1 indicates readmission risk.

From machine learning perspective our task is to learn a classifier from a training data set $\langle \mathbf{x}_i, y_i \rangle_{i=1}^{N}$ which can be used to assign the readmission risk flag to a new patient represented as $\mathbf{x}_*$. Note that the input vector $\mathbf{x}_i$ is highly sparse for this data with nearly 36 non-zeros out of total 11,884 elements on an average.

5 Methodology

We use a *logistic regression* (LR) model [6] as the classifier, which, is the most widely used model in the context of readmission prediction [8]. The LR model, for binary classification tasks, computes the probability of the target y to be 1 (readmission risk), given the input

variables, $\mathbf{x}$ as:

$$p(y=1|\mathbf{x}) = \frac{1}{1+\exp(-\boldsymbol{\beta}^\top \mathbf{x})} \tag{5.1}$$

Where $\boldsymbol{\beta}$ is the LR model parameter (regression coefficients). We assume that $\mathbf{x}$ includes a constant term corresponding to the intercept. Thus $\boldsymbol{\beta}$ has the dimensionality $D+1$ where $D = 11,884$.

The model parameter $\boldsymbol{\beta}$ are learnt from a training data set $(\langle \mathbf{x}_i, y_i \rangle_{i=1}^N)$ by optimizing the following objective function:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{N} \log(1 + exp(-y_i \boldsymbol{\beta}^\top \mathbf{x}_i)) + \lambda \Omega(\boldsymbol{\beta}) \tag{5.2}$$

where the first term refers to the training loss and the second terms is a regularization penalty imposed on the solution; λ being the regularization parameter. Different forms of regularization penalties have been used in the past, including the widely used l_1 and l_2 norms [18]. While l_2 norm $(\Omega(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_2 = (\sum_j \beta_j^2)^{1/2})$ is typically used to ensure stable results, l_1 norm $(\Omega(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1 = \sum_j |\beta_j|)$ is used to promote sparsity in the solution, i.e., most coefficients in $\boldsymbol{\beta}$ are 0.

However, l_1 regularizer does not explicitly promote structural sparsity. Given that the features used in predicting readmission risk have a well-defined structure imposed by the ICD-9 standards, we explore regularizers that leverage this structure for model learning:

Sparse Group Regularizer This regularizer (also refered to as Sparse Group LASSO or SGL) assumes that the input features can be arranged into K groups (non-overlapping or overlapping) [3]. The SGL regularizer is given by:

$$\Omega(\boldsymbol{\beta}) = \alpha \|\boldsymbol{\beta}\|_1 + (1-\alpha) \sum_{k=1}^{K} \|\boldsymbol{\beta}_{G_k}\|_2 \tag{5.3}$$

where $\boldsymbol{\beta}_{G_k}$ are the coefficients corresponding to the group G_k. The above penalty function favors solutions which selects only a few groups of features (group sparsity). For the task of readmission prediction, we divide the features corresponding to 11,881 diagnosis codes into 19 non-overlapping groups, based on the top level groupings in the ICD-9-CM classification (See Table 1). The demographic and insurance plan features are grouped into an additional group.

Tree Structured Group Regularizer This regularizer, also referred to as Tree Structured Group LASSO (TSGL), explicitly uses the hierarchical structure imposed on the features. The TSGL regularizer is given by:

1	Infectious And Parasitic Diseases
2	Neoplasms
3	Endocrine, Nutritional And Metabolic Diseases, And Immunity Disorders
4	Diseases Of The Blood And Blood-Forming Organs
5	Mental Disorders
6	Diseases Of The Nervous System And Sense Organs
7	Diseases Of The Circulatory System
8	Diseases Of The Respiratory System
9	Diseases Of The Digestive System
10	Diseases Of The Genitourinary System
11	Complications Of Pregnancy, Childbirth, And The Puerperium
12	Diseases Of The Skin And Subcutaneous Tissue
13	Diseases Of The Musculoskeletal System And Connective Tissue
14	Congenital Anomalies
15	Certain Conditions Originating In The Perinatal Period
16	Symptoms, Signs, And Ill-Defined Conditions
17	Injury And Poisoning
18	Supplementary Classification Of Factors Influencing Health Status And Contact With Health Services
19	Supplementary Classification Of External Causes Of Injury And Poisoning

Table 1: Top level disease groups in the ICD-9-CM classification

$$\Omega(\boldsymbol{\beta}) = \sum_{i=0}^{D} \sum_{j=1}^{N_i} \|\beta_{G_j^i}\|_1 \tag{5.4}$$

where G denotes the tree constructed using the hierarchy of the diagnosis codes. G_j^i denotes the j^{th} node in the tree at the i^{th} level. Thus G_1^0 denotes the root level, and so on. For readmission risk prediction, we consider a three internal level hierarchy with 1193 nodes at level 3, 183 nodes at level 2, and 20 nodes at level 1.

6 Results

In this section we present our findings by applying logistic regression classifier for the task of readmission prediction on the MDW data described earlier. The full data set consists of 11,774 patients with 4,580 patients with readmission risk flag as *true* and 7,194 patients with readmission risk flag as *false*. We first compare the performance of the different regularization strategies to the classification task using the *F-measure* (harmonic mean of precision and recall for the *readmission = yes* class) as our evaluation metric. We also report the area under the ROC-curve (AUC) for each classifier. For each strategy, we run 10 experiments with random 60-40 splits for training and test data, respectively. The optimal values for the regularization parameter for each strategy are chosen using cross-validation. We use

the MATLAB package, SLEP [14], for the structured regularization experiments.

6.1 Comparing Different Regularizations In the past, researchers have shown that logistic regression typically outperforms other methods for predicting readmissions. Here we compare the performance of different regularization methods discussed in Section 5. The results are summarized in Table 2 and Figure 2.

Regularization	F1 Measure		AUC
	Mean	Std.	
l2	0.5364	0.0044	0.6889
l1	0.5343	0.0092	0.7152
SGL	0.5997	0.0027	0.7236
TSGL	**0.6487**	**0.0027**	**0.7185**

Table 2: Comparison of Different Regularization Strategies

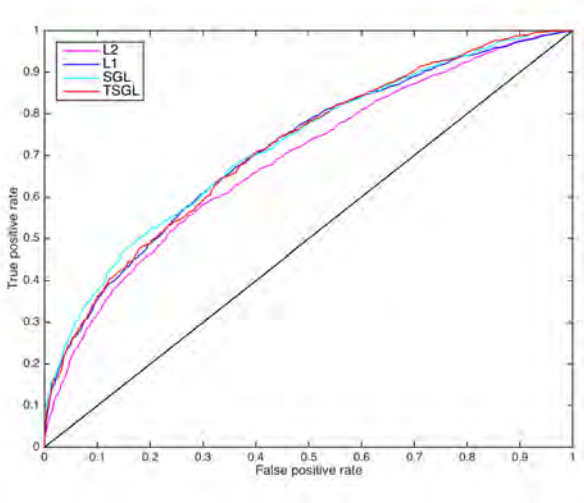


Figure 2: ROC for Different Regularization Strategies

The results indicate that leveraging the structure in the features (disease codes) results in an improvement in the performance of the logistic regression classifier. However, based on the F1-scores, it is not clear which of the structured regularization methods (SGL or TSGL) is better.

6.2 Interpretability of models Table 2 shows a clear evidence that leveraging the structure in the disease codes allows for a better classifier to predict readmission risk. We now focus on the interpretability of the resulting model under the different regularization mechanisms. The weights for the logistic regression model learnt under the four different types of regularizers is shown in Figure 3. We first note that l2 regularizer, for obvious reasons, does not produce a sparse solution (45% zero weights), while the other three regularizers induce significant sparsity (l1 - 92%, SGL - 97%, and TSGL - 94%). However, the structured regularizers are able to achieve structured sparsity which is consistent with the ICD-9-CM hierarchy.

While SGL achieves higher sparsity, the TSGL solution "aligns" better with the ICD-9-CM hierarchy. To verify this, we measure the sparsity of the coefficients at different levels of the hierarchy, as shown in Table 3. We observe that the TSGL provides better sparsity at higher levels of the hierarchy.

Level	SGL	TSGL
0	11537	11181
1	997	1021
2	110	119
3	4	6

Table 3: Number of nodes with all zero coefficients at different levels of the ICD-9-CM hierarchy. Level 0 corresponds to the leafs and level 3 corresponds to the coarsest level.

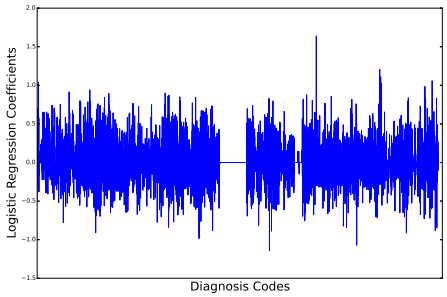
7 Discussions

Section 6 shows that leveraging the hierarchical information in the ICD-9-CM classification improves the predictive capability of logistic regression while promoting the structural sparsity to allow for better interpretability. In this section, we study the learnt model from the healthcare perspective. The focus is to understand the factors that impact readmission risk using the coefficients of the model learnt using the TSGL constraint.

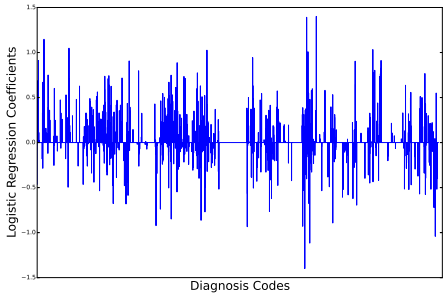
As shown in Figure 3d, the model is well-informed by the ICD-9-CM hierarchy. In all, there are only 437 non-zero coefficients. We choose the top 40 diagnosis codes with highest absolute coefficients. The most important diagnosis codes are listed in Table 4. The codes are grouped by their higher order functions.

The top codes listed in Table 4 reveal several valuable insights into the issue of readmission. For instance, while the role of age is understandable (older patients tend to get readmitted more), the impact of the insurance plan supports existing belief that disease management and population health activities provided by MCO would be associated with readmissions [10].

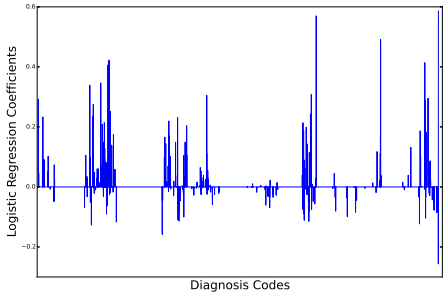
The next group of diagnosis codes pertain to infections and complications that happen during the hospital stay of the patients. This information is especially useful for hospitals to identify the key improvement av-



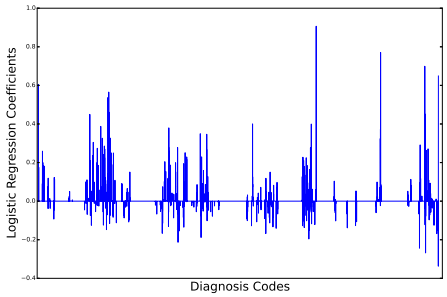
(a) l2



(b) l1



(c) SGL



(d) TSGL

Figure 3: Distribution of logistic regression coefficient values with different regularizers. The diagnosis codes (features) are arranged in the same order as the ICD-9-CM classification.

enues in the hospital operations to reduce the readmission rates. For example, it is well known that post-operative infections result in the patients returning to the hospital shortly after getting discharged. We are able to identify the same issue through this analysis.

While the role of chronic diseases in readmissions is well-understood in the literature, our results indicate that the chronic diseases play less of role, compared to other diagnosis codes.

However, the most important findings of our study correspond to last three groups of diagnosis codes in Table 4. Diagnosis codes related to mental health related issues were some of the most important positive factors in predicting readmissions. This is a valuable insight for hospitals in creating post-discharge strategies for such patients. This includes post discharge counseling and home visits.

Another key factor in readmission is substance abuse. In fact, the role of mental health and substance abuse in readmissions has long been studied in the healthcare community [13]. Many of the substance abuse issues might not be the direct cause of readmissions but might be indirect indicators of the lack of social and family support to assist patient recovery after the hospital discharge.

A direct evidence of the socio-economic effect on readmissions is given by the last set of diagnosis codes which includes lack of housing and malnutrition. These factors had very high positive weights indicating their importance. However, in most existing studies such factors were not considered, primarily because of the lack of relevant data. Through this study we discovery that insurance claims data itself contains elements that can be used to assess the socio-economic background of a patient.

8 Analyzing Sub-populations with Chronic Diseases

To further understand the role of various diagnosis codes in predicting readmissions in the context of major chronic diseases, we studied subpopulations of patients suffering from one of 9 diseases listed in Table 5. Most of these diseases have unique symptoms and treatments and many hospitals and other medical facilities are specialized in one of these diseases. Hence, it is important to understand if the important factors for readmission prediction are different from the entire population. Table 6 the average performance of logistic regression classifier (using cross-validation) on subpopulations corresponding each disease code. The subpopulations are created by considering only those patients who have had at least one admission for a given chronic disease. The results match the values obtained for the entire popula-

Age
Insurance Plan
Complications of surgery and medical care
Other unknown and unspecified cause of morbidity and mortality
Hematoma complicating a procedure
Disruption of external operation (surgical) wound
Other postoperative infection
Renal dialysis status
Aortocoronary bypass status
Care involving other specified rehabilitation procedure
Encounter for antineoplastic chemotherapy
Intestinal infection due to Clostridium difficile
Chronic Diseases
Anemia of other chronic disease
Systolic heart failure, unspecified
Chronic diastolic heart failure
Hypotension, unspecified
Paralytic ileus
Acute kidney failure
Pressure ulcer, lower back
Other ascites
Bacteremia
Hypopotassemia
Mental Disorders
Paranoid type schizophrenia, chronic with acute exacerbation
Unspecified schizophrenia, chronic
Major depressive affective disorder
Bipolar I disorder
Other personality disorders
Suicidal ideation
Substance Abuse
Alcohol withdrawal
Acute alcoholic intoxication in alcoholism, unspecified
Other and unspecified alcohol dependence, unspecified
Opioid type dependence, unspecified
Combinations of drug dependence excluding opioid type drug, unspecified
Sedative, hypnotic or anxiolytic abuse, continuous
Long-term (current) use of aspirin
Socio-economic Factors
Lack of housing
Unspecified protein-calorie malnutrition

Table 4: Most important diagnosis codes for predicting readmission risk

	Number of patients	
Major Disease Group	**Readmission**	**Total**
Coronary Artery Disease, CAD	300	658
Heart Failure, HF	736	1692
Diabetes, DM	1267	3321
Chronic Kidney Disease, CKD	528	1327
Asthma, ASTH	1036	2832
Hyperlipidemia-lipid disorder, LD	1448	4190
Hypertension, HTN	2405	6407
Chronic obstructive lung Disease, COPD	1075	2798
Depression, DEP	2268	5598

Table 5: Details of subpopulations suffering from a major chronic disease

tion.

Major disease group	F1 Measure	Std.
CAD	0.6588	0.0288
HF	0.6457	0.0250
DM	0.6200	0.0183
CKD	0.6239	0.0209
ASTH	0.5815	0.0110
LD	0.5631	0.0091
HTN	0.5966	0.0057
COPD	0.6031	0.0183
DEP	0.6098	0.0050

Table 6: Performance of TSGL based logistic regression classifier on chronic disease subpopulations

Next, we study the diagnosis codes which are most important in predicting readmission risk. Generally, we observe that while a few codes are common to all disease types, there are certain disease codes which are exclusively unique to each of the subtype. The diagnosis codes that have a strong impact on readmissions, independent of the type of chronic disease the patient suffers from, include postoperative infection and suicidal ideation. On the other hand, for every chronic disease, there are certain diagnosis codes that are unique, i.e., they are not a factor in any other chronic disease. Such codes can be significant for organizations that specialize in such diseases. Some of these codes are indicative of other chronic diseases or comorbidities. For example, for Diabetes patients, the presence of *transient cerebral ischemia* is an important predictor of readmission risk. Similarly, the presence of *pancreatic disease* codes in a COPD patient is an important risk factor. Interestingly, the unique diagnosis codes identified for each of the major chronic diseases primarily contain other chronic disease related codes, thereby indicating a strong impact

Entire population
Age
Insurance Plan
Suicidal Ideation
Other Postoperative Infection
Diabetes
Unspecified transient cerebral ischemia
Abnormality of gait
Schizophrenic disorders, residual type, chronic
Methicillin susceptible Staphylococcus aureus septicemia
Hematemesis
Chronic hepatitis C without mention of hepatic coma
Hyperlipidemia
Persistent vomiting
Diverticulitis of colon (without mention of hemorrhage)
Screening for malignant neoplasms of cervix
Hypertension
Candidiasis of mouth
Anemia of other chronic disease
Injury of face and neck
Coronary Artery Disease
Chronic obstructive asthma, unspecified
Asthma,unspecified type, unspecified
Diabetes with peripheral circulatory disorders
Urinary tract infection, site not specified
Epistaxis
Chronic kidney disease, Stage III (moderate)
Cerebral embolism with cerebral infarction
Chronic Kidney Disease
Hydronephrosis
Diabetes mellitus
Orthostatic hypotension
Other diseases of lung
Ventricular fibrillation
Acute respiratory failure
Chronic Obstructive Lung Disease
Secondary malignant neoplasm of bone and bone marrow
Thrombocytopenia, unspecified
Unspecified acquired hypothyroidism
Acute pancreatitis

Table 7: Most important diagnosis codes unique for few selected chronic disease specific subpopulations. Codes that are related to the same disease are ignored.

of comorbidity in readmissions.

9 Conclusions

In the last decade, there have been numerous studies that link factors pertaining to a patient's hospital stay to the risk of readmission. However most studies have been on a focused cohort, limited to one or few hospitals. However, we show here that similar results can be achieved using claims data, which has fewer elements but provides a large population coverage; the entire state of New York for this study. Even with the large volume of data, the predictive algorithms are not accurate enough (~ 0.60 F1-score) to be used as decision making tools. However, model interpretation can reveal insights which can inform the strategies for reducing and/or eliminating readmissions.

A patient's disease history is typically expressed using diagnosis codes, which can take as many as 18000 possible values, with many more possibilities in the next generation ICD-10 disease classification. With so many possible features, ensuring model interpretability is a challenge. However, using structured sparsity inducing models, such as the tree sparse group LASSO, used in this paper, one can ensure that the truly important factors can be identified. In this case study, we discover several such interesting factors.

In particular, we conclude that while in-hospital events such as infections are important, behavioral factors such as mental disorders and substance abuse and socio-economic factors, such as lack of housing or malnutrition at home are equally important. Targeted strategies, such as phone calls and home visits, will need to be developed to handle such situations. In Section 8, we analyze subpopulations specific to chronic diseases and show that similar methodology can reveral disease specific factors for readmissions.

References

[1] Promoting greater efficiency in medicare. MPA Committee Report to Congress, 2007.

[2] R. Amarasingham, B. J. Moore, Y. P. Tabak, M. H. Drazner, C. A. Clark, S. Zhang, W. G. Reed, T. S. Swanson, Y. Ma, and E. A. Halm. An automated model to identify heart failure patients at risk for 30-day readmission or death using electronic medical record data. *Medical Care*, 48(11):981–989, 2010.

[3] F. R. Bach. Consistency of the group lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9:1179–1225, 2008.

[4] S. A. Choudhry, J. Li, D. Davis, C. Erdmann, R. Sikka, and B. Sutariya. A public-private partnership develops and externally validates a 30-day hospital readmission risk prediction model. *Online J Public Health Inform.*, 5(2), 2013.

[5] P. H. Conway and D. M. Berwick. Improving the rules for hospital participation in medicare and medicaid. *JAMA*, 306(20):2256–2257, 2011.

[6] D. R. Cox. The regression analysis of binary sequences. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 215–242, 1958.

[7] J. Donze, D. Aujesky, D. Williams, and J. L. Schnipper. Potentially avoidable 30-day hospital readmissions in medical patients: Derivation and validation of a prediction model. *JAMA Internal Medicine*, 173(8):632–638, 2013.

[8] J. Futoma, J. Morris, and J. Lucas. A comparison of models for predicting early hospital readmissions. *Journal of Biomedical Informatics*, 56:229 – 238, 2015.

[9] D. He, S. C. Mathews, A. N. Kalloo, and S. Hutfless. Mining high-dimensional administrative claims data to predict early hospital readmissions. *J Am Med Inform Assoc.*, 21(2):272–279, 2014.

[10] S. Hewner, J. Y. Seo, S. E. Gothard, and B. Johnson. Aligning population-based care management with chronic disease complexity. *Nursing Outlook*, 62(4), 2014.

[11] S. Hewner, Y.-W. B. Wu, and J. Castner. Comparative effectiveness of risk-stratified care management in reducing readmissions in medicaid adults with chronic disease. *Journal for healthcare quality : official publication of the National Association for Healthcare Quality*, March 2015.

[12] D. Kansagara, H. Englander, A. Salanitro, D. Kagen, C. Theobald, M. Freeman, and S. Kripalani. Risk prediction models for hospital readmission: a systematic review. *Journal of American Medical Association*, 306(15), 2011.

[13] M. Lindsey, W. Patterson, K. Ray, and P. Roohan. Potentially preventable hospital readmissions among medicaid recipients with mental health and/or substance abuse health conditions. *Statsitical Brief, New York State Department of Health*, 3, 2007.

[14] J. Liu, S. Ji, and J. Ye. *SLEP: Sparse Learning with Efficient Projections*. Arizona State University, 2009.

[15] Y. Niu. Regression Models for Readmission Prediction Using Electronic Medical Records. Master's thesis, Wayne State University, Detroit, Michigan, 2013.

[16] A. Pfuntner, L. M. Wier, and C. Steiner. Costs for hospital stays in the united states, 2011. *Healthcare Cost and Utilization Project (HCUP) Statistical Briefs*, 168, 2013.

[17] A. Singh, G. Nadkarni, J. Guttag, and E. Bottinger. Leveraging hierarchy in medical codes for predictive modeling. In *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, BCB '14, pages 96–103, 2014.

[18] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1994.

[19] S. Yu, A. v. Esbroeck, F. Farooq, G. Fung, V. Anand, and B. Krishnapuram. Predicting readmission risk with institution specific prediction models. In *Proceedings of the 2013 IEEE International Conference on Healthcare Informatics*, pages 415–420, 2013.

[20] P. Zhao, G. Rocha, and B. Yu. The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*, 37(6A):3468–3497, 12 2009.

A Data and Knowledge Driven Randomization Technique for Privacy-Preserving Data Enrichment in Hospital Readmission Prediction

Milan Vukicevic * Sandro Radovanovic * Gregor Stiglic † Boris Delibasic *
Sven Van Poucke ‡ Zoran Obradovic §

Abstract

In health care predictive analytics, limited data is often a major obstacle for developing highly accurate predictive models. The lack of data is related to various factors: minimal data available as in rare diseases, the cost of data collection, and privacy regulation related to patient data. In order to enable data enrichment within and between hospitals, while preserving privacy, we propose a system for data enrichment that adds a randomization component on top of existing anonymization techniques. In order to prevent information loss (inclusive loss of predictive accuracy of the algorithm) related to randomization, we propose a technique for data generation that exploits fused domain knowledge and available data-driven techniques as complementary information sources. Such fusion allows the generation of additional examples by controlled randomization and increased protection of privacy of personally sensitive information when data is shared between different sites. The initial evaluation was conducted on Electronic Health Records (EHRs), for a 30-day hospital readmission prediction based on pediatric hospital discharge data from 5 hospitals in California. It was demonstrated that besides ensuring privacy, this approach preserves (and in some cases even improves) predictive accuracy.

Keywords: virtual examples, electronic health records, hospital readmission

1 Introduction

Healthcare predictive analytics have a potential for high-impact applications for many stakeholders. Hospitals can benefit from healthcare predictive analytics by monitoring of quality indicators, planning of healthcare capacities, optimization of supply levels etc. Insurance companies can define adequate charging policies; medical doctors can optimize treatment using decision support in diagnostics while patients can receive better quality of care, assessment of real costs by different hospitals etc. Prediction of 30-day hospital re-admission takes a special place in predictive analytics research [22]. Timely identification of potential unplanned readmissions can have a high impact on the improvement of healthcare services for patients, by reducing the need for unnecessary interventions and hospital visits. In addition, hospital readmission is considered as a major indicator of quality of care for hospitals, with significant economic impact. It is reported that readmission rate was 19.6% within 30 days, 34.0% within 90 days, and 56.1% within one year following discharge. According to the Institute for Healthcare Improvement, of the 5 million U.S. hospital readmissions, approximately 76% can be prevented, generating the annual cost of about $25 billion. [21]

Many researchers addressed this problem by building predictive models on secondary healthcare data, but they often failed to develop highly accurate models because of the lack of data. Regulations and privacy concerns often hinder the exchange of healthcare data between hospitals or other healthcare providers [22, 24]. This problem can be solved by two major strategies: secure multi-party computation (SMC) [22, 14], and randomization [8]. In the case of SMC, the sites cooperate to build the global prediction model without sharing the data itself, and these techniques have already shown their usefulness in many application areas [22, 14].

On the other hand, these techniques cannot help in situations where the lack of data originates from long and expensive clinical trials [2] or in the case of data from rarely observed diseases [1]. In such situations a randomization based pre-processing could be applied, where some noise is added to the original data prior to predictive modeling. Still, randomization techniques often hamper the utility of the model [14].

One way to address these problems is the inclusion

*Faculty of Organizational Sciences, University of Belgrade, Belgrade, Serbia. {milan.vukicevic, sandro.radovanovic, boris.delibasic}@fon.bg.ac.rs

†Faculty of Health Sciences, University of Maribor, Maribor, Slovenia. gregor.stiglic@um.si

‡Department of Anesthesiology, Critical Care, Emergency Medicine and Pain Therapy, Ziekenhuis Oost-Limburg, Genk, Belgium. svanpoucke@gmail.com

§Center for Data Analytics and Biomedical Informatics, Temple University, Philadelphia, USA. zoran@ist.temple.edu

of additional training examples created from the current set of examples by utilizing specic knowledge about the task at hand (often called virtual examples VE [16]). Compared to simple randomization techniques incorporation of VE as training examples in machine learning not only preserves model accuracy (and data privacy) but often improves it [25]. This is explained because incorporation of prior knowledge may contain information on a domain not present in the available domain dataset [15, 18] and thus exploits advantages in domain knowledge (knowledge driven) and data driven knowledge as complementary information sources. In the area of healthcare predictive modeling, virtual examples are successfully used for sepsis analysis [17]. VEs are of crucial significance for early sepsis prediction since patients infected by this disease often die in the early stage, and thus, temporal data cannot be gathered. Recently proposed predictive models for addressing this problem are based on VE that use differential equations [2] or medical models [5] as prior knowledge sources. It can be concluded that VE can be useful, and sometimes are the only possible technique to compensate for the lack of data in predictive modeling.

In this paper, we propose a method for VE generation which uses labeled examples and domain knowledge in the form of the ICD-9-CM hierarchy of diseases. The proposed technique is based on perturbation techniques that preserve privacy, but also allow generation of unobserved comorbidities. We consider three perturbation techniques based on apriori probabilities (data driven) and ICD-9-CM hierarchy information (knowledge driven) in order to randomize examples in a controlled manner while preserving privacy and addressing the problem of potentially existing, but non-observed comorbidities in data at hand. Additionally, features that indicate patient identity (patient identification, hospital identification and year of admission) are excluded. The intuition for the inclusion of hierarchy of diseases is based on too specific diagnoses that medical experts can assign. In case of similar symptoms, medical experts can make mistakes and assign false diagnoses. However, such diagnoses often belong to the same group of diagnoses due to their similarity. Therefore, the inclusion of a hierarchy of diseases in order to emphasize diagnoses from the same group or to generate unseen comorbidities can be used as a privacy preserving technique.

2 State of the art

Virtual examples are very popular in areas where data are hard to obtain or where data interchange is impossible due to regulations. This kind of problems is defined as small sample data set problems. This means that sample size is smaller compared to the number of features which leads to a poor generalization of classifiers.

One of the first efforts in investigating the effects of virtual examples is presented in [19]. They compared two ways for incorporation of domain knowledge in learning algorithm. One approach was based on changing the learning algorithm and other by incorporating virtual examples. The first approach may perform better and faster, but it requires significant effort in changing the goal function, or optimization of the learning algorithm. An approach based on virtual examples have two major advantages. First, it improves accuracy of the learning algorithm and second, it can be readily implemented for any learning algorithm. However, the virtual examples based approach increases overall training time and specifically for support vector machines leads to situation where many virtual examples becomes support vectors hence decreasing classification speed.

Virtual examples can be divided into two categories. First, more popular category, is to generate virtual examples by extracting the nontrivial knowledge hidden in the problem being solved. The task of extracting nontrivial knowledge is being formulated as a probability density function estimation [16]. This approach improved performance of pattern recognition by, given a 3D view of an object, generating new images from other angles through mathematical transformations. Extracting prior knowledge is a highly challenging task which requires a lot of efforts. This approach assures rationality, however adaptability to other problems is very low. Our approach can be categorized as the extraction of nontrivial knowledge. However, we added formally specified domain knowledge in the form of ICD-9-CM ontology (hierarchy). This way information about similarity of features is included in the generation of virtual examples.

Another approach, called perturbation, is to generate virtual examples by adding the noise to the original examples. This approach often adds noise using uniform or normal distribution. An interesting perturbation examples are presented in [27] where training samples of the rare class are divided into p groups using the k nearest-neighbor algorithm, then generated virtual examples by averaging every two samples of each group, and leaving the labels unchanged. Compared to the first approach, adaptability of these methods is more evident, but rationality cannot be assured.

There are four problems the virtual generator needs to address: inputs, the strategy of the virtual examples generator, outputs and the number of virtual examples. Most of the virtual example approaches differ in strategy. To the best of our knowledge, there are several

strategies of virtual example generation.

The first strategy randomly picks a sample inside of the hypersphere of a real samples input point, where the hypersphere is defined by uniform or Gaussian distribution. Since the point is selected near an original data point, it is similar, but not the same as an original data point. Further, the output is selected by a weighted average of original data points in a hypersphere or using evolutionary approach. This approach emphasizes that utilizing virtual examples does improve the performance of the classifier. [3]

Another approach of functional virtual population generation [11] is developed for specific types of systems such as manufacturing systems. The process of virtual examples generation starts from one system attribute and generate a specified number of virtual examples in the neighborhood of selected attribute. To test a virtual population, neural network is used, where the real performance of manufacturing system is used as an output. Once accuracy reaches its peak, a different system attribute is selected and the process is repeated. When all system attributes are processed, an integrated virtual population is used for artificial testing of the manufacturing process. Experimentally this approach dramatically improved learning accuracy and scheduling in a manufacturing system. This system inspired our strategy of virtual example generation. Similarly as selecting one system attribute, we select one diagnosis from ICD-9-CM codes (the proposed system is explained in more details in the following section).

The need for virtual examples in a small sample studies is explained in [7]. They elaborate that a classical network cannot recognize a non-linear function with a small sample. Therefore, they utilized the information diffusion principle. This principle asserts that, when an incomplete data set is used to estimate a relationship between features, there must exist reasonable diffusion means to change observations into fuzzy sets to partly fill the gap. They proposed a random generator controlled by probability density function as a diffusion function. Derived patterns are controlled to match using BP networks.

Another paper demonstrationg the importance of virtual examples in small data sets [9]. Small sample size learning cannot achieve high performance with respect to the overall population independent of the learning algorithm employed. However, small sample has a certain distribution, and virtual examples can be derived from density function obtained from intervalized kernel density estimation. This approach is shown to improve the performance of the learning algorithm. However, this approach seemed less suitable with nominal attributes (which is the case in our problem).

Virtual examples are highly utilized in medical domain, especially for the problem of rare disease and microarray analysis, which are formulated as small sample problems. One paper which utilizes virtual examples in order to improve the performance of learning algorithm for cancer identification is [12]. Their procedure for the binary classification problems with small sample data sets consists of three steps. First, a gene selection algorithm, which selects genes based on t-statistic value is employed to reduce dimensionality and improve learning ability (which is expected in gene expression problems due to high noise in attributes). Then, by utilizing group discovery technique, they profile related characteristics of each discriminative gene within a dataset. This step primarily searches for sample grouping (clusters) based on the spatial relationship between each other. As such, outliers are presented as a separate group. It is expected that clusters have the same label. Further, random noise is added to real examples using mean and standard deviation for each cluster. Simulation on both synthetic and real world data sets have shown that performance improved dramatically compared to the original data set. This paper motivated us to use groupings of features, not samples. Since we can utilize domain knowledge in the form of ICD-9-CM ontology (hierarchy) we grouped features which are similar in terms of effects and charging.

Virtual examples are also utilized in scheduling. Because of limited information in early dynamic flexible manufacturing systems, scheduling knowledge is difficult or impossible to obtain. Therefore, virtual examples must be generated and utilized for simulation. In [13] mega-trend-diffusion technique is performed to develop virtual examples. The mega-diffusion method diffuses a set of data using a common diffusion function with the objective to determine possible dataset coverage on a group consideration basis. When the group is found (domain range) samples are randomly selected from group and value of the diffusion function is added. The idea of grouping of data is also implemented in our research, where we grouped data based on hierarchy (domain knowledge). Namely, in the process of selection of samples, we used samples which have a same or similar diagnosis. Our approach is explained in more details in the following section.

Smoothing aiming for better estimation of pdf is used in [26]. Their virtual examples generator estimates parameters of Normal distribution from data. Further, using a random number generator they produce a virtual example. Although, this seems obvious it is mathematically and empirically shown that virtual examples improve the performance of learning algorithms for small size data sets and imbalanced data sets. These re-

sults motivated us to develop domain knowledge-based virtual examples generator.

Metaheuristics have also been used for virtual examples generators. In [10] a genetic algorithm is used for virtual example generation specially designed for small data prediction problems. Their mathematical model optimized mean absolute percentage error of linear regression function with constraints. The acceptable value of each attribute was determined with lower and upper bound. Virtual examples were defined as units in genetic algorithms, which were optimized in each iteration. The output of virtual example is defined based on real world samples with defined upper and lower bound. If the output does not satisfy these conditions the process is repeated. This way virtual examples are generated with an optimization procedure which reduces the error of learning algorithm. However, since virtual examples use class information adding the noise is essential in order to prevent overfitting.

Medical records including rare diseases are one of the most challenging prediction tasks where virtual examples generator is needed in order to obtain acceptable performance of learning algorithm.

In [5] a population of virtual patients is generated by random initialization of some parameters and by random initialization of the states initial conditions. Further, a patient is tracked over time using ordinary differential equations and based on results it can be either in survival group or in non-survival group. This random initialization and random selection of states, both using pdf from real data, have shown promising results. Another paper [2] produced in silico or virtual patients for sepsis prediction. Virtual examples were created using dynamical equation, but each of the patients has a unique set of parameters and therefore unique response to the CLP induction of sepsis. This is especially important since sepsis is highly progressive disease and early prediction is a must. As in majority of virtual example generator papers parameters are randomly sampled from predefined intervals and and if the likelihood for sepsis is high enough over time then the virtual patient is accepted as valid. It has been shown that this approach in combination with domain knowledge improves performance of prediction compared to a data driven approach. Therefore, we find this motivating to include formally written domain knowledge in order to improve performance of learning algorithm. In same domain (sepsis prediction) there is another approach for virtual patients generation which shown promising results

In [23] a feasibility based programming method is used as a virtual example generator. Model optimize mean absolute prediction error. Inputs are chosen randomly while outputs are defined using a genetic algorithm and backup propagation neural networks based feasibility-based programming model, with a constraint on output (must be inside lower and upper bound). When a virtual example is created latent spectral features are extracted which simplify model (thus reducing model training time). It has been shown that this approach improves the performance of learning algorithm for shell vibration and acoustic spectral data of a laboratory-scale ball mill.

This paper extends the first method for generating virtual examples which utilizes structured domain knowledge in the form of ontology (hierarchy) [25]. The hypothesis in this work is that using higher level concepts for probability smoothing and selection of diagnoses (as a step of virtual example construction) would positively influence readmission prediction and that this approach would enable data sharing between hospitals.

3 Proposed System

In order to address problems discussed above we propose a system for data enrichment and sharing of information about EHRs between hospitals that adds an additional layer of privacy protection into existing predictive modeling systems (Figure 1).The process of privacy protection starts with traditional anonymization techniques, which map personal and hospital identity into encrypted form. Additionally, time and duration of hospital visits are presented in relative form (number of days from initial admission), while exact dates are removed. Even though these techniques can substantially reduce the risk of patient identification, the state of the art predictive techniques theoretically can still identify the person based on procedures, diagnoses, and other data that cannot be encrypted if they serve as a basis for collaborative building and evaluation of predictive models. In order to increase privacy protection and allow data sharing and building of the more accurate predictive models, we propose a data enrichment mechanism that is based on randomization. However, data enrichment based on simple probabilistic randomization most often reduces the predictive performance, because of additional noise that is added to data. In order to prevent data quality loss by randomization we introduce a mechanism for fusion of data randomization techniques with the domain knowledge sources (ontologies or rules), and thus, randomization of the original data in a controlled manner.

We consider three types of EHR randomization: *a priori*, *knowledge-based* and *hybrid*. For the purpose of clarity, this will be more thoroughly explained later in the text. After anonymization and randomization,

this additional example can be used for data enrichment within or between hospitals. Further, each hospital can build predictive models on enriched data (generated on its own or by other hospitals) and these models can be used for assessment of the risk of readmission for new patients. Finally, predictive models (classification, regression, etc.) can be built on enriched data sources and applied for many different problems in healthcare e.g. prediction of re-admission risk, a number of admissions in hospitals, cost-to-charge ratios, etc. In this research, we built and evaluated predictive models for readmission risk prediction. These models should serve as decision support for medical doctors when making a decision about diagnoses and/or therapy. High readmission risk can indicate that diagnosis or therapy are not adequate for the given patient and that doctor should re-examine the patient, or send him to additional testing in order to prevent potential readmission. The proposed system is depicted on Figure 1.

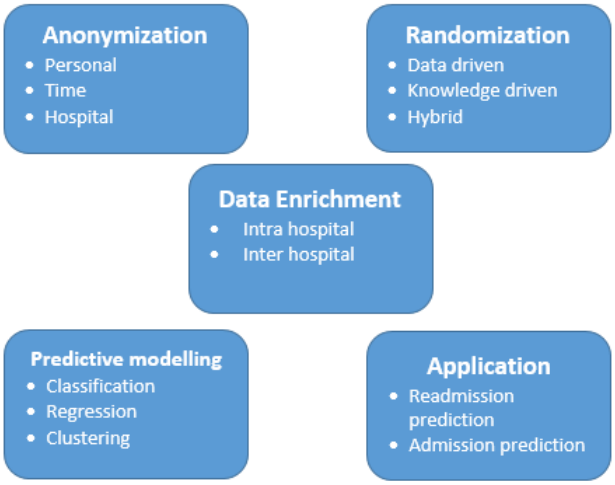


Figure 1: DSS for privacy-preserving sharing of data

In the further text, we explain in more detail the procedures for enabling data sharing through a priory (probability) based and knowledge guided randomization. These techniques are similar to one recently proposed [25], which was previously used for the generation of rare diseases and improved generalization of predictive algorithms. Here it will be used as a general knowledge-based randomization mechanism that allows more secure data sharing.

The additive (a priori) randomization approach uses a technique for smoothing the probabilities of every diagnosis, in a similar way as Laplace smoothing in the Naive Bayes algorithm. For each new VE, we start the generation process by selecting a diagnosis based on a priori probabilities of all diagnoses that are smoothed (increased) with parameter . The initial disease may be selected based on the highest probability of appearance (if most common disease from the hospitals should be shared), or inverted probabilities (if rare diseases should be shared). When the first feature (disease) is selected, the next disease (comorbidity) is selected in the following way: First, the comorbidity subset (CS) is formed with all diagnoses that have comorbidities with the previously selected diagnosis. Next, features are chosen based on λ-updated probabilities from CS. This procedure is iteratively repeated by forming CS based on conditional probabilities of comorbidities for already selected features. It is intuitively clear that this procedure will result in feature distribution that is similar to the original data. Namely, all new features will have the same or reduced set of features compared to the original dataset, where privacy will be preserved, but there is no chance of generating unseen comorbidities.

Knowledge-based randomization - enables generation of features (i.e. comorbidities) that are not observed in the original dataset. This generation can preserve privacy, but also, could be useful in situations when hospitals did not have patients with a specific set of diseases (and it is known that such a set can appear in the future). Of course, by using simple randomization such VE cannot be generated, and thus the process of randomization has to be guided by some form of domain knowledge.

In this study, we use hierarchical ICD-9 (excerpt of hierarchy is given in Figure 2) classification of diseases as a knowledge source. The ICD-9 codes are organized in a hierarchy where an edge represents an is-a relationship between a parent and its children. Hence, the codes become more specific as we go down the hierarchy [20]. When leveraging the ICD-9 hierarchy for generating virtual examples, we can assume that the child nodes have a correlated relationship with the feature of interest (selected feature). There are about 15,000 diagnostic codes in the ICD-9-CM hierarchy. Each three-digit diagnostic code is associated with a hierarchy tree. In this paper, we refer to it as a top-level diagnostic code. Figure 2 shows a part of hierarchy within the top-level (most general) diagnostic code that represents infectious and parasitic diseases. Top-level can be represented as a set of lower level concept group of diagnoses, which present more specific diagnoses. Further, that set of diagnoses can be specified to more specific concepts (five digit codes). Hierarchy used in this paper is Clinical Classification Software (CCS) which clusters patient diagnoses and procedures into clinically meaningful categories. [4]

When leveraging the ICD-9 hierarchy for generating virtual examples, we can assume that the child nodes

Algorithm 1 Pseudo-code for VE generator

Inputs: dataset **D**, # examples **n**, smoothing λ, continue parameter **cp**, number of examples **k**
Output: list of virtual examples **VE**

VE $= \emptyset$ *//initialize list of virtual examples*
while k virtual examples are created **do**
  set **CS** $=$ **D** *//create comorbidity subset CS*
  V $= \emptyset$ *//initialize virtual example*
  while cp is *true* **do**
    //calculate probabilities of diagnoses in CS
    //smooth probabilities of every ICD-9 code
    //smooth probabilities of similar diagnoses

$$\mathbf{P} = \frac{|X| + \lambda \times |X| + \lambda \times |X_{cs}|}{n_{cs}}$$

    if *first step* **then**
      //invert probabilities

$$\mathbf{P} = \frac{1}{\mathbf{P}}$$

    end if
    //add disease i to V
    //using roulette wheel selection
    Add(V, i)
    //select CS with examples having at least
    //one diagnosis from three level group of
    //selected diagnosis
    CS $= D_{cs}$
    //calculate ratio of examples in CS with
    //higher number of diagnoses and number of
    //examples with lower number of diagnosis

$$ratio = \frac{|CS_{>}| + \lambda \times |CS_{>}|}{|CS| + 2\lambda \times |CS_{>}|}$$

    if *random number* $\geq$ ratio **then**
      cp $=$ *false*
    end if
  end while
  //roulette wheel selection for other features
  //excluding hospital and date of admission

  //add virtual example to list
  Add(VE, V)
end while

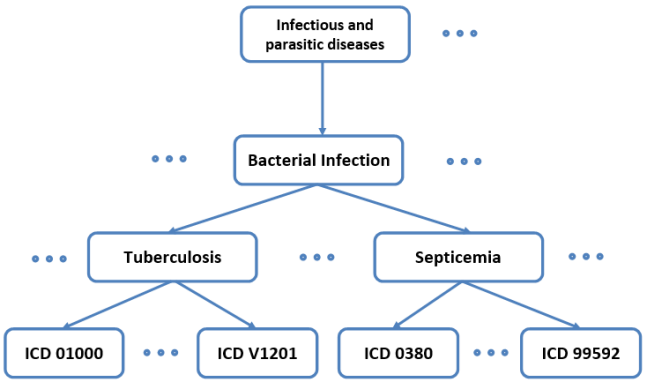


Figure 2: ICD-9 hierarchy of diseases

and a parent node are both correlated with the feature of interest (in this case, the risk of 30-day hospital readmission). So the main idea is to generate VEs with similar readmission outcome for diseases or comorbidities (combination of diseases) from the same hierarchy group.

The first step is the same as in Additive smoothing: the first diagnosis is chosen from rare diagnoses that are favored for selection. The main contribution is the iterative step, where CS is formed not only from comorbidities with previously selected diseases but all comorbidities of 3-digit hierarchy level that selected diagnoses to originate from. This extends the space of possible diagnosis (now not only comorbidities with one diagnosis are considered, but comorbidities with the hierarchy group) and allows knowledge-guided selection of unseen cases. Intuition behind this approach is that diagnoses from the same hierarchical group are often treated the same way and that on the low level of hierarchy diagnosis could be too specific, since various diagnoses from the same group at symptom level seem to share similar behavioral symptoms and diagnostic criteria [28], meaning that real diagnosis could be overlooked.

This way it is possible to adapt models for the unseen cases, but also to randomize them in a controlled manner and thus preserve privacy when sharing data.

Integrated randomization (Additive and ICD9 based) smoothing combines previously described approaches by executing Additive and ICD9 smoothing, respectively. After execution, feature probabilities are updated by the sum of aforementioned smoothing updates. Further CS is formed the same way as in the ICD9 smoothing. The level of randomization and ICD9 influence is controlled by smoothing parameters that control smoothing levels for each type of smoothing. Users also provide the number of examples to be gen-

Table 1: Accuracy of logistic regression (AUC) when using enriched data of a single hospital versus using an individual hospital data alone or shared data from all hospitals.

# Examples	# Readmitted	% Readmitted	Individual	Shared	Enriched
7884	1,336	16.95	0.695	**0.820**	0.815
6394	1,450	22.68	0.693	**0.793**	0.771
6317	1,064	16.84	0.644	**0.782**	0.762
5103	705	13.82	0.621	0.780	**0.794**
4405	813	18.46	0.636	0.728	**0.761**
7884	1,336	16.95	0.695	**0.825**	0.817
6394	1,450	22.68	0.693	0.802	**0.810**
6317	1,064	16.84	0.644	**0.791**	0.741

erated and a parameter for smoothing variables other than diagnoses. Pseudo-code is given in Algorithm 1.

4 Experimental Evaluation

In this research, we addressed the problem of hospital readmission prediction in situations where EHRs are not shared between hospitals. Our main hypothesis was that the controlled (knowledge guided) randomization of data can provide additional examples that can be shared in a more secure way and increase the performance of predictive models built by each hospital.

4.1 Data In Hospital discharge data from California, State Inpatient Databases (SID), Healthcare Cost and Utilization Project (HCUP), Agency for Healthcare Research and Quality was used [6]. This data tracks all hospital admissions at the individual level, having a maximum of 15 diagnoses for each admission. Since there are over 14,000 ICD-9-CM codes, and using diagnoses as 15 polynomial attributes would be unfeasible for any learning algorithm to handle, we transformed the feature space by presenting each code as a feature. Therefore, we have about 14,000 binary features, where positive value marks the presence of the diagnosis. The final data set was preprocessed as in [22], with 850 input features (diagnoses) and as predictors for single binary output (patient was re-admitted within 30 days or not).

4.2 Experimental setup Since Data from 2009 and 2010 (about 2/3 of the entire data set) were used for training, while data from 2011 was used for testing. As a learning algorithm, we used logistic regression (LR), since it often showed good performance in medical applications, also performing well on this type of data and most importantly, providing interpretable models. Interpretability of models is especially important in Healthcare predictive analytics because of high costs of wrong decisions. We used all pediatric patient data from 8 hospitals with the highest numbers of patients and highest numbers of different diseases and highest number of patients.

Hybrid strategy (both additive and knowledge based randomization) was used in order to generate additional examples in a controlled (knowledge guided) manner. For each hospital, the same number of randomized examples is created, leading to a repository of 30,103 examples that were used for data enrichment of each hospital. In order to show usefulness of enriching data from specific hospitals with virtual examples, we made the following sets of data (on which logistic regression is applied and evaluated):

- **Individual** LR was trained on data from a single hospital.to predict readmission at that hospital.
- **Shared** LR model is developed on integrated data from all hospitals.
- **Enriched** LR was trained on data from an individual hospital enriched with data from VE repository.

Since the data has a high class imbalance (about 20% of all patients were readmitted), we evaluated all models with Area Under Curve (AUC) instead of classification accuracy.

4.3 Results In contrast to medical applications where data sharing is not applied or not allowed, the proposed method can generate additional examples, which can allow developing more accurate and with better generalization power. Since there are a lot of hospitals with a relatively small number of admissions, at these hospitals this method can supplement missing examples. Table 1 shows brief data description and AUC values for each experiment on each hospital (larger values are better and the best performance is presented in bold letters).

It can be seen at Table 1 that sharing the data drastically improves model performance. All models that are built on Original data have AUC less than 0.696, while models on Shared data had AUC performance from 0.728, up to 0.825. Still, such sharing of data is often not possible due to strict data privacy regulations. On the other hand, models built using data from VE repository allow sharing the data without compromising privacy. It can be seen that models that are built on data from VE repository (and original data from each hospital) achieved results comparable to using shared data. Performance on all hospitals was very similar and for hospitals 4, 5 and 7 results were even slightly better.

5 Conclusion and Future Research

In this paper we proposed a method that allows privacy while preserving data sharing between hospitals. The system is based on domain knowledge guided randomization techniques, where domain knowledge is presented in the form of a hierarchy of diagnoses. It is shown that sharing the data through generated virtual examples as such improves model performance for hospital readmission prediction. We conclude that hospitals could reduce costs for readmitted patients by using data sharing and virtual examples.

In future work, we plan to extend the system to other types of domain knowledge sources, such as other hierarchies and ontologies, where additional information about relations between diseases is present.

6 Acknowledgment

This research was partially supported by the U.S. Office of Naval Research grant N00014-15-1-2729, and by SNSF Joint Research project (SCOPES), ID: IZ73Z0_152415.

References

[1] Bavisetty, S., Grody, W. W., & Yazdani, S. (2013). Emergence of pediatric rare diseases: Review of present policies and opportunities for improvement. *Rare Diseases*, *1*(1), e23579.

[2] Cao, X. H., Stojkovic, I., & Obradovic, Z. (2014, January). Predicting Sepsis Severity from Limited Temporal Observations. In *Discovery Science* (pp. 37-48). Springer International Publishing.

[3] Cho, S., Jang, M., & Chang, S. (1997). Virtual sample generation using a population of networks. *Neural Processing Letters*, *5*(2), 21-27.

[4] Elixhauser, A., Steiner, C., & Palmer, L. (2008). Clinical classifications software (CCS). *Book Clinical Classifications Software (CCS)*(Editor ed eds).

[5] Ghalwash, M., & Obradovic, Z. A Data-Driven Model for Optimizing Therapy Duration for Septic Patients. In *Proc. 14th SIAM Intl. Conf. Data Mining, 3rd Workshop on Data Mining for Medicine and Healthcare*, Philadelphia, PA, USA (April 2014).

[6] Healthcare Cost and Utilization Project. (2008). Introduction to the HCUP Kids inpatient database (KID) 2006. Rockville (MD): Agency for Healthcare Research and Quality.

[7] Huang, C., & Moraga, C. (2004). A diffusion-neural-network for learning from small samples. *International Journal of Approximate Reasoning*, *35*(2), 137-161.

[8] Kantarcioglu, M., Nix, R., & Vaidya, J. (2009). An efficient approximate protocol for privacy-preserving association rule mining. In *Advances in Knowledge Discovery and Data Mining* (pp. 515-524). Springer Berlin Heidelberg.

[9] Li, D. C., & Lin, Y. S. (2006). Using virtual sample generation to build up management knowledge in the early manufacturing stages. *European Journal of Operational Research*, *175*(1), 413-434.

[10] Li, D. C., & Wen, I. H. (2014). A genetic algorithm-based virtual sample generation technique to improve small data set learning. *Neurocomputing*,*143*, 222-230.

[11] Li, D. C., Chen, L. S., & Lin, Y. S. (2003). Using functional virtual population as assistance to learn scheduling knowledge in dynamic manufacturing environments. *International Journal of Production Research*, *41*(17), 4011-4024.

[12] Li, D. C., Fang, Y. H., Lai, Y. Y., & Hu, S. C. (2009). Utilization of virtual samples to facilitate cancer identification for DNA microarray data in the early stages of an investigation. *Information Sciences*, *179*(16), 2740-2753.

[13] Li, D. C., Wu, C. S., Tsai, T. I., & Lina, Y. S. (2007). Using mega-trend-diffusion and artificial samples in small data set learning for early flexible manufacturing system scheduling knowledge. *Computers & Operations Research*, *34*(4), 966-982.

[14] Mathew, G., & Obradovic, Z. (2013). Distributed Privacy-Preserving Decision Support System for Highly Imbalanced Clinical Data. *ACM Transactions on Management Information Systems (TMIS)*, *4*(3), 12.

[15] Mirchevska, V., Lutrek, M., & Gams, M. (2014). Combining domain knowledge and machine learning for robust fall detection. *Expert Systems*, *31*(2), 163-175.

[16] Niyogi, P., Girosi, F., & Poggio, T. (1998). Incorporating prior information in machine learning by creating virtual examples. *Proceedings of the IEEE*, *86*(11), 2196-2209.

[17] Radosavljevic, V., Ristovski, K., & Obradovic, Z. (2013). A data-driven acute inflammation therapy. *BMC Medical Genomics*, *6*(Suppl 3), S7.

[18] Radovanovic, S., Vukicevic, M., Kovacevic, A., Stiglic, G., & Obradovic, Z. (2015). Domain knowledge based hierarchical feature selection for 30-day hospital readmission prediction. In *Artificial Intelligence in Medicine* (pp. 96-100). Springer International Publishing.

[19] Schlkopf, S. P., Simard, P., Vapnik, V., & Smola, A. J. (1997). Improving the accuracy and speed of support vector machines. *Advances in neural information processing systems*, *9*, 375-381.

[20] Singh, A., Nadkarni, G., Guttag, J., & Bottinger, E. (2014, September). Leveraging hierarchy in medical codes for predictive modeling. In *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics* (pp. 96-103). ACM.

[21] Srivastava, R., & Keren, R. (2013). Pediatric readmissions as a hospital quality measure. *JAMA*, *309*(4), 396-398.

[22] Stiglic, G., Wang, F., Davey, A., & Obradovic, Z. (2014). Pediatric Readmission Classification Using Stacked Regularized Logistic Regression Models. In *AMIA Annual Symposium Proceedings* (Vol. 2014, p. 1072). American Medical Informatics Association.

[23] Tang, J., Jia, M., Liu, Z., Chai, T., & Yu, W. (2015, August). Modeling high dimensional frequency spectral data based on virtual sample generation technique. In *Information and Automation, 2015 IEEE International Conference* on (pp. 1090-1095). IEEE.

[24] Vest, J. R., & Gamm, L. D. (2010). Health information exchange: persistent challenges and new strategies. *Journal of the American Medical Informatics Association*, *17*(3), 288-294.

[25] Vukicevic, M., Radovanovic, S., Kovacevic, A., Stiglic, G., & Obradovic, Z. (2015). Improving Hospital Readmission Prediction Using Domain Knowledge Based Virtual Examples. In *Knowledge Management in Organizations* (pp. 695-706). Springer International Publishing.

[26] Yang, J., Yu, X., Xie, Z. Q., & Zhang, J. P. (2011). A novel virtual sample generation method based on Gaussian distribution. *Knowledge-Based Systems*, *24*(6), 740-748.

[27] Zhang L. & Chen G.H. (2006), Method for constructing training data set in intrusion detection system, *Computer Engineering and Applications*, *42*(28). 145 146.

[28] Zhou, J., Lu, Z., Sun, J., Yuan, L., Wang, F., & Ye, J. (2013, August). FeaFiner: biomarker identification from medical data through feature generalization and selection. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1034-1042). ACM.

Effect of the Definition of Non-Exposed Population in Risk Pattern Mining

Giulia Toti[1], Ricardo Vilalta[1], Peggy Lindner[2], and Daniel Price[2]

[1]Department of Computer Science, University of Houston
[2]Honors College, University of Houston

Abstract

Odds ratios, and other traditional metrics used to quantify the correlation between illness and risk factors, rely on the correct definition of exposed and non-exposed populations. This definition has always been straightforward in conventional epidemiological studies that focus on the effect of a single risk factor. Current data mining techniques, like association rule mining, allow the evaluation of the effect of combinations of multiple risk factors. In this new scenario, what would be, the optimal definition of non-exposed population?

So far in the literature, the non-exposed group included every subject who was not exposed to *all* of the risk factors under analysis. Alternatively, we may decide to include in the non-exposed group subjects who were not exposed to *any* of them. A study to determine which definition should be favored in differing circumstances is currently missing. In this paper, we discuss possible advantages and disadvantages in using one definition or the other. We also show the differences in results obtained when the two definitions are implemented in an association rule mining algorithm and used to extract rules from a group of datasets. We conclude that these differences should not be ignored and proper attention should be given to finding the correct definition of non-exposed population in risk assessment studies.

1 Introduction.

In recent years we have observed an exponential growth in the amount of data available in the medical field. This trend creates an opportunity for modern data mining techniques, which can be employed to extract meaningful and useful information from massive repositories [11]. Techniques that can extract and report the information in the form of rules are particularly favored, because they are readily interpretable for all health practitioners, even those who do not have a background in data analytics. The most popular algorithms for extraction of rules from data can be grouped into two families: Decision Trees and Association Rule Mining (ARM). In this paper, we will focus on ARM.

Association rule mining was originally designed to find frequent associations between items in large databases [1]. Since then, different formulation of ARM have been studied, and many have also been applied in clinical environments. One of the first applications was presented by Brossette *et al.* in 1998, to study the association between hospital infections and public health surveillance [4]. Other publications include studies on chronic hepatitis, septic shock, heart disease, association deficit disorder, cancer prevention, response to drugs and general lifestyle risk behaviors [6, 12, 17, 18, 19, 20, 21, 22]. In its early formulation, ARM was designed to find rules with high support and confidence, that is, groups of elements that appear frequently in the dataset and that are highly correlated. However, sometimes associations of interest in the medical domain can be infrequent and not particularly highly correlated. Therefore, it was necessary to introduce new metrics to evaluate the interestingness of a rule. A list of these metrics and an evaluation of their effectiveness is presented in [18].

The list of objective measures available to evaluate rules includes Risk Ratio (RR) and Odds Ratio (OR). These two measures are largely used in the field of medicine and public health to establish a correlation between one or more factors and the health outcome under study. The factors implicated in the health outcome may differ widely (from genetic, to demographic, to environmental) and are generally called *exposures*. By computing risk ratio and odds ratio, it is possible to compare the exposed and non exposed populations to determine if one of them has higher chances to develop the outcome under study.

In 2009, Li *et al.* [13] presented a variant of association rule mining that abandoned the traditional support-confidence framework in favor of a pattern search guided by risk ratio. The proposed method was more efficient in covering the search space, and produced a smaller number of rules. But the number of rules in the output could still be too large for easy interpretation. Later, another paper by Li *et al.* [14] presented a method to prune redundant rules based on overlapping of the confidence interval of the odds ratio. The odds ratio is usually reported with its confidence interval to show the accuracy of the estimate. Li *et al.* used confidence intervals to determine if a rule and its parent are statistically different. If the confidence intervals do not overlap, the rules must carry different

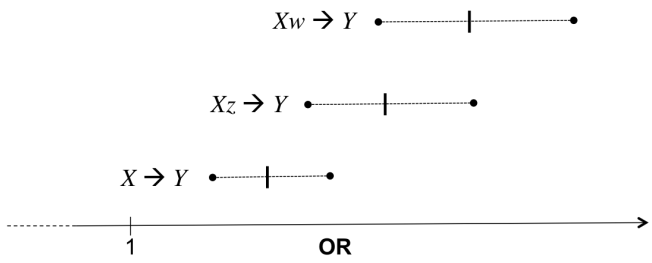


Figure 1: Schematic representation of OR confidence interval of different rules. Rule $X \to Y$ is the parent. By adding other exposures to the parent rule, we obtaine the new rules $Xz \to Y$ and $Xw \to Y$. Because only the confidence interval of $Xw \to Y$ does not overlap with the parent rule, only this new association is statistically different. $Xw \to Y$ brings new relevant information, while $Xz \to Y$ should be pruned.

information; otherwise, they are considered equivalent and the subrule is pruned. A schematic representation of this concept is visible in Figure 1.

The odds ratio measures the correlation between the exposure under study and a particular health outcome by comparing two groups of subjects, exposed and non-exposed (Eq. 2.2). Traditional epidemiological studies analyze one risk factor at a time. While other factors can be included in the data collection to control for confounders and interactions; the interest is usually limited to one new exposure that has not been studied before. In this way, defining exposed and non-exposed populations becomes straightforward. But in ARM, multiple risk factors are often combined to form a rule, therefore a new question arises: given a set of exposures, which subjects should be included in the non-exposed group?

So far in the literature [14, 18], researchers have chosen to define non-exposed subjects as all those subjects other than those exposed to *all* the factors included in the rule. This definition is also implemented in popular software for ARM such as the *arules* package in R [8, 9]. An alternative definition, which to our knowledge has not been discussed before, includes in the non-exposed population only those subjects who have not been exposed to *any* of the factors included in the rule. In this paper, we will discuss possible advantages and disadvantages of the two definitions and we will show their impact when the pruning criteria described in [14] is used.

2 Problem definition.

We will start this section with some of the foundations of rules and rule mining. A rule represents an association between two sets of items, i.e. X and Y. The notation $X \to Y$ indicates that when X occurs Y also occurs with a probability $P(Y|X)$. In association rule mining, rules are extracted from large binary databases. The columns of the database represent the set of possible items $I = \{i_1, i_2, ..., i_m\}$. A subset of items $X \subseteq I$ is called an *itemset*. The rows of the database represent all instances, or transactions, that occurred in the dataset. In a medical study, each row represents a different subject of study, columns are used to represent characteristics or conditions of the corresponding subject. Possible items could be $\{Male\}$ or $\{Age : 30 \div 40\}$ or $\{Smoker\}$. Medical data are mined to find significant rules such as $\{Smoker, Age : 30 \div 40\} \to \{Lung\ cancer\}$. Similar rules may have low confidence (not all smokers in their 30s have lung cancer!), but help finding a significant change in risk for exposed and non-exposed population.

Not every possible combination of itemsets (X, Y) forms an interesting rule. Different criteria have been defined to differentiate meaningful rules from the rest. The most common, introduced by Agrawal in its first ARM formulation [1], forms the support-confidence framework. The support of a rule represents how often the items of the rule occur together in the dataset ($supp(XY) = P(Y \wedge X)$). The confidence of a rule measures the chance of finding the itemset Y (also called *consequent* or *RHS, right hand side*) in a transaction, given the presence of the itemset X (also called *antecedent* or *LHS, left hand side*). Therefore, the confidence is simply a conditional probability:

$$conf(X \to Y) = P(Y|X) = \frac{supp(XY)}{supp(X)} \tag{2.1}$$

The support-confidence framework requires selected rules to have support and confidence larger than some minimum thresholds, normally imposed by the user according to the situation. But in a public health study this framework represents a limitation. Some of the interactions between exposures and health outcome can be infrequent or not very strong, but still significant, especially if they capture an important difference between who is exposed and who is not. Fortunately other metrics can be used to determine if a rule is significant for medical purposes. One of them is the odds ratio:

$$\begin{aligned} \mathrm{OR}(X \to Y) &= \frac{P(X \wedge Y)/(1 - P(X \wedge Y))}{P(\neg X \wedge Y)/(1 - P(\neg X \wedge Y))} = \\ &= \frac{supp(XY)/supp(X\neg Y)}{supp(\neg XY)/supp(\neg X\neg Y)} \\ &= \frac{supp(XY)supp(\neg X\neg Y)}{supp(\neg XY)supp(X\neg Y)} \end{aligned} \tag{2.2}$$

The odds ratio allows us to compare health outcomes in two populations differentiated by some exposure(s). A positive (negative) correlation between outcome and exposure exists if the OR is greater (less) than 1. OR = 1 indicates no correlation. Epidemiological studies normally report odds ratios with their confidence interval (CI) $[exp(log(\text{OR}) - \omega), exp(log(\text{OR}) + \omega)]$, where

$$\omega = z_{\alpha/2}\sqrt{\frac{1}{supp(XY)} + \frac{1}{supp(\neg XY)} + \frac{1}{supp(X\neg Y)} + \frac{1}{supp(\neg X\neg Y)}} \tag{2.3}$$

z is the critical value of the confidence interval, and it is typically equal to 1.96, for a 95% level of significance.

ARM can be used to mine rules of the form $X \to Y$ with a significant odds ratio (that is, an odds ratio which confidence interval does not cross 1). X represents a set including one or more exposures and Y is the health outcome under study. The great value of ARM is that it allows us to explore the impact of all possible combinations of exposures and report only those that produce an interesting OR.

Unfortunately, this application, like many other forms of ARM, is affected by the problem of redundant rules. A redundant rule is a rule whose LHS could be simplified by reducing the number of items without any loss of information. For example, the rule $\{Pregnant\} \to \{Age : 20 \div 40\}$ is just as informative as the rule $\{Pregnant, Female\} \to \{Age : 20 \div 40\}$. Or, in the case of a medical study, adding an exposure may not change the odds of having the health outcome. Consider for example the rules $\{Smoker, Female\} \to \{Lung\ cancer\}$ and $\{Smoker\} \to \{Lung\ cancer\}$, resulting in the same odds ratio. Clearly, smoking is responsible for the health outcome. The fact that some subjects were females and smokers did not worsen their odds, even if women and men have different levels of risk. Therefore, the simpler rule should be preferred. Not controlling for redundant rules can cause the number of output rules to grow exponentially and make the results impossible to understand. Kotsiantis and Kanellopoulos [10] offer a good overview of association rule mining techniques and open questions, including a paragraph on redundant association rules. The most popular methods include selection of k best rules [2, 5], mining only maximal itemsets [3, 7], and integration of external knowledge to facilitate the search. This latter became particularly popular in mining relationships in gene expression data [16, 15].

Li *et al.* built an algorithm [14] based on the following assumption: if adding an exposure to a rule does not produce a significant change in OR, the rule should not be reported. The odds ratio between two rules is significantly different if their 95% confidence intervals do not overlap. The assumption seems reasonable, but it is affected by how the non-exposed population is defined in the presence of multiple risk factors. Changing the non-exposed population results in different odds ratio, and as a consequence must be designed carefully. The definition of the exposed population when the antecedent X includes more than one exposure is straightforward: all risk factors must be present at the same time. However, in the current literature, $\neg X$ includes every other possible scenario. This may create some confusion in the interpretation of the rule, because the comparison group is non-homogenous (it includes partially exposed and completely non-exposed subjects). We also argue that it results in a wrong comparison between rules. Consider the rule $\{X\} \to \{Y\}$ and its child $\{Xw\} \to \{Y\}$, where w is a single new exposure ($w \notin X$). We observe the confidence intervals of the rule to determine if they are statistically different. The odds ratio of the parent rule is computed against a non-exposed population composed by the union of $(\neg Xw)$ and $(\neg X\neg w)$. But the odds ratio of the child rule includes in the non-exposed population $(\neg Xw)$, $(\neg X\neg w)$, and also $(X\neg w)$. The non-exposed group of the child rule is not included in the non-exposed group of its parents. This makes the rule intrinsically different.

The alternative definition of non-exposed population, which includes only subjects that have not been exposed to any of the risk factors under examination $((\neg X\neg w))$, results in easily interpretable rules, because the non-exposed group is homogenous. It also offers a more consistent comparison between child and parent rule, because now the non-exposed group of the child rule is a subset of the parent non-exposed group. However, this definition may be problematic as it reduces significantly the size of the non-exposed group, thus reducing the power of the analysis. Furthermore, the comparison with a completely non-exposed population results in higher odds ratios that must be interpreted carefully.

We performed a set of tests to observe the differences in performance obtained by the two definitions of non-exposed group. The following section describes the data used for the experiments and the different tests conducted, followed by a summary of results and discussion.

3 Experimental setting.

3.1 Method. We used a basic A Priori association rule mining algorithm to extract rules from the data (described in details in the following section). We conducted our experiment using the Rstudio environment and the *arules* package [8, 9]. We used three different variations of the basic algorithm to evaluate the effect of pruning and of using different population definitions:

- **Traditional (Trad):** the first method uses the traditional definition of non-exposed population, that is, any subject who has not been exposed to all the risk factors included in the LHS. No pruning criteria is used to filter redundant rules.
- **Traditional + Pruning (TradP):** this method adds to the traditional definition of non-exposed population a pruning criteria of redundant rules based on overlapping of 95% CI.
- **Alternative + Pruning (AltP):** the last method uses the same CI based pruning criteria, but the non-exposed population used to compute the OR is limited to subjects who have not been exposed to any of the risk factors included in the LHS.

All methods have low thresholds for minimum support (1%) and confidence (0.0%) to preserve a large number of rules and observe the differences in the results. The rules that satisfy the requirements of minimum support and confidence were checked for statistically significant OR confidence interval. Only rules with an interval that does not cross 1 were included in the output (for all three methods).

3.2 Data. First, we tested the three methods described in the previous section on six synthetic datasets including 20,000 subjects and 51 features (one indicating whether the subject is a case or a control, the other 50 describing the exposure history). A single rule was embedded in each of the six datasets. By knowing in advance what rule should be found in each dataset, it was possible to evaluate the performance of each algorithm. Embedded rules have different lengths (from 1 to 3 risk factors in combination) and different strengths (weak, $P(Y|X) = 0.4$ or strong, $P(Y|X) = 0.8$). All the features not included in the embedded rule have no impact on the outcome and are potential sources of noise. A baseline probability $P(Y|\neg X) = 0.1$ was introduced to create a population of exposed controls, the absence of which would result in infinite OR. Table 1 offers a summary of the six datasets.

We also tested the three methods on a more complex synthetic dataset designed to have a controlled interaction between exposure and health outcome. The

ID	Rule length	Strength	Cases
1	1	weak	3795
2	1	strong	6197
3	2	weak	2486
4	2	strong	3184
5	3	weak	2124
6	3	strong	2342

Table 1: :
Description of the six single-rule synthetic datasets used in the experiment. Each dataset was embedded with a rule of different length and strength. The last column reports how many of the 20,000 subjects included in each datasets are cases.

data represent a case-control study including 3220 cases and 16780 controls. The database includes six exposures designed to have a different impact on the chances of developing the health outcome. Features are named for ease of understanding. However, the data is not representative of a real clinical study. We gave subjects the following features across the database:

- Age; continuous, uniform distribution from 20 to 80 years.
- Gender; binary (male = 1), p(male)=0.5.
- Smoker; continuous, from 0 to 30 cigarettes per day; p(0 = non smoker) = 0.6; remaining 40% is uniformly distributed.
- Systolic blood pressure (SBP); continuous, normal (mu = 130, sigma = 25).
- Diabetes; binary (diabetes = 1), p(diabetes) = 0.2.
- Daily exercise; categorical (none = 0, light = 1, intense = 2), uniformly distributed.

The features have been designed to have different impact on the chances of contracting the disease. Every subject starts from a baseline probability of 5%. Exposures can have a gradual impact or only act after a certain threshold. They can also be affected by other exposures. Here is the complete list:

- Age: the probability increases by 0.0025 by year of age, starting at 0 for age = 20 and ending at +0.15 for age = 80.
- Gender: no effect.
- Smoker: the impact of cigarettes has been designed as a step function. No impact up to 20 cigarettes per day, then the probability of developing the health outcome rises by 0.4 (+40%).

- High SBP and diabetes: these two features have no impact unless they happen together (diabetes = true and pressure $\geq$ 150). If this condition is verified, the probability of the event goes up by 0.2 (+20%).
- Exercise reduces the risk of cases by 0.2 if light and 0.4 if intense. However, exercise has no effect in case of high blood pressure.

The database described above includes 5 embedded meaningful rules: 3 caused by single exposures (Age, Smoker and Exercise), and 2 caused by interaction between exposures (high SBP with Diabetes, and high SBP with Exercise). A good rule miner should capture all these rules and avoid other less meaningful rules. Less meaningful rules can be divided into two categories: rules caused by simultaneous presence of two or more risk factors, and truly redundant rules. The first category includes those rules that do not represent true interaction between risk factors, but produce a different odds ratio because of their simultaneous presence. For example, we expect the rule $\{Age, Smoker\} \rightarrow \{Event\}$ to result in a higher odds ratio than its single parent rules. Although the rule is not representative of a real interaction between risk factors, it can still be of interest for the study; therefore, we do not penalize methods that output these associations. Truly redundant rules include risk factors whose removal would result in no changes in odds ratio. For example, we expect the rule $\{Male, Smoker\} \rightarrow \{Event\}$ to have approximatively the same OR of $\{Smoker\} \rightarrow \{Event\}$, because gender has no impact. In this case, the longer rule has no added utility and should be avoided.

4 Results.

We recorded the number of rules reported by the different methods when they were used to mine the six one-rule datasets. Ideally, the output should be limited to the one embedded rule. However, this is highly unlikely because of the noise in the data and correlations introduced when embedding the rule. A good output should include the embedded rule and limit the number of other associations.

Every method was able to find the embedded rule in all of the six datasets. The total number of rules found was variable, as visible in Figure 2. Because of the low support and confidence thresholds used and the absence of a pruning criterion for redundant rules, the Traditional method reports a very high number of rules, sometimes over a thousand. This proves that pruning for redundancy can be very useful in lowering the number of output rules, when other selection criteria are missing or less strict.

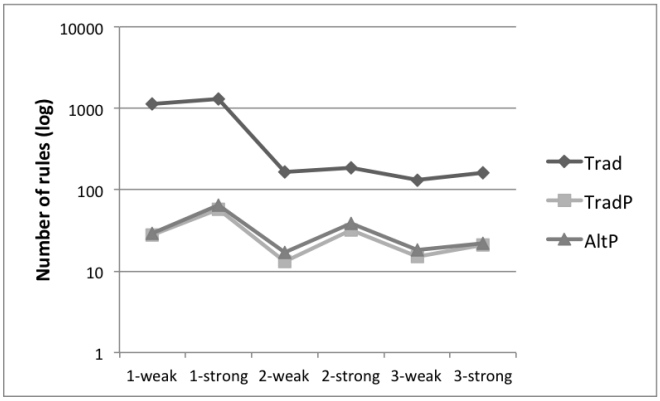


Figure 2: Number of rules found by the different methods in the six one-rule datasets. Datasets are labeled on the x-axis using length of embedded rule and strength of association.

TradP and AltP report a very similar number of rules, although the first method appears slightly more effective at filtering rules and returns in average 3.83 rules less than AltP, per trial. No remarkable difference was found in the overall value of the reported odds ratios and p-values.

When tested on the more complex synthetic dataset, Trad output 23 associations, including the embedded five. 9 rules represented additive effects between risk factors. And 9 of the 23 reported rules were redundant, as they were composed by simpler rules and the risk factor $\{Male\}$, which we know by design has no effect. Again, the high number of redundant rules output shows that this method alone is not effective for the task.

TradP output 14 associations, including the 5 embedded in the synthetic set. 7 rules represented additive effects between risk factors. Two redundant rule were also included: $\{Male, Smoker\} \rightarrow \{Event\}$, and $\{Male, Diabetes\} \rightarrow \{Event\}$.

Alt3 reports a total of 14 rules: the 5 most significant plus 9 additive effects. No redundant rules are reported. A summary of the rules found by each algorithm is visible in Figure 3.

5 Conclusions.

We confirmed that mining with no pruning criteria produces a high number of redundant rules, thus proving the necessity of a process for their elimination. TradP and AltP were both effective in reducing the number of rules and the size of their output is almost identical. However, AltP appeared to be slightly more effective at eliminating redundant rules in a more complex scenario. TradP produced some undesired results, in

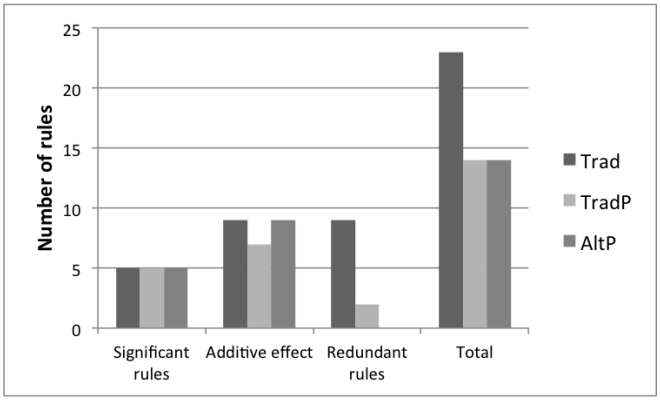


Figure 3: Number of rules found by the different methods. The first three groups of columns represent rules of different quality. The significant rules are important and should be preserved. Additive effects are tolerable. Redundant rules should be avoided.

the form of the rules $\{Male, Smoker\} \rightarrow \{Event\}$, and $\{Male, Diabetes\} \rightarrow \{Event\}$. As suspected, comparing the combination $\{Male, Smoker\}$ against a mixed non-exposed population resulted in an OR different from the parent rules (it is more than $\{Male\}$, but less than $\{Smoker\}$), tricking the algorithm into thinking they were significantly different. If the comparison were made against the uniform non-exposed population, the OR would be similar to the rule with the single $\{Smoker\}$ risk factor and would be pruned. Similar circumstances occurred for $\{Male, Diabetes\} \rightarrow \{Event\}$.

AltP was the only one capable of avoiding all truly redundant rules when mining the more complex database, thanks to a more consistent comparison between populations of child rules with their parents. However, it reported a slightly higher number of rules in the tests done using the one-rule datasets.

TradP appeared to be more resistant against interaction and produced fewer rules caused by additive effects between exposures than AltP, possibly because comparing non-homogenous populations may require more significant differences to be present to produce the necessary change in the odds ratio.

This experiment shows that different definitions of non-exposed groups can be used when using ARM for risk estimate. The differences in using one or the other definition may seem unimportant in these simple mining scenarios, however they represents on a small scale the risk of using the wrong method when mining association rule in large medical databases. In the future, the three methods should be tested on real datasets to better understand their performance when mining perturbed data. We currently do not know what is causing the differences in performance over the proposed datasets. We believe that exploring this question would be beneficial for the development of medical data mining for risk evaluation and of interest for the participants of the workshop.

References

[1] R. Agrawal, T. Imieliński, and A. Swami. Mining association rules between sets of items in large databases. *SIGMOD Rec.*, 22(2):207–216, June 1993.

[2] M. Atzmueller and F. Puppe. Sd-map – a fast algorithm for exhaustive subgroup discovery. In *Knowledge Discovery in Databases: PKDD 2006*, pages 6–17. Springer, 2006.

[3] J. Bayardo and R. J. Efficiently mining long patterns from databases. *SIGMOD Rec.*, 27(2):85–93, jun 1998.

[4] S. Brossette, A. Sprague, J. Hardin, K. Waites, W. Jones, and S. Moser. Association rules and data mining in hospital infection control and public health surveillance. *Journal of Americal Medical Informatics Association*, 5(4):373–81, 1998.

[5] R. Cai, Z. Hao, W. Wen, and H. Huang. Kernel based gene expression pattern discovery and its application on cancer classification. *Neurocomputing*, 73(13-15):2562–2570, aug 2010.

[6] J. Chen, H. He, G. Williams, and H. Jin. Temporal sequence associations for rare events. In H. Dai, R. Srikant, and C. Zhang, editors, *Advances in Knowledge Discovery and Data Mining*, volume 3056 of *Lecture Notes in Computer Science*, pages 235–239. Springer Berlin Heidelberg, 2004.

[7] K. Gouda and M. J. Zaki. Genmax: An efficient algorithm for mining maximal frequent itemsets. *Data Min. Knowl. Discov.*, 11(2):223–242, nov 2005.

[8] M. Hahsler, C. Buchta, B. Gruen, and K. Hornik. *arules: Mining Association Rules and Frequent Itemsets*, 2015.

[9] M. Hahsler, B. Gruen, and K. Hornik. arules – A computational environment for mining association rules and frequent item sets. *Journal of Statistical Software*, 14(15):1–25, October 2005.

[10] S. Kotsiantis and D. Kanellopoulos. Association rules mining: A recent overview. *GESTS International Transactions on Computer Science and Engineering*, 32(1):71–82, 2006.

[11] N. Lavrac. Selected techniques for data mining in medicine. *Artificial Intelligence in Medicine*, 16(1):3–23, May 1999.

[12] D. Lee, K. Ryu, M. Bashir, J.-W. Bae, and K. Ryu. Discovering medical knowledge using association rule mining in young adults with acute myocardial infarction. *Journal of Medical Systems*, 37(2), 2013.

[13] J. Li, A. W. chee Fu, and P. Fahey. Efficient discovery of risk patterns in medical data. *Artificial Intelligence in Medicine*, 45:77–89, 2009.

[14] J. Li, J. Liu, H. Toiovonen, K. Satou, Y. Sun, and B. Sun. Discovering statistically non-redundant subgroups. *Knowledge-Based Discovery*, 67:315–327, 2014.

[15] Y.-C. Liu, C.-P. Cheng, and V. S. Tseng. Discovering relational-based association rules with multiple minimum supports on microarray datasets. *Bioinformatics*, 27(22):3142–3148, 2011.

[16] R. Martinez, N. Pasquier, and C. Pasquier. Mining association rule bases from integrated genomic data and annotations. In *Computational Intelligence Methods for Bioinformatics and Biostatistics*, pages 78–90. Springer, 2008.

[17] J. Nahar, K. Tickle, A. Ali, and Y.-P. Chen. Significant cancer prevention factor extraction: An association rule discovery approach. *Journal of Medical Systems*, 35(3):353–367, 2011.

[18] M. Ohsaki, Y. Sato, H. Yokoi, and T. Yamaguchi. A rule discovery support system for sequential medical data in the case study of a chronic hepatitis dataset. In *Proceedings of the ECML/PKDD-2003 discovery challenge workshop*, pages 154–165, 2002.

[19] C. Ordonez, N. Ezquerra, and C. Santana. Constraining and summarizing association rules in medical data. *Knowledge and Information Systems*, 9(3):1–2, 2006.

[20] J. Paetz and R. Brause. A frequent patterns tree approach for rule generation with categorical septic shock patient data. In J. Crespo, V. Maojo, and F. Martin, editors, *Medical Data Analysis*, volume 2199 of *Lecture Notes in Computer Science*, pages 207–213. Springer Berlin Heidelberg, 2001.

[21] S. Park, S. Jang, H. Kim, and S. Lee. An association rule mining-based framework for understanding lifestyle risk behaviors. *PLoS One*, 9(2), February 2014.

[22] Y. Tai and H. Chiu. Comorbidity study of adhd: applying association rule mining (arm) to national health insurance database of taiwan. *International Journal of Medical Informatics*, 78(12):75–83, December 2009.

Knowledge Transfer with Medical Language Embeddings*

Stephanie L. Hyland[†‡] Theofanis Karaletsos[‡] Gunnar Rätsch[‡]

Abstract

Identifying relationships between concepts is a key aspect of scientific knowledge synthesis. Finding these links often requires a researcher to laboriously search through scientific papers and databases, as the size of these resources grows ever larger. In this paper we describe how distributional semantics can be used to unify structured knowledge graphs with unstructured text to predict new relationships between medical concepts, using a probabilistic generative model. Our approach is also designed to ameliorate data sparsity and scarcity issues in the medical domain, which make language modelling more challenging. Specifically, we integrate the medical relational database (SemMedDB) with text from electronic health records (EHRs) to perform knowledge graph completion. We further demonstrate the ability of our model to predict relationships between tokens not appearing in the relational database.

1 Introduction

The accelerating pace of scientific progress presents both challenge and opportunity to researchers and healthcare providers. Reading and comprehending the ever-growing body of literature is a difficult but necessary part of knowledge discovery and synthesis. This is particularly important for biomedical research, where therapeutic breakthroughs may rely on insights derived from disparate subfields. Curating literature at such breadth and scale is infeasible for individuals, necessitating the development of domain-specific computational approaches.

We present here a method using *language embeddings*. Such an embedding is a representation of the tokens of a language (such as words, or objects in a controlled vocabulary) as elements of a vector space. Semantic similarity is then captured by vector similarity, typically through Euclidean or cosine distance. The dimensionality of the space is typically much less than the size of the vocabulary, so this procedure allows tokens to be represented more compactly while also capturing semantics. Such representations can be used as features in downstream language-processing tasks. In our case, we aim to exploit the embedding *itself* to discover new relationships between tokens. This is possible because our embedding procedure defines a probability distribution over token-relationship-token triples, allowing for questions such as 'is `abdominal pain` more likely to be `associated with` `acute appendicitis` or `pulmonary tuberculosis`?', or 'how is `radium` related to `carcinoma`?'[1]

The tokens of interest are chiefly Concept Unique Identifiers (CUIs) from the Unified Medical Langauage System (UMLS) [3]. These represent discrete *medical concepts*, which may require several words to describe, for example: `C0023473: chronic myelogenous leukemia`. We consider it more meaningful and interesting to consider relationships between CUIs rather than words themselves, when possible. We exploit the exisence of `SemMedDB` [9], a database of semantic predications in the form of subject-relationship-object triples, where the subjects and objects are such CUIs. These were derived from PubMed abstracts using the tool SemRep [16]. We combine this structured data with unstructured text consisting of clinical notes written by physicians at Memorial Sloan Kettering Cancer Center (MSKCC).

2 Related Work

Neural language models [2] are an approach to learning embeddings which use a word's representation to *predict* its surrounding context. This relies on the fact that words with similar meanings have similar contexts (the distributional hypothesis of language [17]), which forces their representations to be similar. Intriguingly, it was observed [13] [4] that the geometry of the resulting space preserved *functional relationships* between terms. An example is a consistent offset vector existing between 'Berlin' and 'Germany', and 'Dublin' and 'Ireland', seemingly representing the relationship `capital city of country`. This property has been exploited to perform knowledge-base completion, for example [4] [18]

*Supported by the Memorial Hospital and the Sloan Kettering Institute (MSKCC; to G.R.). Additional support for S.L.H. was provided by the Tri-Institutional Training Program in Computational Biology and Medicine.

[†]Tri-Institutional Training Program in Computational Biology and Medicine, Weill Cornell Medical College

[‡]Computational Biology Program, Memorial Sloan Kettering Cancer Center (MSKCC)

[1]These are real examples from `SemMedDB`.

[20], however these approaches have restricted their attention to edge-discovery *within* a knowledge graph. To *extend* such a graph we therefore developed a model [7] which can combine structured and unstructured data sources while explicitly modelling the types of relationships present in the structured data.

Despite the popularity of language embeddings in the broader natural language processing (NLP) community, the biomedical domain has yet to fully exploit them. Pedersen et *al.* [15] highlight the need to perform *domain-specific* NLP and discuss measures of semantic relatedness. Other recent applications include using representations of nominal elements of the EHR to predict hospital readmission [11], identifying adverse drug reactions [6], and clinical concept extraction [8].

3 Approach

3.1 Model We briefly describe the `bf` model; see our earlier paper [7] for more details. This is a probabilistic generative model over directed subject-relationship-object triples (S, R, O). Subject and object are both tokens from the vocabulary (e.g., UMLS CUIs), although following [12] and [5] we give them independent representations. This is formulated mathematically through an energy function,

$$\mathcal{E}(S,R,O|\Theta) = -\frac{\mathbf{v}_O \cdot G_R \mathbf{c}_S}{\|\mathbf{v}_O\| \|G_R \mathbf{c}_S\|} \tag{3.1}$$

Entities S and O are represented as vectors, while each representation R corresponds to an *affine transformation* on the vector space. Intuitively, our energy function is the cosine distance between (the representation of) O and S *under the context of* R, where this context-specific similarity is achieved by first transforming the representation of S by the affine transformation associated to R.

This energy function defines a Boltzmann probability distribution over (S, R, O) triples,

$$P(S,R,O|\Theta) = \frac{1}{Z(\Theta)} e^{-\mathcal{E}(S,R,O|\Theta)} \tag{3.2}$$

where the denominator is the partition function, $Z(\Theta) = \sum_{s,r,o} e^{-\mathcal{E}(s,r,o|\Theta)}$. Equation 3.2 defines the probability of observing a triple (S, R, O), given the embedding Θ, which is the set of all vectors $\{\mathbf{c}_s, \mathbf{v}_o\}_{s,o \in \text{tokens}}$ and matrices $\{G_r\}_{r \in \text{relationships}}$.

3.2 Training To learn the embedding (the parameters Θ consisting of all word vectors $\mathbf{c}_s$, $\mathbf{v}_o$, and the relationship matrices G_r), we maximise the joint probability of a set of *true* triples (S, R, T) under this model. Likely pairs have a high cosine similarity (low energy) in the context of their shared relationship, requiring similar vector representations. We employ stochastic maximum likelihood for learning, approximating gradients of the partition function using persistent contrastive divergence [19].

In all cases, we perform early stopping using a held-out validation set. The hyperparameters of the model are as follows: vector dimension is 100, batch size is 100, we use 3 rounds of Gibbs sampling to get model samples, of which we maintain one persistent Markov chain. The learning rate is 0.001 and we use a l_2 regulariser with strength 0.01 on G_r parameters. To make learning more stable, we use Adam [10] with hyperparameters as suggested in the original paper.

3.3 Prediction Equation 3.2 defines a joint distribution over triples. However, we are often interested in *conditional* probabilities: given a pair of entities S and O, which R most likely exists between them (if any)? Such a distribution over R (or equivalently S, O) can easily be derived from the joint distribution, for example:

$$P(R|S,O;\Theta) = \frac{e^{-\mathcal{E}(S,R,O|\Theta)}}{\sum_r e^{-\mathcal{E}(S,r,O|\Theta)}} \tag{3.3}$$

The cost of calculating the conditional probability is at worst linear in the size of the vocabulary, as the (generally intractable) partition function is not required.

4 Experiments

4.1 Data preparation We train the model on two types of data: *unstructured* (`EHR`) and *structured* (`SemMedDB`). The unstructured data is a corpus of de-identified clinical notes written by physicians at MSKCC. We process raw text by replacing numbers with generic tokens such as `HEIGHT` or `YEAR`, and removing most punctuation. In total, the corpus contains 99,334,543 sentences, of which 46,242,167 are unique. This demonstrates the prevalence of terse language and sentence fragments in clinical text; for example the fragment `no known drug allergies` appears 192,334 times as a sentence. We identify CUIs in this text by greedily matching against strings associated with CUIs (each CUI can have multiple such strings). This results in 45,402 unique CUIs, leaving 270,100 non-CUI word tokens. We note that the MetaMap [1] tool is a more sophisticated approach for this task, but found it too inefficient to use on a dataset of our size. To generate (S, R, O) triples, we consider two words in a `appears in a sentence with` relationship if they are within a five-word window of each other.

The structured data (`SemMedDB`) consists of `CUI-relationship-CUI` statements, for example

C0027530(Neck) is LOCATION OF C0039979(Thoracic Duct) or C0013798(Electrocardiogram) DIAGNOSES C0026269(Mitral Valve Stenosis). These were derived from PubMed abstracts using SemRep [16]. SemMedDB contains 82,239,653 such statements, of which 16,305,000 are unique. This covers 237,269 unique CUIs.

Since the distribution of CUI/token frequencies has a long tail in both data sources, we threshold tokens by their frequency. Firstly, tokens (words of CUIs) must appear at least 100 times in either dataset, and then at least 50 times in the pruned datasets. That is, in the first round we remove sentences (in EHR) or statements (in SemMedDB) containing 'rare' tokens. In addition, the 58 relationships in SemMedDB also exhibit a long-tailed frequency distribution, so we retain only the top twenty.

From this pool of (S, R, O) triples (from EHR and SemMedDB) we create fixed test sets (see next subsection) and smaller datasets with varying relative abundances of each data type, using 0, 10, 50, 100, 500, and 1000 thousand training examples. The final list of tokens has size $W = 45,586$, with 21 relationships: twenty from SemMedDB and an additional `appears in sentence with` from EHR. Of the W tokens, 7,510 appear in both data sources. These overlapping tokens are critical to ensure embeddings derived from the knowledge graph are consistent with those derived from the free text, allowing information transfer.

4.2 Knowledge-base completion

Experimental design As the model defines conditional distributions for each element of a triple given the remaining two (Equation 3.3), we can test the ability to predict new components of a knowledge graph. For example, by selecting the best R given S and O, we predict the relationship (the type of edge) between tokens S and O.

Without loss of generality, we describe the procedure for generating the test set for the R task. We select a random set of S, O pairs appearing in the data. For each pair, we record all entities r which appear in a triple with them, removing these triples from the training set. The $S, O \to \{r_i\}_i$ task is then recorded in the test set. Evidently, there may be many correct completions of a triple; in this case we expect the model to distribute probability mass across all answers. How best to evaluate this is task-dependent; we consider both the *rank* and the *combined probability mass* in these experiments.

Results Figure 1 shows results for the task of predicting R given S and O. The model produces a ranking of all possible Rs (high probability $\to$ low rank) and we report the mean reciprocal rank of the

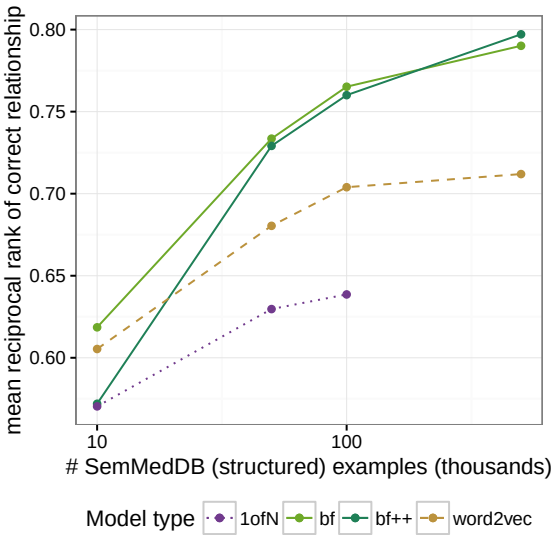


Figure 1: With more evidence from the knowledge graph, the model better predicts the correct relationship for a given (S, O) pair. `bf++` has an additional 100,000 triples from EHR: with little structured data, so much off-task information is harmful, but provides some benefit when there is enough signal from the knowledge graph. Baselines are a random forest taking $[f(S) : f(O)]$ as an input to predict the label R, where the feature representation f is either a 1-hot encoding (`1ofN`) or 200-dimensional `word2vec` vectors trained on PubMed. `1ofN` proved too computationally expensive for large data.

lowest-ranked correct answer over the test set. We use this metric to evaluate the utility of these predictions in *prioritising* hypotheses to test: we would like *any* correct answer to be ranked highly, and don't apply a penalty for a failure to capture alternative answers. Results for our model are marked by `bf`[2] and `bf++`. The latter model uses an additional 100,000 training examples from the EHR: these are 'off-task' information. As a baseline we consider a random forest trained to predict R given the concatenation $[f(S) : f(O)]$, where the representation f is either: a) `1ofN`: each token has a binary vector of length W ($W = 45,586$), b) `word2vec`: each token has a 200-dimensional vector obtained by running `word2vec` [12] trained on PubMed [14]. We note that the PubMed corpus contains over 2 billions tokens, far more data than was available to `bf`. We additionally trained `TransE` [4] on this data, but it proved unsuited to the task (data not shown).

As we can see, adding examples from SemMedDB improves performance for all model types, but `bf` seems to make better use of the additional data. In spite of

[2] `bf` stands for 'brí-focal', which means *word meaning* in Irish.

its very large input vector size ($2W = 91172$), `1ofN` struggles, likely as it treats all tokens as independent entities. We note that for `bf++`, performance is *degraded* when the amount of structured data is low. This is consistent with earlier observations on non-medical data [7], as the quantity of 'off-task' information added is in this case comparable to that of 'on-task'. Interestingly however, the model appears slightly *better able* to exploit more structured data when some 'semantic background' is provided by `EHR`.

4.3 Information transfer

Experimental design As mentioned, the model is capable of combining structured and unstructured data. In [7] we observed that classification performance on a knowledge base could be improved by addition of unstructured data. However, the task in that case was quite 'easy'; the model simply needed to differentiate between true and false triples. Here we consider the harder problem of correctly selecting *which* entity would complete the triple.

In addition to possibly improving performance, access to unstructured data provides the opportunity to *augment* the knowlede base. That is, we can predict relationships for tokens *not appearing* in `SemMedDB`. This uses the joint embedding of all tokens into one vector space, regardless of their data source. The geometric action of the relationships learned from `SemMedDB` can then be applied to the representation of any token, such as those uniquely found in `EHR`. We note that this procedure amounts to *label transfer* from structured to unstructured examples, which can be understood as a form of semi-supervised learning.

To generate ground truth for this task, we select some tokens $\{T_i\}$ (these could appear as S or O entities) found in both `SemMedDB` and `EHR` and remove them from `SemMedDB`, recording them to use in the test set. Put another way, as in the previous setting, during the 'random' selection of S, O (still wlog) pairs, we make sure all of these recording them to use in the test set. Put another way, as in the previous setting, during the 'random' selection of S, O (still wlog) pairs, we make sure all T_i in the deletion list are included, alongside any other tokens which appear in a `SemMedDB`-derived relationship with them. The task is then to use purely *semantic* similarity gleaned from `EHR` to place these tokens in the embedding space such that the action of relationship operators is still meaningful.

Results Figure 2 shows results on all three tasks (predicting S, R, O given the remaining two), as a function of the *type of test example*. The right column of results is for test entities involving at least one element *not appearing* in `SemMedDB`. As we are now interested

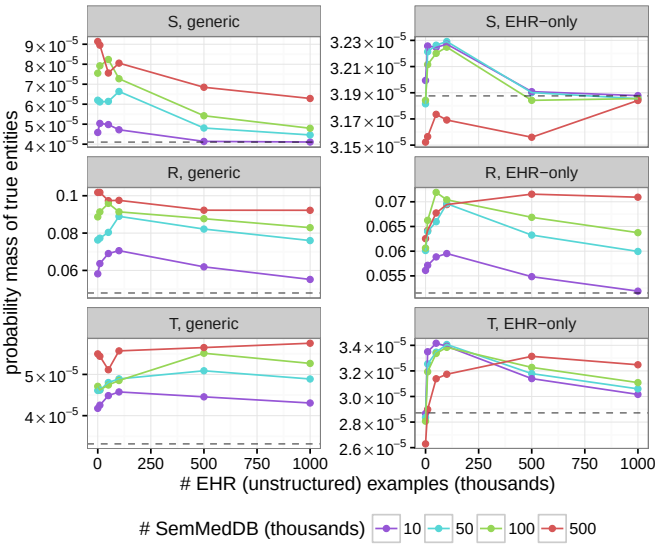


Figure 2: Total probability mass assigned to correct answers for all tasks. The right column shows results for test triples where at least one of S and O is found *only* in `EHR`, and therefore represents the *knowledge transfer* setting. Information about relationships found in `SemMedDB` must be transferred through the joint embedding to enable these predictions. Grey dotted lines represent a random-guessing baseline.

in the *embeddings themselves* we report the probability mass of true entities, feeling this better captures the information contained in the embeddings. That is, it is no longer sufficient for the model to correctly predict *a single* answer, we want it to assign appropriate probability mass to *all* correct answers. The dotted grey lines demonstrate the random baseline, where all tokens are equally likely. The probability mass assigned by the baseline is therefore equal to k/W (or k/R) where k is the average number of correct options in that task type.

There are several observations to be made here:

- Most of the time, performance is best with a non-zero, but *relatively small* amount of `EHR` data (x-axis). This supports our observations that off-task information improves embeddings, but can 'drown out' signal if it dominates relative to the on-task examples. This can be improved by including a pre-factor on gradient contributions from the off-task data to adjust their contribution relative to the structured examples, as demonstrated in our previous work [7].
- The `EHR-only` setting is much harder, as anticipated. In the case of S and O it is comparable to the random baseline. For R however, the model successfully assigns probability mass when there is enough `SemMedDB` data available.

- The S and O tasks are not symmetric. The S task features slightly more correct options on average than O (1.87 and 1.5 respectively, for the `generic` task), but this does not account for the difference in proportional performance relative baseline, especially at low `EHR` abundance. A possible explanation is the energy function (Equation 3.1): it does not treat S-type and O-type variables identically. However, experiments using the Frobenius norm of G_R in the denominator of $\mathcal{E}$ did not remove asymmetry, so it is likely that the tasks are simply not equivalent. This could arise due to bias in the directionality of edges in the knowledge graph.

We conclude that it is possible to use the joint embedding procedure to predict R for pairs of S, O entities even if they do not appear in `SemMedDB`. For the harder S and O tasks, the model generally succeeds in improving visibly over baseline, but its assignments are still quite 'soft'. This may reflect premature stopping during training (most results reported were before 50 epochs had elapsed), an insufficiently powerful model formulation, or an excess of noise in the training data. Many predicates in `SemMedDB` are vague, and some relationships lend themselves to a one-to-many situation, for example `part of`, or `location of`. A core assumption in our model is that a token with fixed vector representation can be transformed by a single affine transformation to be similar to its partner in a relationship. Many-to-one (or vice-versa) type relationships requires that multiple unique locations must be mapped to the same point, which necessitates a rank-deficient linear operator or a more complex transformation function (one which is locally-sensitive, for example). Future work in relational modelling must carefully address the issue of many-to-many and hierarchical relationships.

5 Discussion

Distributed language representations have seen limited application in healthcare to date, but present a potentially very powerful tool for analysis and discovery. We have demonstrated their use in knowledge synthesis and text mining using a probabilistic generative model which combines structured and unstructured data. These embeddings can further be used in downstream tasks, for example to reduce variation in language use between doctors (by identifying and collapsing similar terms), for 'fuzzy' term-matching, or as inputs to *compositional* approaches to represent larger structures such as sentences, documents, or even patients. Expressive knowledge representations such as these will be facilitate richer clinical data analysis in the future.

References

[1] Aronson, A. R. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proceedings of the AMIA Symposium* (2001), American Medical Informatics Association, p. 17.

[2] Bengio, Y., Ducharme, R., Vincent, P., and Janvin, C. A neural probabilistic language model. *J. Mach. Learn. Res. 3* (Mar. 2003), 1137–1155.

[3] Bodenreider, O. The unified medical language system (umls): integrating biomedical terminology. *Nucleic Acids Research 32* (2004), D267–D270.

[4] Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., and Yakhnenko, O. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems (NIPS)* (2013), pp. 2787–2795.

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

[6] Henriksson, A. Representing clinical notes for adverse drug event detection. In *Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis* (2015), Association for Computational Linguistics, pp. 152–158.

[7] Hyland, S. L., Karaletsos, T., and Rätsch, G. A generative model of words and relationships from multiple sources. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence* (2016).

[8] Jonnalagadda, S., Cohen, T., Wu, S., and Gonzalez, G. Enhancing clinical concept extraction with distributional semantics. *Journal of biomedical informatics 45*, 1 (2012), 129–140.

[9] Kilicoglu, H., Shin, D., Fiszman, M., Rosemblat, G., and Rindflesch, T. C. Semmeddb: a pubmed-scale repository of biomedical semantic predications. *Bioinformatics 28*, 23 (2012), 3158–3160.

[10] Kingma, D., and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[11] Krompass, D., Esteban, C., Tresp, V., Sedlmayr, M., and Ganslandt, T. Exploiting latent embeddings of nominal clinical data for predicting hospital readmission. *KI - Künstliche Intelligenz 29*, 2 (2014), 153–159.

[12] Mikolov, T., Chen, K., Corrado, G., and Dean, J. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[13] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems (NIPS)* (2013), pp. 3111–3119.

[14] Moen, S. P. F. G. H., and Ananiadou, T. S. S. Distributional semantics resources for biomedical text processing. *LBM* (2013).

[15] Pedersen, T., Pakhomov, S. V., Patwardhan, S., and Chute, C. G. Measures of semantic similarity and relatedness in the biomedical domain. *Journal of Biomedical Informatics 40*, 3 (2007), 288 – 299.

[16] Rindflesch, T. C., and Fiszman, M. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. *Journal of Biomedical Informatics 36*, 6 (2003), 462 – 477. Unified Medical Language System.

[17] Sahlgren, M. The distributional hypothesis. *Italian Journal of Linguistics 20*, 1 (2008), 33–53.

[18] Socher, R., Chen, D., Manning, C. D., and Ng, A. Reasoning with neural tensor networks for knowledge base completion. In *Advances in Neural Information Processing Systems (NIPS)* (2013), pp. 926–934.

[19] Tieleman, T. Training restricted boltzmann machines using approximations to the likelihood gradient. In *International Conference on Machine Learning (ICML)* (2008), pp. 1064–1071.

[20] Weston, J., Bordes, A., Yakhnenko, O., and Usunier, N. Connecting language and knowledge bases with embedding models for relation extraction. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2013), pp. 1366–1371.

IDEA: Integrative Detection of Early-stage Alzheimer's disease

Wei Ye* Bianca Wackersreuther* Christian Böhm* Michael Ewers† Claudia Plant‡

Abstract

Data integration and selecting only the relevant information for solving biological and environmental problems is one of the most important challenges in today's data mining. One urgent problem in the medical community is to support the classification of dementia caused by Alzheimer's disease and even its detection in the predementia phase to optimize the medical treatment of a disease that accounts for 60 to 80 percent of dementia cases and affects more than 35 million people world-wide. In this paper we present IDEA, a fully automated, easy-to-use and clinically interpretable diagnostic software for early-stage Alzheimer's. The main contribution of our framework is that it allows for a combined analysis of various feature types such as neuroimaging data sourcing from different modalities, and non-image data that consist of numerical and categorical values, resulting in high classification accuracy results. Using advanced information theory, we select only subsets out of the rich pool of information that build high-predictive feature combinations. In an extensive medical case-study on a large real-world data set, we show that already small feature subsets are adequate to derive significant classification accuracies. And, as IDEA usually determines more than one suitable feature set, it even can be used for an optimized analysis process by selecting the assessment tools that produce minimal cost (in terms of money or stress for the patients) without loosing accuracy.

1 Introduction

Analyzing mixed-type attributes or also known as integrative data mining is among the top 10 challenging problems in data mining research identified in panel discussions [19] and position papers [25]. Moreover, it is essential for solving many of the other top 10 challenges, including data mining in social networks and data mining for biological and environmental problems. In this paper, we address the application of integrative data mining for the detection of early-stage patterns for Alzheimers's disease (AD) dementia, by a combined analysis of different medical imaging modalities together with multiple numerical and categorical attributes, resulting from neuropsychological tests or genetic and biochemical screenings.

AD is the most common form of dementia, that usually develops slowly and includes gradual onset of cognitive impairment in episodic memory and at least one other domain [16]. Although, there is currently no cure for Alzheimer's that stops the disease from progressing, medical treatment can temporarily slow down the worsening of dementia symptoms. However, the benefit of this treatment strongly correlates with a reliable early detection of AD in predementia stages such as mild cognitive impairment (MCI). But, cerebral or cognitive changes are only of subtle degree at MCI stages, and therefore much harder to detect.

Usually AD is diagnosed on the basis of a patient's medical history and a variety of cognitive tests. Most of these tests produce sets of continuous numerical values or categorize a certain screening result in predefined bins. In order to exclude other cerebral pathology or subtypes of dementia, advanced medical imaging techniques, like initially computed tomography (CT) and then magnetic resonance imaging (MRI), are often used. Structural MRI detects tissue changes in the grey and white matter of the human brain. Cognitive task-related changes in brain activity and basal brain activity during resting state are assessed by functional MRI (fMRI). The positron emission tomography (PET) visualizes and quantifies abnormal structures called plaques caused by the protein amyloid-beta (Aβ) in the brains of patients with AD, even in stages of MCI or complete presymptomatic states. Figure 1 shows a hypothetical model of the predicted utility during the progression of AD for different biomarkers, following the studies of Jack et al. [14].

Consequently, we do not rely on single test modes in this project, but rather combine different sources to determine individual risk profiles. We develop IDEA, a new Integrative Detection framework for Early-stage AD patterns. We exploit an unprecedented amount of heterogeneous knowledge sources, including multimodal neuroimaging, biochemical markers and neuropsychological tests. However, the essential effort (in terms of money, time and stress factor for the patients) for collecting the data strongly depends on the different data acquisition tools. Consequently, we select a set of relevant key features yielding best possible classification results concerning both accuracy and cost-effectiveness based on an information-theoretic driven feature selection, and pro-

*Department of Computer Science, LMU, Munich, Germany
{ye, wackersreuther, boehm}@dbs.ifi.lmu.de
†Institute for Stroke and Dementia Research, LMU, Munich, Germany
{michael.ewers@med.uni-muenchen.de}
‡Faculty of Computer Science, University of Vienna, Austria
{claudia.plant@univie.ac.at}

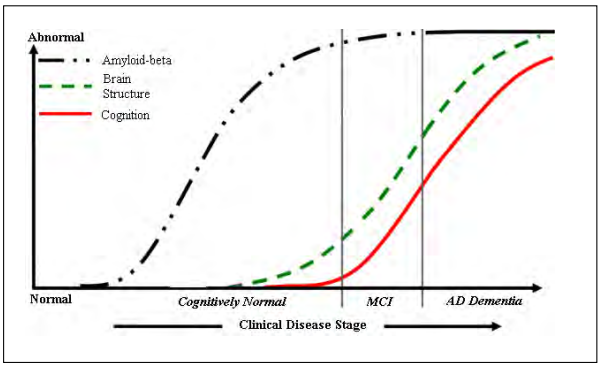


Figure 1: Predicted utility of various biomarkers during the progression of Alzheimer's.

vide a suggestion for the most promising association of different assessment tools. Therefore, IDEA provides two main contributions.

1. A combined analysis of image and non-image data achieves more accurate prediction results.
2. Unavailable measures (due to any reason) can be replaced by equivalent sets of feature combinations.

The rest of this paper is organized as follows: Section 2 gives a brief survey of the large previous work on integrative data mining and related research for early-stage detection of Alzheimer's disease. Section 3 presents our new diagnosis framework which performs heterogeneous data mining for image, numerical and categorical data to achieve high accurate risk profiles for Alzheimer's disease. Section 4 documents a medical case-study, where we present each processing step on a real-world data set provided by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (`http://adni.loni.usc.edu/`). Finally, Section 5 summarizes the paper.

2 Related Work

In this section, we survey relevant research in the field of data integration and describe related classification approaches for neuroscience application.

Integrative Data Mining. Several papers address the problem of finding dependencies among heterogeneous data. Most integrative clustering approaches, as for instance the algorithms K-Prototypes [13], CFIKP [26], CAVE [12], and K-means-mixed [1] rely on the basic algorithmic paradigm of K-means [11]. While K-means focuses on clustering numerical data, the aforementioned approaches typically use several different optimization goals, one for each data type. Whenever these goals disagree, a manually chosen weighting factor has to decide how to resolve this tie situation. But, it is not trivial to select a suitable weighting factor that is valid for different clusters or for a complete clustering process (while the clusters evolve). Moreover, such approaches implicitly assume the independence between attributes of different types. More advanced solutions, like INTEGRATE [2] or INCONCO [20], consider the task of learning weighting factors and even the number of expected clusters K to detect dependencies between attributes (of the same or different type) as part of the overall clustering process.

The proposed ideas for integrative clustering can be easily applied for a classification scenario. But none of these approaches are suitable for the combination of numerical, categorical and *imaging* data. Rather, we present a solution for this clinically relevant task without the need of challenging parameter settings by using advanced information-theoretic techniques.

Classification of Neuroimaging Data for Early Stage AD Detection. Pattern classifiction methods on the basis of high-dimensional neuroimaging data sets are promising tools to aid the clinical diagnosis of Alzheimer's fully automatically. Support vector machines (SVM) have been applied in a number of studies to exploit structual or functional MRI and PET images for the early diagnosis of AD in MCI and healthy controls [7, 15] and also have been applied to multicenter MRI data sets [5]. However, the cross-validation results of SVM derived patterns show only limited robustness for the prediction of clinical progression in MCI. Other classification algorithms such as Bayes statistics and voting feature intervals show clinically acceptable accuracy ($> 85\%$) for the detection of AD dementia, but insufficient accuracy for the prediction of AD dementia at the MCI stage [21, 3]. A major reason for the limited clinical applicability for the early detection is the inherent heterogeneity of brain changes that are characteristic of AD. In keeping with the diagnostic guidelines, we propose here to source different types of measures including neuroimaging, biochemical markers, genetic features and neuropsychological tests.

The most related approach is the work by Shuo Xiang et al. [24], that examines AD prediction on the basis of heterogeneous data with the focus on missing values. However, besides balancing missing attributes, IDEA tries to find an optimal set of independent features by identifying redundant information sources.

3 Integrative Detection of Early-stage AD Patterns

The first step of the integrative diagnosis framework IDEA is selecting the most informative features of each data modality (neuroimaging, numerical or categorical). This step deserves high diligence, because selecting subsets of strong discriminating features is indispensable for reliable classification results.

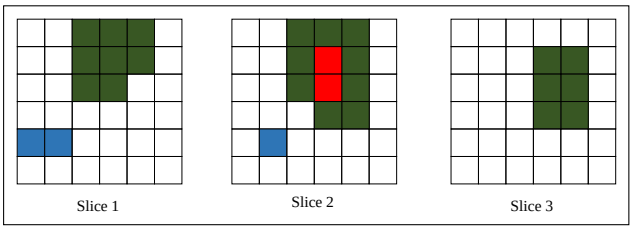


Figure 2: Example for a 3-dimensional DBSCAN used for density-based clustering of neuroimaging data to find brain regions with high-discriminatory power.

3.1 Feature Selection

With the Information Gain (IG) [22, 10], we perform class separation based on the concept of entropy, as IG rates the interestingness of a specific attribute (e.g. one voxel of neuroimage scan) for the separation. To formalize the IG, first the definition of the entropy of the class distribution is needed.

DEFINITION 3.1. (ENTROPY OF CLASS DISTRIBUTION) *Given a class c_i (e.g. AD patients) and its corresponding class probability $p(c_i)$, the entropy of the class distribution is defined as follows:*

$$H(C) = \sum_{c_i \in C} p(c_i) \cdot log_2(p(c_i)).$$

$H(C)$ corresponds to the required amount of bits to predict the class of an unknown subject and scales between 0 and 1.

The entropy of the class distribution before and after observing an attribute a refers to the information gain (IG) of a and is formally defined as follows:

DEFINITION 3.2. (INFORMATION GAIN) *Given an attribute a (e.g. a voxel), the information gain of a is:*

$$IG(a) = H(C) - H(C|a).$$

In the case of $k = 2$ (e.g. if we consider the classes MCI and AD), IG scales between 0 and 1, where 0 means that the attribute a provides no information on class label of the subject. An IG of 1 means that the class label of all subjects can be derived from the corresponding attribute a without any error.

We can compute the IG for each attribute type, regardless of being an image, numerical or categorical attribute. For features with continuous values (e.g. voxel intensities), we apply the discretization algorithm by Fayyad and Irani [8], which divides the attribute range into class pure intervals, where the IG of the split defines the cut points. To avoid a disproportional high number of cut points, the MDL principle is used to determine the optimal number and location of the cut points. For all attributes, regardless of arising in an image data or not, we hereby calculate class-separation information without the need for data format transformations, which means that we combine the different data types without loss. Only features that have an IG value above a specified threshold IG_{opt} are kept for further processing.

However, the huge amount of information present especially in the neuroimaging data (each image consists of more than two million voxels) poses a major problem for the automated analysis including noisy data and replicability, irrelevant information, and costs in terms of data acquisition and processing time. For this purpose, we apply a density-based clustering approach on the spatially complex imaging data. Thereby, we receive connected brain *regions* which are much more informative for further processing than single voxels.

3.2 Clustering of Neuroimaging Data

In general, clustering algorithms aim at deriving a partitioning of the data into groups (clusters) such that similar objects are grouped together. To identify groups of adjacent voxels that commonly share high IG values, and to remove noise in the imaging data, we use a variant of the well-established density-based clustering approach DBSCAN [6] as recommended in the paper of Plant et al. [21]. Density-based clustering algorithms are designed to find clusters of arbitrary shape in noisy data.

The notion of the original DBSCAN algorithm, which was designed for clustering data objects represented by feature vectors, is defined as follows. An object O is called **core object** if it has at least $MinPts$ objects in its ϵ-range, i.e. $|N_\epsilon(O) \geq MinPts|$, where $N_\epsilon(O) = \{O' | dist(O, O') \leq \epsilon\}$. An object O is **directly density reachable** from another object P w.r.t. ϵ and $MinPts$ if P is a core object and $O \in N_\epsilon(P)$. An object O is **density-reachable** from an object P w.r.t. ϵ and $MinPts$ if there exists a sequence of objects $O_1, \cdots, O_n$ such that $O_1 = P$ and $O_n = O$ and O_{i+1} is directly density-reachable w.r.t. ϵ and $MinPts$ from O_i for $1 \leq i \leq n$. Two objects O and P are **density-connected** w.r.t. ϵ and $MinPts$ if there exists an object Q such that both O and P are density-reachable from Q. A **density-based cluster** is the maximum set of density-connected objects, i.e. the transitive closure of the density reachability relation.

To adapt this algorithm to the setting of neuroimage data, where each object is represented by 3-dimensional voxels, a **core voxel** is a voxel, which is surrounded by at least six voxels that commonly share an IG value higher than IG_{opt}. Figure 2 illustrates an example. It shows three sequent slices in the brain, each of which contains 6×6 voxels. Colored voxels (red, blue or green) indicate voxels with high IG-values. The red voxels are core voxels w.r.t. $\epsilon = 1$ and $MinPts = 6$. The blue voxels are noise and the green voxels are density-reachable w.r.t. to the given values of ϵ and $MinPts$.

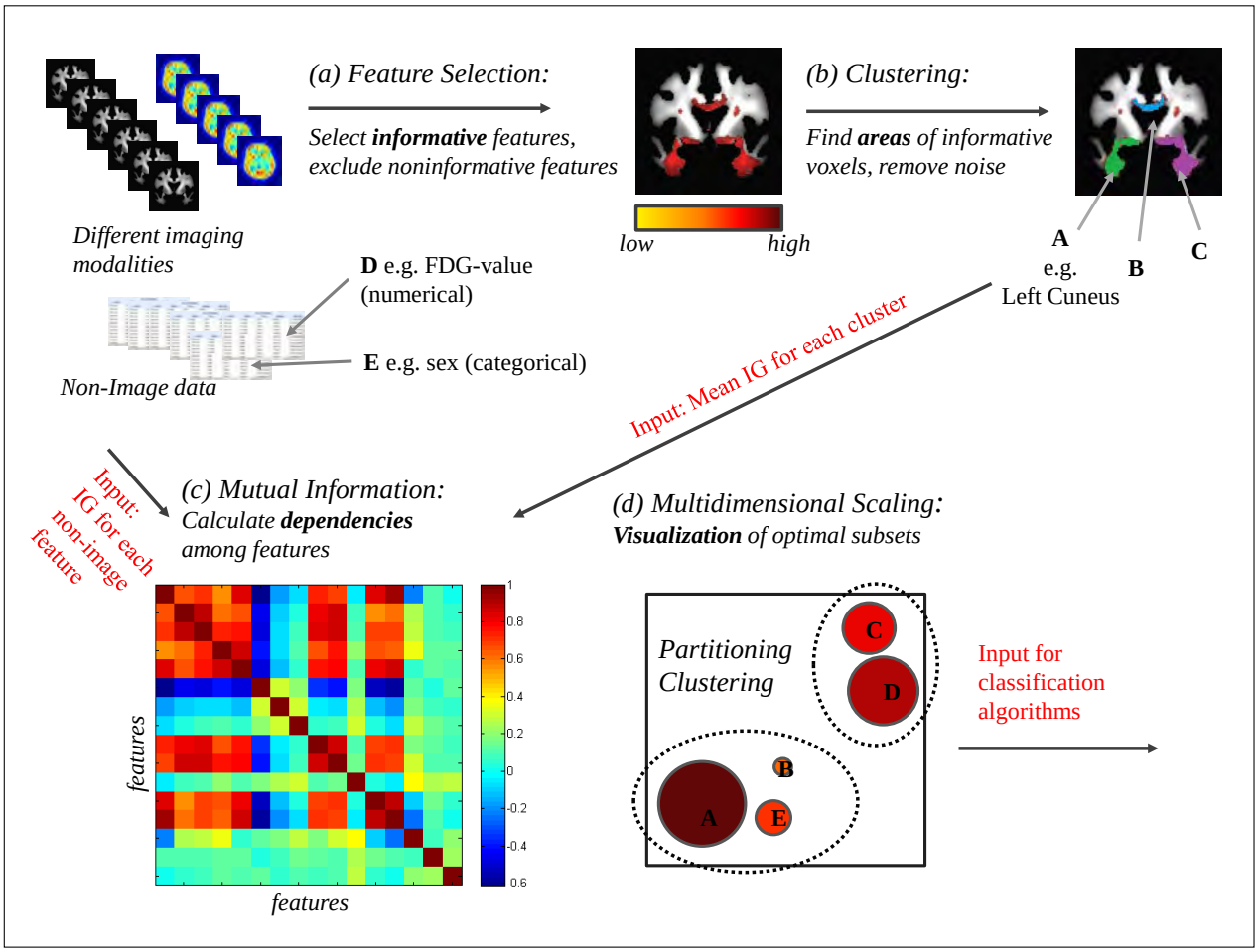


Figure 3: Data analysis stream from heterogeneous data sources to the visualization of optimal feature sets. Input of the combined analysis are the IG values of numerical and categorical non-image data and a representative IG-value per informative voxel cluster for each neuroimage modality. The calculation of the pairwise mutual information leads to feature subsets that provide maximum information for the classification process, visualized by multidimensional scaling. The feature sets, determined by partitioning clustering, serve as input for the classification algorithms.

After selecting single informative features, IDEA computes dependencies among all possible pairs of attributes regardless of being a clustered neuroimaging feature or being a numerical or categorical non-image assessment value in the next step.

3.3 Calculating Dependencies Among Features

To build sets of informative features, we use the concept of mutual information (MI) as suggested by Peng et al. [17]. Hereby, IDEA rates the information dependencies among the different attributes. Informative brain regions that result from the aforementioned feature selection step are represented by the mean values of the corresponding voxels. MI is not limited to real-valued random variables like the correlation coefficient, but rather MI is more general and determines how similar the joint distribution of two random variables x and y is.

DEFINITION 3.3. (MUTUAL INFORMATION) *Given two random variables x and y, their mutual information is defined in terms of their probabilistic density functions $p(x)$, $p(y)$ and $p(x, y)$:*

$$MI(x;y) = \int \int p(x,y) log \frac{p(x,y)}{p(x)p(y)} dxdy.$$

The resulting MI-matrix forms a metric space, which enables us to determine irrelevant or redundant information sourcing from the various analysis methods for the clinical diagnosis of AD. This means, that the clinician might chose only one assessment modality out of multiple redundant features to

reduce cost, or receives a recommendation for further tests that maximize the accuracy of the classification.

3.4 Visualization of Feature Subsets

Finally, the results of the MI-matrix are representable in 2-dimensional space to facilitate the application of our heterogeneous data mining approach in the clinical environment. For this purpose, we use a standard technique in statistics called multidimensional scaling (MDS) [4]. For a measure of the global similarity among a set of features (in our case the MI matrix), MDS provides a spatial configuration, in which the distances between the objects match their proximities as closely as possible. Each object in the spatial configuration (each point in the visual mapping) is one assessable attribute, its radius visualizes its IG, which is an additional criterion for an optimal subset configuration. The smaller the distance between two objects is, the higher is the amount of redundant information. Therefore, an optimal subset of measures consists of attributes with large radius and high distance to each other.

In order to build sets of informative features, IDEA performs partitioning clustering (e.g. K-means [11]), where each cluster represents one source of independent information. As each cluster usually contains several attributes, we select the features of one cluster according their IG-values. If one attribute can not be assessed (due to expensive costs or accessability) a feature in its direct neigborhood is chosen instead.

3.5 Summary and Implementation Details

Figure 3 summarizes the overall workflow for our integrative diagnosis tool IDEA. After identifying the most informative voxels in all imaging modalities in step (a), a clustering algorithm groups these voxels into areas of interest in step (b) that can be mapped to real anatomical brain regions, e.g. 'Left Cuneus'. Together with the non-image data (e.g. the FDG-value and the sex of the subject), the pairwise mutual information is calculated in step (c). By use of multidimensional scaling, the pairwise dependencies are visualized. This can be used to decide which measure should be assessed to achieve best accuracy with minimal number of tests. In our example, 'Left Cuneus' (feature A) and the FDG-value (feature D) provide the highest IG (radius sizes of the circles correspond to IG values) and therefore should be favored. However, if A is not an option for any reason, feature E is closest to A and thus the best alternative, as E and A share a lot of common information, while C (higher IG) is redundant to D. The detected feature sets are the input data for the classification algorithms.

The implementation of IDEA roughly consists of three parts. Part (1) determines the best IG threshold value for each fold of image data. Part (2) is dedicated to masking the training data and test data in each fold, and part (3) integrates data from different sources. For the first step, we store the candidate IG threshold values in a vector $\boldsymbol{t}$, and select the optimal value by 10-fold cross validation. For part (2), we perform an IG-based feature selection on the training data and mask the test data in each fold, i.e. we keep voxels in the test data which have the same positions as those kept in the training data. Part (3) is the core part of IDEA. Here, each image cluster is represented by its mean image intensity value. We combine the mean value matrix with non-image data and compute pairwise MI. After applying partitioning clustering in the space returned by MDS on MI-matrix, each cluster is represented by the feature with highest IG value. Finally, IDEA performs Support Vector Machine (SVM) classification with polynomial kernel [23] on selected features.

4 Experimental Evaluation

In this section, we present our medical case-study for early-stage AD pattern detection on an open-source data set.

4.1 The Data

We evaluate IDEA on a study that was conducted in the years 2005 to 2007 and attended by 395 participants. The corresponding data set is obtained from the Alzheimer's Disease Neuroimaging Initiative (`http://adni.loni.usc.edu/`). It includes scans for 98 healthy control subjects (HC), 201 patients with MCI (amnestic w.r.t. the study by *Petersen et al.* [18]) and 96 patients with clinically probable AD dementia (referring to *McKhann et al.* [16]). All subjects underwent volumetric MRI (performed on a T1 MRI scanner) and PET, resulting in $121 \times 145 \times 121$ voxels per scan. In addition, the data set provides information for multiple clinical examinations. Table 1 summarizes eight non-image attributes we used for further processing, including demographic variables (e.g. age and sex), biochemical measures (e.g. FDG), genetics (e.g. ApoE genotype) and neuropsychological test scores (e.g. MMSE). The epsilon 4 allele of APOE is the strongest known genetic risk factor for AD with a two- to three-fold increased risk for AD in people with one allele of this kind, rising up to approximately 12-fold in those with two alleles.

4.2 IG-based Feature Selection

Our medical case-study includes three different settings, namely AD vs. HC, AD vs. MCI and HC vs. MCI. To process the neuroimaging data, all scans were randomly divided and stratified w.r.t. the diagnosis into ten folds using 10-fold cross-validation. For each experiment, we also used 10-fold cross-validation on the training data to select a suitable information gain threshold IG_{opt} in a range of $0.02, 0.04, \cdots, 0.5$. To determine relevant brain regions, IDEA performs density-based clustering (cf. Section 3.2)

Table 1: Demographic, biochemical, genetic and neuropsychological variables for the different groups. For each numerical attribute, we report mean and standard deviation of the underlying values. For each categorical variable, we specify the number of subjects in each category.

Attribute	Type	HC	MCI	AD
Age	numerical	$\mu_{Age} = 74.75$ $\sigma_{Age} = 6.90$	$\mu_{Age} = 75.50$ $\sigma_{Age} = 6.60$	$\mu_{Age} = 75.30$ $\sigma_{Age} = 6.61$
Sex	categorical	female: 37 (37.76 %) male: 61 (62.24 %)	female: 64 (31.84 %) male: 137 (68.16 %)	female: 38 (39.58 %) male: 58 (60.42 %)
Years of education	numerical	$\mu_{Education} = 15.95$ $\sigma_{Education} = 3.02$	$\mu_{Education} = 15.76$ $\sigma_{Education} = 2.87$	$\mu_{Education} = 14.61$ $\sigma_{Education} = 3.20$
Race	categorical	white: 90 (91.84 %) black: 7 (7.14 %) asian: 1 (1.02 %)	white: 187 (93.03 %) black: 10 (4.98 %) asian: 4 (1.99 %)	white: 89 (92.71 %) black: 5 (5.21 %) asian: 2 (2.08 %)
Marital status	categorical	never married: 6 (6.12 %) married: 71 (72.45 %) divorced: 8 (8.16 %) widowed: 13 (13.27 %)	never married: 3 (1.49 %) married: 151 (75.12 %) divorced: 18 (8.96 %) widowed: (14.43 %)	never married: 3 (3.13 %) married: 83 (86.46 %) divorced: 4 (4.17 %) widowed: 6 (6.25 %)
Number of ApoE4 alleles	categorical	0: 73 (74.49 %) 1: 23 (23.47 %) 2: 2 (2.04 %)	0: 94 (46.77 %) 1: 81 (40.30 %) 2: 26 (12.94 %)	0: 33 (34.38 %) 1: 48 (50.00 %) 2: 15 (15.63 %)
FDG value	numerical	$\mu_{FDG} = 6.09$ $\sigma_{FGD} = 0.76$	$\mu_{FDG} = 5.85$ $\sigma_{FDG} = 0.76$	$\mu_{FDG} = 6.06$ $\sigma_{FDG} = 0.64$
MMSE-Score	categorical	none: 90 (91.84 %) ($28 \leq MMSE \leq 30$) mild: 8 (8.16 %) ($25 \leq MMSE \leq 27$) moderate: 0 (0.00 %) ($20 \leq MMSE \leq 24$) severe: 0(0.00 %) ($MMSE < 20$)	none: 92 (45.77 %) mild: 93 (46.27 %) moderate: 16 (9.96 %) severe: 0 (0.00 %)	none: 0 (0.00 %) mild: 39 (40.63 %) moderate: 56 (58.33 %) severe: 1 (1.04 %)

MMSE: The Mini Mental State Examination (also known as Folstein test is a 30-point neuropsychological questionnaire, used in clinical and research settings to measure general cognitive impairment [9].

with a parametrization of $MinPts = 4$ voxels and $\epsilon = 1$ voxel. We only keep robust clusters that are detected across all folds. Figure 4a shows ten robust clusters detected in MRI data for the setting AD vs. HC. The two identified clusters of the PET data are illustrated in Figure 4b, respectively. Single informative voxels, which distinguish AD patients from HC are spread all over the brain (162,532 voxels in MRI and 110,117 voxels in PET). To interpret the detected clusters, we map them to real brain regions according their anatomical location information using the Talairach Daemon software available at `http://www.talairach.org`. This mapping is presented in Table 2.

Only a few features (49 voxels in MRI and 675 voxels in PET) classify HC from MCI. For AD vs. MCI, 64,265 voxels in MRI and 37 voxels in PET have an IG value above IG_{opt}. Consequently, IDEA did not detect any informative neuroimaging clusters for AD vs. MCI and HC vs. MCI.

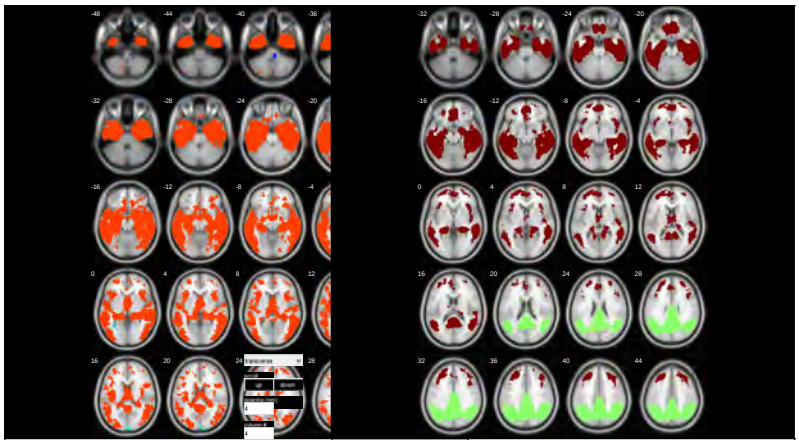
Finally, Table 3 summarizes the IG values of each attribute for the non-image data (cf. Table 1). For further processing, IDEA selects all attributes with an IG value higher than zero.

Table 3: IG values for each attribute of the non-image data for the settings AD vs. HC, AD vs. MCI and HC vs. MCI, respectively.

	AD vs. HC	AD vs. MCI	HC vs. MCI
Age	0.06	0.00	0.05
Sex	0.00	0.00	0.00
Years of education	0.00	0.00	0.00
Race	0.00	0.00	0.00
Marital status	0.00	0.00	0.00
Number of ApoE4 alleles	0.12	0.00	0.15
FDG value	0.41	0.15	0.07
MMSE-Score	0.83	0.49	0.20

4.3 Dependencies among Features

For all features identified in the aforementioned section and each experimental setting, we calculate the pairwise MI and visualize it using MDS, as described in Sections 3.3 and 3.4. Figure 5a shows the MI-matrix of informative attributes



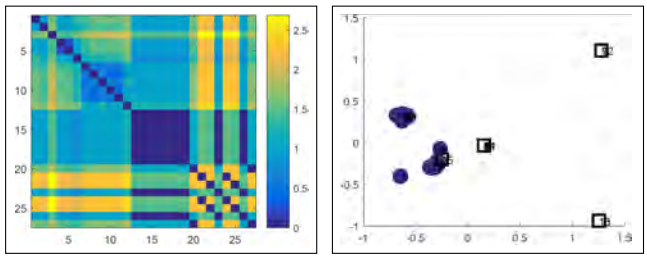
(a) MRI: AD vs. HC **(b)** PET: AD vs. HC

Figure 4: Selected informative clusters appearing in all folds of MRI and PET data for the AD vs. HC study.

Cluster ID	Cluster size	Brain region
MRI data		
1	4	Left Cerebellar Tonsil
2	4	Left Cingulate Gyrus
3	4	Right Precuneus
4	36	Right Medial Frontal Gyrus
5	51	Right Precentral Gyrus
6	66	Left Parahippocampal Gyrus
7	236	Right Cerebellar Tonsil
8	355	Right Superior Parietal Lobule
9	576	Left Cuneus
10	161,200	Right Middle Temporal Gyrus
PET data		
1	37,197	Left Precuneus
2	72,920	Left Middle Temporal Gyrus

Table 2: Mapping of detected clusters in the neuroimaging data to real brain regions using the Talairach Daemon software for the study AD vs. HC.

sourcing from neuroimaging scans and non-image data for the class of MCI patients in one fold. Figure 5b illustrates the corresponding dependencies by MDS. The depicted distance of two objects in this plot, directly correlates with their joined degree of information. Hence, it is obvious that some features provide redundant information. By partitioning clustering, IDEA determines *different* kind of information. To represent the discriminatory attributes, shown in Figure 5, only five features (one of each cluster) are adequate to achieve strong classification results (cf. Section 4.4).



(a) MI-matrix of merged attributes sourcing from neuroimaging and non-imaging data.

(b) 2D-representation by MDS. Circles indicate neuroimaging attributes, squares formalize non-image features.

Figure 5: Calculation and illustration of dependencies among features for the group of MCI patients.

4.4 Classification Results

For classification, we use the WEKA implementation (available at `http://www.cs.waikato.ac.nz/ml/weka`) of the Support Vector Machine (SVM) with polynomial kernel. For each classification result, we report accuracy (acc), sensitivity (sen) and specificity (spec). Table 4 presents the results on neuroimaging data w.r.t. using all voxels of the detected clusters versus the mean value of the underlying voxels of each cluster.

The next experiments document the benefit of an integrative classification procedure as performed by IDEA. Again, we distinguish between image cluster representations by all voxels or mean values. The classification results described by accuracy, sensitivity and specificity are represented in Table 5. The accuracy of AD vs. HC of MRI and PET image data combined with the non-image attributes is approximately the same due to the number of features of image data dominate the number of non-image attributes. However, when combining mean value of clusters with the informative non-image features, the classification results are improved above 90%.

As stated in the aforementioned section, IDEA automatically provides small feature sets that achieve accurate classification results. For this experiment, we evaluate the classification results on a set of features that was built by partitioning clustering with $k = 5$ and an IG-driven feature selection for each cluster on the data illustrated in Figure 5b. The corresponding results are presented in Table 6. Compared with Table 5, where we were using all available attributes, selecting the right set of (few) features yield to similar classification accuracies.

5 Conclusion

With IDEA, we presented a data mining framework for Integrative Detection of Early-stage Alzheimer's disease based on multimodal neuroimaging and heterogeneous non-image data types. The combination of information gain, mutual information, multidimensional scaling and clustering enables us to find feature combinations that have a high potential to predict Alzheimer's at an early stage. In near future, we per-

Table 4: Classification results on neuroimaging data using all voxels of a cluster vs. using the mean value of the voxels to represent a cluster.

	AD vs. HC	AD vs. MCI	HC vs. MCI
MRI data			
acc (all)	0.8029	0.6959	0.6556
acc (mean)	0.7458	0.7095	0.6723
sen (all)	0.8067	0.3700	0.1433
sen (mean)	0.69221	0.1656	0
spec (all)	0.7978	0.8510	0. 9055
spec (mean)	0.7978	0.9650	1
PET data			
acc (all)	0.8763	0.7128	0.6956
acc (mean)	0.7513	0.7024	0.6723
sen (all)	0.8422	0.3478	0.2900
sen (mean)	0.7378	0.2089	0
spec (all)	0.9100	0.8857	0.8900
spec (mean)	0.7600	0.9355	1

Table 5: Classification results on neuroimaging data in combination with non-image data using all voxels of a cluster vs. using the mean value of the voxels to represent a cluster.

	AD vs. HC	AD vs. MCI	HC vs. MCI
MRI data			
acc (all)	0.8789	0.9187	0.9433
acc (mean)	1	0.9292	0.9667
sen (all)	0.8667	0.8289	0.9600
sen (mean)	1	0.8189	0.9900
spec (all)	0.8900	0.9600	0.9352
spec (mean)	1	0.9800	0.9555
PET data			
acc (all)	0.8866	0.9359	0.9199
acc (mean)	0.9950	0.9392	0.9532
sen (all)	0.8633	0.8189	0.9089
sen (mean)	0.9900	0.8300	0.9589
spec (all)	0.91	0.9900	0.9255
spec (mean)	1	0.9900	0.9505

form a big data study on a second data set contributed by partners of the Institute for Stroke and Dementia Research (ISD), University of Munich based on a compact data representation. Here, we again expect new insight to the development and diagnosis of a disease that causes problems with memory, thinking and behavior for a multitude of elderly people. Furthermore, we currently work on a user-optimized graphical presentation based on scatter-plots that enable the medical scientists to rate the individual risk profile of a particular subject.

Table 6: Classification results after feature selection on neuroimaging data using average voxels of a cluster in combination with non-image data

	AD vs. HC	AD vs. MCI	HC vs. MCI
MRI data			
acc	0.8497	0.9121	0.9067
sen	0.8622	0.7433	0.8800
spec	0.8389	0.9900	0.9207
PET data			
acc	0.7971	0.9160	0.8700
sen	0.7867	0.7778	0.8400
spec	0.8067	0.9800	0.8800

References

[1] Ahmad, A., Dey, L.: A k-mean clustering algorithm for mixed numeric and categorical data. Data & Knowledge Engineering 63(2), 503–527 (2007)

[2] Böhm, C., Goebl, S., Oswald, A., Plant, C., Plavinski, M., Wackersreuther, B.: Integrative Parameter-Free Clustering of Data with Mixed Type Attributes. In: Advances in Knowledge Discovery and Data Mining, 14th Pacific-Asia Conference, PAKDD 2010, Hyderabad, India, June 21-24, 2010. Proceedings. Part I. pp. 38–47 (2010)

[3] Böhm, C., Oswald, A., Plant, C., Wackersreuther, B.: A Data Mining Framework for Classification of High-resolution Magnetic Resonance Images. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2008), Workshop on Mining Medical Data, Las Vegas, USA (2008)

[4] Borg, I., Groenen, P.J.: Modern Multidimensional Scaling: Theory and Applications. Springer (2005)

[5] Dyrba, M., Ewers, M., Wegrzyn, M., Kilimann, I., Plant, C., Oswald, A., Meindl, T., Pievani, M., Bokde, A., Fellgiebel, A., Filippi, M., Hampel, H., Kloppel, S., Hauenstein, K., Kirste, T., Teipel, S., the EDSD study group: Robust automated detection of microstructural white matter degeneration in Alzheimer's disease using machine learning classification of multicenter DTI data. PLOS ONE 8(5), 1–14 (2013)

[6] Ester, M., Kriegel, H., Sander, J., Xu, X.: A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), Portland, Oregon, USA. pp. 226–231 (1996)

[7] Fan, Y., Batmanghelich, N., Clark, C., Davatzikos, C., ADNI: Spatial patterns of brain atrophy in MCI patients, identified via high-dimensional pattern classification, predict subsequent cognitive decline. NeuroImage 39(4), 1731–1743 (2008)

[8] Fayyad, U.M., Irani, K.B.: Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning. In: Proceedings of the 13th International Joint Conference on Artificial Intelligence. Chambéry, France, August 28 - September 3, 1993. pp. 1022–1029 (1993)

[9] Folstein, M.F., Folstein, S.E., McHugh, P.R.: Mini-Mental State (a practical method for grading the state of patients for the clinician). Journal of Psychiatric Research 12(3), 189–198

(1975)
[10] Hall, M.A., Holmes, G.: Benchmarking Attribute Selection Techniques for Discrete Class Data Mining. IEEE Transactions on Knowledge and Data Engineering 15(6), 1437–1447 (2003)
[11] Hartigan, J.A.: Clustering Algorithms. John Wiley & Sons, Inc., New York, NY, USA, 99th edn. (1975)
[12] Hsu, C., Chen, Y.: Mining of mixed data with application to catalog marketing. Expert Systems with Applications 32(1), 12–23 (2007)
[13] Huang, Z.: Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. Data Mining and Knowledge Discovery 2(3), 283–304 (1998)
[14] Jack, C.R., Knopman, D.S., Jagust, W.J., Shaw, L.M., Aisen, P.S., Weiner, M.W., Petersen, R.C., Trojanowski, J.Q.: Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. The Lancet Neurology 9(1), 119–128 (2010)
[15] Kloppel, S., Stonnington, C., Chu, C., Draganski, B., Scahill, R., Rohrer, J., Fox, N., Jack, C., Ashburner, J., Frackowiak, R.: Automatic classification of MR scans in Alzheimer's disease. Brain 131(3), 681–689 (2008)
[16] McKhann, G., Drachman, D., Folstein, M., Katzman, R., Price, D., Stadlan, E.M.: Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. Neurology 34(7), 939–944 (1984)
[17] Peng, H., Long, F., Ding, C.H.Q.: Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(8), 1226–1238 (2005)
[18] Petersen, R.C., Doody, R., Kurz, A., Mohs, R.C., Morris, J.C., Rabins, P.V., Ritchie, K., Rossor, M., Thal, L., Winblad, B.: Current Concepts in Mild Cognitive Impairment. Archives of neurology 85(12), 1985–1992 (2001)
[19] Piatetsky-Shapiro, G., Djeraba, C., Getoor, L., Grossman, R., Feldman, R., Zaki, M.J.: What are the grand challenges for data mining?: KDD-2006 panel report. SIGKDD Explorations 8(2), 70–77 (2006)
[20] Plant, C., Böhm, C.: INCONCO: INterpretable Clustering Of Numerical and Categorical Objects. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, August 21-24, 2011. pp. 1127–1135 (2011)
[21] Plant, C., Teipel, S.J., Oswald, A., Böhm, C., Meindl, T., Mourao-Miranda, J., Bokde, A.W., Hampel, H., Ewers, M.: Automated detection of brain atrophy patterns based on MRI for the prediction of Alzheimer's disease. NeuroImage 50(1), 162–174 (2010)
[22] Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann (1993)
[23] Vapnik, V.N., Chervonenkis, A.Y.: Theory of Pattern Recognition [in Russian]. Nauka (1974)
[24] Xiang, S., Yuan, L., Fan, W., Wang, Y., Thompson, P.M., Ye, J.: Multi-source learning with block-wise missing data for Alzheimer's disease prediction. In: The 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2013, Chicago, IL, USA, August 11-14, 2013. pp. 185–193 (2013)
[25] Yang, Q., Wu, X.: 10 Challenging Problems in Data Mining Research. International Journal of Information Technology & Decision Making 5(4), 597–604 (2006)
[26] Yin, J., Tan, Z.: Clustering Mixed Type Attributes in Large Dataset. In: Parallel and Distributed Processing and Applications, Third International Symposium, ISPA 2005, Nanjing, China, November 2-5, 2005, Proceedings. pp. 655–661 (2005)

Identifying Significance of Discrepancies in Radiology Reports

Arman Cohan* Luca Soldaini* Nazli Goharian*
Allan Fong† Ross Filice‡ Raj Ratwani†

Abstract

At many teaching hospitals, it is common practice for on-call radiology residents to interpret radiology examinations; such reports are later reviewed and revised by an attending physician before being used for any decision making. In case there are substantial problems in the resident's initial report, the resident is called and the problems are reviewed to prevent similar future reporting errors. However, due to the large volume of reports produced, attending physicians rarely discuss the problems side by side with residents, thus missing an educational opportunity. In this work, we introduce a pipeline to discriminate between reports with *significant discrepancies* and those with *non-significant discrepancies*. The former contain severe errors or mis-interpretations, thus representing a great learning opportunity for the resident; the latter presents only minor differences (often stylistic) and have a minor role in the education of a resident. By discriminating between the two, the proposed system could flag those reports that an attending radiology should definitely review with residents under their supervision. We evaluated our approach on 350 manually annotated radiology reports sampled from a collection of tens of thousands. The proposed classifier achieves an Area Under the Curve (AUC) of 0.837, which represent a 14% improvement over the baselines. Furthermore, the classifier reduces the False Negative Rate (FNR) by 52%, a desirable performance metric for any recall-oriented task such as the one studied in this work.

1 Introduction

A key aspect of the education of resident radiologists is the development of the necessary skills to interpret radiology examinations and report their findings. Reports are later examined by an experienced attending physician, who revises eventual interpretation errors or minor mistakes. In case the attending performs substantial edits to the report, we say that *significant discrepancies* exist between the initial and the revised report. These discrepancies are due to potential erroneous image interpretation of the resident. Prevention of such errors is essential to the education of the radiology residents as well as the patient care. On the other hand, if a report has been edited to only address minor errors or style issues, we say that *non-significant discrepancies* exists. In Figure 1, examples of significant and non-significant discrepancies are shown (each example is a small section of a much longer report).

Researchers have studied the frequency of discrepancies in radiology reports [28, 24], as well as their impact on resident learning and patient care [23]. Moreover, recent studies have also determined that residents produce less reports that need to be significantly edited by attending radiologists as their experience increase [9].

The large volume of radiology reports generated each day makes manual surveillance challenging; thus, in recent years, systems to identify reports that have major discrepancies have been introduced. Sharpe, et al. [25] proposed an interactive dashboard that highlights the differences between reports written by residents alongside the version edited by attending radiologists. Kalaria and Filice [11] used the number of words differing between the preliminary and final report to measure the significance of the discrepancies. However, deviation detected using this measure does not fully capture the difference between reports with significant discrepancies and non-significant ones, as dissimilarities in the writing styles between residents and attending radiologists can also cause differences in word counts.

We propose an accurate and effective two-stage pipeline to distinguish between significant and non-significant discrepancies in radiology reports. In other words, given a set of preliminary radiology reports with the respective final reports, we identify those with significant discrepancies. The first stage of our pipeline

*Information Retrieval Lab, Computer Science Department, Georgetown University
†National Center for Human Factors in Healthcare, MedStar Health
‡Department of Radiology, MedStar Georgetown University Hospital

	Significant discrepancies	Non-significant discrepancy
Preliminary report (resident radiologist)	*"No acute hemorrhage. No extra-axial fluid collections. ~~The differentiation of gray and white matter is normal~~."*	*"Postsurgical changes related to right thoracotomy with surgical packing material and hemorrhagic blood products in the right lower chest."*
Final report (attending radiologist)	*"Subtle hypodensities in the inferolateral left frontal lobe and anterolateral left temporal lobe likely represent acute cortical contusions. No acute hemorrhage. No extra-axial fluid collections. Small area of encephalomalacia in the right parietal lobe."*	*"Postsurgical changes related to right thoracotomy with surgical packing material and large amount of hemorrhagic blood products in the right lower chest."*

Figure 1: Example of significant and non-significant discrepancies between reports. The stroked-through text has been removed from the preliminary report by the attending radiologist, while the underlined sections have been added.

employs an ontology of radiology terms and expressions to identify reports with no significant differences. The remaining reports are then separated by a Support Vector Machine (SVM) classifier. We evaluate the impact of a diverse set of textual, statistical, and assessment score features on the performance of the second-stage classifier. Some of these features have been previously used to assess the quality of the text summarization and machine translation systems. Results illustrate significant improvement over the baseline (up to +14.6% AUC, -52% FNR) and show the effectiveness of the proposed approach. Our focus on false negative rate is motivated by the fact that each missed significant discrepancy is a missed opportunity to educate a resident about a significant error in interpreting an examination.

To summarize, the main contributions of this work are as follows:

- We introduce an approach for automatically classifying the type of discrepancies between preliminary and final radiology reports.
- We explore the use of summarization and machine translation evaluation metrics as features identifying reports with significant discrepancies.
- We provide extensive evaluation of different aspects of the proposed pipeline.

2 Related Works

A related–yet ultimately different–problem to the one studied in this paper is the classification of radiology reports based on their content. In this task, which falls under the text classification domain, the goal is to classify radiology reports into a discrete set of predefined categories. For example, Nguyen and Patrick [19] aimed at grouping radiology reports into cancerous or non-cancerous cases using an SVM. Chapman, et al. [4] presented a system for detecting reports with mediastinal findings associated with inhalational anthrax. Percha, et al. [21] classified reports by breast tissue decomposition using a rule based classification scheme. Johnson, et al. [10] proposed a hybrid approach that combines rules with SVM to classify radiology reports with respect to their findings. Bath, et al. [3] introduced a classifier to determine the appropriate radiology protocol among those available for each disease. Their semi-supervised system takes advantage of the UMLS[1] ontology.

Researchers have also proposed methods for quantifying or comparing the quality of text in various domains. For example, Louis and Nenkova [15] introduced a model for classifying sentences in news articles into general/specific depending on the level of the information carried by each sentence. Their classifier uses word, syntax, and language modeling features. Feng, et al. [7] explored a range of text features such as discourse properties, language modeling features, part-of-speech-based features, and syntactic features to quantify text complexity. Zeng-Treitler, et al. [29] proposed a system to grade the readability of health content; their tool employs lexical, syntactic, semantic and stylistic characteristics to accomplish such goal. Ashok, et al. [2] proposed an SVM classifier based on part of speech and lexical distributions, sentiment features, and grammatical properties to predict the success of novels. Lastly, Louise and Nenkova [16] proposed a model for predicting the

appropriate length for a textual content in response to a specific information need.

Another line of related work is detecting plagiarism; systems designed for such task are concerned with determining if a given document was plagiarized from another source. To do so, current approaches in literature attempt to capture the significance of differences between a suspicious text and a source document (e.g., [1, 22, 27]). Most of the previous efforts in plagiarism detection are centered on the retrieval aspect to find the original source of plagiarized content; thus, they focus on information and passage retrieval. Our problem differs from plagiarism detection in that our system takes as input a a candidate-source pair (preliminary and final reports) and attempts at classifying the significance of differences between them; instead, in plagiarism detection, the goal is the retrieval of source document.

3 Methodology

We propose a two stage pipeline for classification of type of discrepancies in radiology reports based on their significance. The overview of our approach is shown in Figure 2. In first stage, we utilize a heuristic based on domain ontology to identify non-significant discrepancies. In next stage, reports that are labeled as significant by the heuristic are processed by a classifier that exploits a variety of textual features. Specifically, we adapt features that are originally used to evaluate text summarization and machine translation systems to our problem. The following sections provide details about each one of these two stages.

3.1 Stage 1: Domain ontology. We first link the significance of the discrepancies to the differences between the domain specific concepts in the reports. To extract domain specific concepts, we use RadLex[1], which is a comprehensive ontology of radiology terms and expressions with about 68K entries.

The domain specific concepts between the preliminary report and the final report are then compared. There might be cases in which there are no difference between the concepts of radiology reports but in one report some concepts are negated. As an example, consider these two sentences: " ... *hypodensities in the inferolateral left frontal lobe ...*" and "*... no hypodensity in the inferolateral left frontal lobe ...*". Although the radiology concepts are identical, the negation might indicate significant discrepancy. Therefore, we also consider the negations in which the RadLex concepts appear to prevent false classification. To detect negations, we use the dependency parse tree of the sentences and a set of seed negation words (*not* and *no*). That is, we mark a radiology concept as negated if these seed words are dependent on the concept. If the RadLex concepts of the reports are identical and the negations are consistent, we classify the type of changes as non-significant. We call this stage, the RadLex heuristic (As indicated in Figure 2). A more comprehensive negation detection algorithm (*NeGex* [5]) was also evaluated; however, its results did not show any significant improvement.

The RadLex heuristic highly correlates with human judgments in identifying non-significant changes, as shown in Section 4.2. However, this simple heuristic is not accurate for detecting the significant discrepancies. In other words, if RadLex terms or their associated negations are not consistent, one can not necessarily classify the report as significant.

3.2 Stage 2: Classification using textual features. In this section, we detail a binary classifier designed to address the shortcoming of the RadLex heuristic, we propose a binary classifier. The classifier uses diverse sets of textual features that aim to capture significance of discrepancies in radiology reports. The features that we use include surface textual features, summarization evaluation metrics, machine translation evaluation metrics, and readability assessment scores. We briefly explain each of these feature sets and provide the intuition behind each one of them.

3.2.1 Surface Textual Features. Previous work used word count discrepancy as a measure for quantifying the differences between preliminary and final radiology reports [11]. We use an improved version of the aforementioned method as one of the baselines. That is, in addition to the word count differences, we also consider the character and sentence differences between the two reports as an indicator of significance of changes.

3.2.2 Summarization evaluation features. ROUGE[1] [14], one of the most widely used set of metrics in summarization evaluation, estimates the quality of a system generated summary by comparing it to a set of human generated summaries. ROUGE has been proposed as an alternative to manual evaluation of the quality of system generated summaries which can be a long and exhausting process. Rather than using ROUGE as evaluation metric, we exploit it as a feature for comparing the quality of the preliminary radiology report with respect to the final report. Higher ROUGE scores indicate that the discrepancies between the preliminary and the final reports are less significant.

[1] https://www.nlm.nih.gov/research/umls/
[1] http://www.rsna.org/radlex.aspx

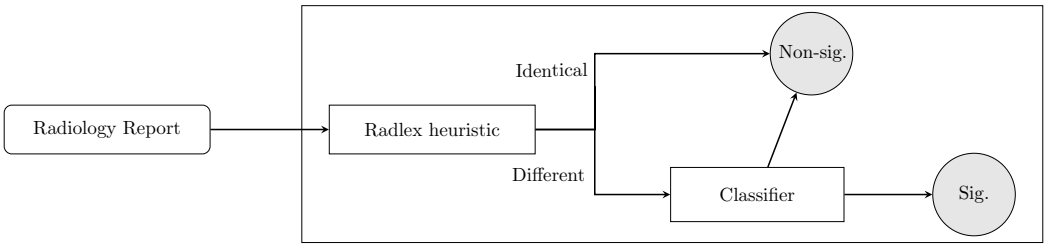


Figure 2: Overview of the proposed approach. The radiology reports are first classified by the Radlex heuristic. If there is no Radlex difference between a preliminary and the associated final report, the case is classified as non-significant discrepancy (*Non-sig* in the figure). Otherwise the case is sent to the a binary classifier for further analysis. The classifier which works based on several textual features, classifies the reports as having either significant (*Sig.* in the figure) or non-significant discrepancies

We utilize the following variants of ROUGE:

ROUGE-N: In our setting, ROUGE-N is the N-gram precision and recall between the preliminary and final report, where N is the gram length (e.g., N=1 indicates a single term, N=2 a word bigram, and so on.) We consider ROUGE-1 to ROUGE-4.

ROUGE-L: This metric compares the two reports based on the Longest Common Subsequence (LCS). Intuitively, longer LCS between the preliminary and the final report shows that the quality of the two reports are closer and therefore differences between the two are less significant.

ROUGE-S: ROUGE-S computes the skip-bigram co-occurrence statistics between the two reports. It is similar to ROUGE-2 except that it allows gaps between the bigrams. Skip-grams are used in different NLP application; they consider additional n-grams by skipping middle tokens. Applying skip-bigrams without any threshold on the distance between tokens often results in incorrect matches (e.g. we do not want to consider all "the the" skip-bigrams in a sentence with multiple "the" expressions). To prevent this, we limit the maximum allowed distance to 10.

3.2.3 Machine translation evaluation features. The Machine Translation (MT) evaluation metrics quantify the quality of a system-generated translation against a given set of reference or gold translations. We consider the final report as the reference and evaluate the quality of the preliminary report with respect to it. Higher scores indicate a better quality of the preliminary report, showing that the discrepancies between the preliminary and final versions are less significant. In detail, we use the following MT metrics: BLEU [20], Word Error Rate and METEOR [6].

BLEU (Bi-Lingual Evaluation Understudy): In our setting, BLEU is an n-gram based comparison metric for evaluating the quality of a candidate translation with respect to several reference translations. It is conceptually similar to ROUGE-N, except being precision-oriented. Specifically, BLEU combines a modified n-gram-based precision and a so-called "Brevity Penalty" (BP), which penalizes short sentences with respect to the reference. Here, we use the BLEU score of the preliminary report with respect to the final report as a feature that indicates the quality of the preliminary report.

Word Error Rate (WER): WER is another commonly used metric for the evaluation of machine translation [26]. It is based on the minimum edit distance between the words of a candidate translation versus reference translations; we consider WER as the following formula:

$$\text{WER} \stackrel{def}{=} (100 \times (S + I + D)/N)$$

where N is the total number of words in the preliminary report; S, I, and D are the number of Substitutions, Insertions, and Deletions made to the preliminary report to yield the final report.

Metric for Evaluation of Translation with Explicit word Ordering (METEOR): METEOR is a metric for evaluation of machine translation that aligns the translations to the references. Here, we want to find the best alignment between the preliminary report and the final report. In addition to exact matches between terms, METEOR also accounts for synonyms and paraphrase matches between the words and sentences which are not captured by previous features such as

[1]Recall-Oriented Understudy for Gisting Evaluation

		RadLex	A	B
non-significant	RadLex	1.0	0.964	0.942
	A	0.964	1.0	0.906
	B	0.942	0.906	1.0
count=139	Fleiss κ = 0.880			
significant	RadLex	1.0	0.557	0.492
	A	0.557	1.0	0.934
	B	0.492	0.934	1.0
count=61	Fleiss κ = 0.468			

Table 1: Agreement rate between the RadLex heuristic and two annotators A and B. Agreement for significant and non-significant reports are separately presented. Both raw agreement rates as well as Fleiss κ between the annotators and the RadLex heuristic are shown.

Baselines	F-1	FNR	AUC	ACC
Sf (Improved v. of [11])	0.650	0.329	0.642	0.633
RL	0.690	0.355	**0.746**	**0.707**
Sf+RL	**0.694**	**0.329**	0.730	0.700
Our methods	F-1	FNR	AUC	ACC
Rd	0.568	0.421	0.594	0.553
BL	0.709	0.184*	0.757	0.660
M	0.604	0.368	0.627	0.580
Rg	0.767*	0.197*	0.838*	0.753*
Rg+BL	0.739*	0.237*	0.831*	0.727*
Rg+M	0.775*	0.184*	0.847*	0.760*
Rg+WER	0.702	0.211*	0.746	0.660
Rg+BL+M	0.780*	0.184*	**0.843***	0.767*
Rg+BL+M+RL	0.769*	0.211*	0.841*	0.760*
Rg+BL+M+RL+Rd	**0.797***	**0.171***	0.837*	**0.787***

Table 2: F-1 score (F1) and False Negative Rate (FNR) for significant reports as well as overall Area Under the Curver (AUC) and Accuracy (ACC) based on different set of features. The top part of the table shows the baselines and the bottom part shows our proposed features. Sf: Surface features – character, word and sentence differences; RL: RadLex concepts and their associated negation differences; Rd: Readability features; M: METEOR; BL: BLEU. Rg: ROUGE. Asterisk (*) shows statistically significant improvement over all baselines (two-tailed student t-test, $p < 0.05$).

ROUGE. We use both the WordNet [18] synonyms and RadLex ontology synonyms for calculation of the METEOR score.

3.3 Readability assessment features. To quantify complexity of textual content and the style of the reports, we use readability assessment features. Here, "style" refers to reporting style of the radiology reports, such as lexical and syntactic properties. In detail, we use the Automated Readability Index (ARI) [12] and the Simple Measure Of Gobbledygook (SMOG) index [17]. These two metrics are based on distributional features such as the average number of syllables per word, the number of words per sentence, or binned word frequencies. In addition to these statistics, we also consider average phrase counts (noun, verb and propositional phrases) among the features.

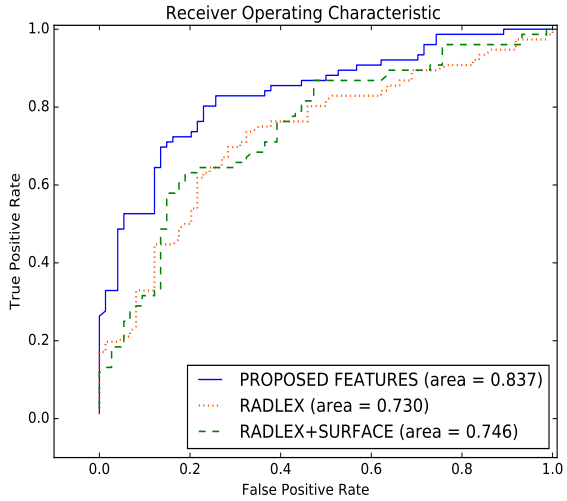
4 Empirical Results

4.1 Experimental setup We use a collection of radiology reports with discrepancies obtained from a large urban hospital for evaluation. These reports contain two main textual sections: *findings*, which contains the full interpretation of the radiology examination, and *impression*, which is a concise section that highlights important aspects of the report. We use both sections for evaluation of our proposed pipeline. We use 10 fold cross validation for evaluating the proposed classification scheme.

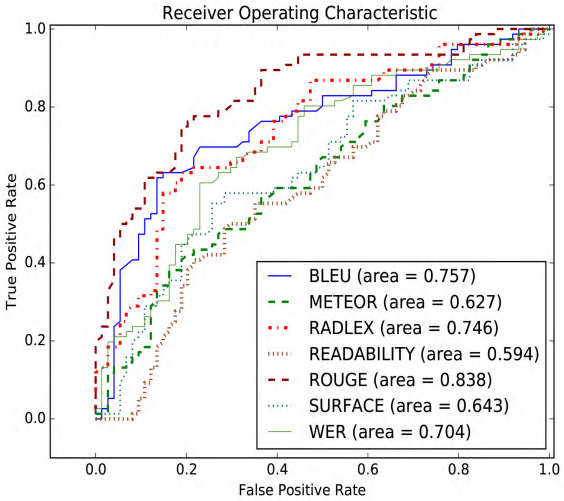
4.2 Classification using RadLex ontology. As explained in Section 3, we first classify the reports using the RadLex ontology and the negation differences between the preliminary and final versions of the report. We ran this method on 200 randomly sampled reports from the dataset; two annotators were asked to label the reports based on significance of discrepancies. The annotators were allowed to label a case as "not-sure" if they could not confidently assign a label for the report. The agreement rates between the annotators and the RadLex heuristic is shown in Table 1. As illustrated, RadLex heuristic is highly correlated with human judgments and the Fleiss κ for non-significant reports is above 0.8, which can be interpreted as perfect agreement [13, 8]. However, the simple RadLex heuristic's performance for the reports that it labels as significant is low. Thus, we conclude that RadLex concept differences between the reports do not necessarily mean that the changes between them is significant. As we show in next section, the proposed classification scheme with the textual features can solve this problem for reports with RadLex differences.

4.3 Classification by textual features. To evaluate our proposed classification approach, a radiologist manually identified types of discrepancies of 150 randomly sampled radiology reports that include RadLex concept differences.

4.3.1 Feature analysis. Table 2 shows the cross validated classification results using the set of features described in Section 3. We use an SVM classifier with linear kernel. We report F-1 score and false negative rates for significant reports, and the overall



(a) Comparison of the proposed pipeline with the baselines



(b) Comparison of individual features.

Figure 3: ROC curves

area under the curve and accuracy. We consider the following baselines: (i) Surface textual features including character, word and sentence differences between the reports (Indicated as "Sf" in the table). (ii) RadLex concepts and associated negation differences (Indicated as "RL"). (iii) Surface textual features along with RadLex concepts and negation differences (RL+Sf). Results based on different sets of features are presented. We experimented with all possible combinations of features; for the sake of brevity, we only report combination of features of significance.

We observe that majority of the proposed features outperform the baseline significantly. One feature set performing worse than the baseline is the readability features. As described in Section 3.3, readability features mostly capture the differences between the reporting styles, as well as the readability of the written text. However, the reporting style and readability of the preliminary and final report might be similar although their content differs. For example, some important radiology concepts relating to a certain interpretation might be contradictory in the preliminary and final report while they both follow the same style. Thus, the readability features on their own are not able to capture significant discrepancies. However, when used with other features such as ROUGE, they are able to capture style differences that are not realized by other features especially in insignificant change category. This causes the performance of combined metrics to increase.

ROUGE features are able to significantly improve over the baseline. When we add METEOR features, we observe a further improvement over ROUGE alone. This is likely due to the fact that METEOR considers synonyms in aligning the sentences as well, which is not captured by ROUGE. However, we note that METEOR by itself underperforms the baseline. We attribute this to the concept drift that may have been caused by consideration of synonyms in METEOR as observed in high FNR of METEOR. The highest scores are achieved when we combine METEOR, ROUGE, BLEU, RadLex and readability features. We attribute the high performance of this setting to different aspects of reporting discrepancies captured by each of the features. ROC curve differences between our best performing features and the baseline (Figure 3a) further shows the effectiveness of our approach. Individual effects of features in terms of ROC curves are also compared in Figure 3b. As shown, ROUGE features are the most informative for identifying significant discrepancies.

4.3.2 Sections of the report. We evaluated which sections of the radiology report have more influence on the final significance of the discrepancies. As explained in Section 4.1, the reports have two main sections: *findings* and *impression*. As shown in table 3, *impression* section features have higher F-1 scores (+6.68%), lower false negative rates (-31.8%) and higher accuracy (+4.5%) than *findings* section. This is expected, since *impression* contains key points of the report. However, the best results are achieved when both sections are considered, thus indicating that the *findings* section contains valuable information that are not present in the *impression*.

Sections	F-1	FNR	AUC	ACC
Impression	0.772	0.197	0.821	0.760
Findings	0.725	0.289	0.817	0.727
All	0.797	0.171	0.837	0.787

Table 3: Comparison of the results based on features extracted from different sections of the reports.

4.4 Error Analysis. We examined the cases that our approach incorrectly classified. First, many of the false positive cases (i.e., reports that were incorrectly flagged as having significant discrepancies) were due to unnecessarily long length of preliminary reports. We saw that in many cases, the preliminary report, especially in *impression* section, contains extra information that is later removed by the attending editor. In these cases, when almost half of the preliminary report is removed in the final version, our classification scheme fails to classify them as insignificant. According to the domain expert annotator, however, those removed sections do not convey any critical information. Since our features are mostly considering lexical overlaps between the reports, they fail to capture these special cases.

Second, we noticed that some of the false negative cases were due to only slight changes between the two reports. An example is illustrated below which shows a snippet from the preliminary and the final reports:

- **preliminary report**: "*Worsening airspace disease at the left base represents aspiration.*"
- **final report** "*Worsening airspace disease at the left base could represent aspiration.*"

This small change in the report is interpreted as a significant discrepancy between the two reports by the domain expert. Since there is only a slight change between the two reports and the term *could* is not a domain specific term, our features fail to detect this case as significant. In this special case, the term *could* changes a specific interpretation from a definite fact to a possibility, thus can be considered as significant discrepancy.

Although the proposed approach misclassifies these cases, such discrepancies are very rare. In future work, we will focus on designing features that can capture significance of discrepancies in such cases.

5 Conclusions and future work

Identifying significance of discrepancies in radiology reports is essential for education of radiology residents and patient care. We proposed a two-stage pipeline to distinguish between significant and non-significant discrepancies in radiology reports. In the first stage we adopted a heuristic based on the RadLex domain ontology and negations in radiology narratives. In the second stage, we proposed a classifier based on several features including summarization and machine translation evaluation, and text readability features for classification of the reports. We validated our approach using a real world dataset obtained from a large urban hospital. We showed the effectiveness of our proposed pipeline which gains statistically significant improvement (+14.6% AUC, -52% FNR) over the several baselines. A provisional patent based on the proposed approach has been filed at United States Patent and Trademark Office (application number 62280883).

We only focused on the binary classification of changes into two categories: significant and non-significant. Future work will be concerned with exploring the problem of categorizing changes into multiple levels of significance.

Error analysis revealed some rare cases that our features are not designed to capture. Such cases are mostly due to either very small textual differences between the reports that imply significant discrepancy or huge textual differences that do not reflect any significant discrepancies. One natural extension is to design features that can capture such cases. For example, one can consider differences between modality of the reports.

An important goal in detecting significant discrepancies is to prevent future similar problems. One intuitive direction to follow would be clustering discrepancies based on certain textual descriptors. Thus, finding common problems in the collection of initial reports can further promote patient care and resident education.

Acknowledgments

The authors thank Amit Kalaria for helping in annotation and Ophir Frieder for discussions and comments.

References

[1] A. Abdi, N. Idris, R. M. Alguliyev, and R. M. Aliguliyev, *Pdlk: Plagiarism detection using linguistic knowledge*, Expert Systems with Applications, 42 (2015), pp. 8936–8946.

[2] V. G. Ashok, S. Feng, and Y. Choi, *Success with style: Using writing style to predict the success of novels*, Poetry, 580 (2013), p. 70.

[3] A. Bhat, G. Shih, and R. Zabih, *Automatic selection of radiological protocols using machine learning*, in Proceedings of the 2011 workshop on Data mining for medicine and healthcare, ACM, 2011, pp. 52–55.

[4] W. W. Chapman, G. F. Cooper, P. Hanbury, B. E. Chapman, L. H. Harrison, and M. M. Wagner, *Creating a text classifier to detect radiology reports describing mediastinal findings associated with inhalational anthrax and other disorders*, Journal of the American Medical Informatics Association, 10 (2003), pp. 494–503.

[5] W. W. Chapman, D. Hilert, S. Velupillai, M. Kvist, M. Skeppstedt, B. E. Chapman, M. Conway, M. Tharp, D. L. Mowery, and L. Deleger, *Extending the negex lexicon for multiple languages*, Studies in health technology and informatics, 192 (2013), p. 677.

[6] M. Denkowski and A. Lavie, *Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems*, in Proceedings of the Sixth Workshop on Statistical Machine Translation, Association for Computational Linguistics, 2011, pp. 85–91.

[7] L. Feng, M. Jansche, M. Huenerfauth, and N. Elhadad, *A comparison of features for automatic readability assessment*, in Proceedings of the 23rd International Conference on Computational Linguistics: Posters, Association for Computational Linguistics, 2010, pp. 276–284.

[8] A. M. Green, *Kappa statistics for multiple raters using categorical classifications*, in Proceedings of the 22nd annual SAS User Group International conference, vol. 2, 1997, p. 4.

[9] G. Issa, B. Taslakian, M. Itani, E. Hitti, N. Batley, M. Saliba, and F. El-Merhi, *The discrepancy rate between preliminary and official reports of emergency radiology studies: a performance indicator and quality improvement method*, Acta Radiologica, 56 (2015), pp. 598–604.

[10] E. Johnson, W. C. Baughman, and G. Ozsoyoglu, *Mixing domain rules with machine learning for radiology text classification*, (2014).

[11] A. D. Kalaria and R. W. Filice, *Comparison-bot: an automated preliminary-final report comparison system*, Journal of digital imaging, (2015), pp. 1–6.

[12] J. Kincaid, R. Fishburne, R. Rogers, and B. Chissom, *Derivation of new readability formulas*, tech. report, Technical report, TN: Naval Technical Training, US Naval Air Station, Memphis, TN, 1975.

[13] J. R. Landis and G. G. Koch, *The measurement of observer agreement for categorical data*, biometrics, (1977), pp. 159–174.

[14] C.-Y. Lin, *Rouge: A package for automatic evaluation of summaries*, in Text summarization branches out: Proceedings of the ACL-04 workshop, vol. 8, 2004.

[15] A. Louis and A. Nenkova, *Automatic identification of general and specific sentences by leveraging discourse annotations.*, in IJCNLP, 2011, pp. 605–613.

[16] A. Louis and A. Nenkova, *Verbose, laconic or just right: A simple computational model of content appropriateness under length constraints*, EACL 2014, (2014), p. 636.

[17] G. H. McLaughlin, *Smog grading: A new readability formula*, Journal of reading, 12 (1969), pp. 639–646.

[18] G. A. Miller, *WordNet: a lexical database for English*, Communications of the ACM, 38 (1995), pp. 39–41.

[19] D. H. Nguyen and J. D. Patrick, *Supervised machine learning and active learning in classification of radiology reports*, Journal of the American Medical Informatics Association, 21 (2014), pp. 893–901.

[20] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, *Bleu: a method for automatic evaluation of machine translation*, in Proceedings of the 40th annual meeting on association for computational linguistics, Association for Computational Linguistics, 2002, pp. 311–318.

[21] B. Percha, H. Nassif, J. Lipson, E. Burnside, and D. Rubin, *Automatic classification of mammography reports by bi-rads breast tissue composition class*, Journal of the American Medical Informatics Association, 19 (2012), pp. 913–916.

[22] M. Potthast, M. Hagen, M. Völske, and B. Stein, *Crowdsourcing interaction logs to understand text reuse from the web.*, in ACL (1), 2013, pp. 1212–1221.

[23] A. T. Ruutiainen, D. J. Durand, M. H. Scanlon, and J. N. Itri, *Increased error rates in preliminary reports issued by radiology residents working more than 10 consecutive hours overnight*, Academic radiology, 20 (2013), pp. 305–311.

[24] A. T. Ruutiainen, M. H. Scanlon, and J. N. Itri, *Identifying benchmarks for discrepancy rates in preliminary interpretations provided by radiology trainees at an academic institution*, Journal of the American College of Radiology, 8 (2011), pp. 644–648.

[25] R. E. Sharpe Jr, D. Surrey, R. J. Gorniak, L. Nazarian, V. M. Rao, and A. E. Flanders, *Radiology report comparator: a novel method to augment resident education*, Journal of digital imaging, 25 (2012), pp. 330–336.

[26] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, *A study of translation edit rate with targeted human annotation*, in Proceedings of association for machine translation in the Americas, 2006, pp. 223–231.

[27] E. Stamatatos, M. Potthast, F. Rangel, P. Rosso, and B. Stein, *Overview of the pan/clef 2015 evaluation lab*, in Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer, 2015, pp. 518–538.

[28] J. Walls, N. Hunter, P. M. Brasher, and S. G. Ho, *The depictors study: discrepancies in preliminary interpretation of ct scans between on-call residents and staff*, Emergency radiology, 16 (2009), pp. 303–308.

[29] Q. Zeng-Treitler, L. Ngo, S. Kandula, G. Rosemblat, H.-E. Kim, and B. Hill, *A method to estimate readability of health content*, Association for Computing Machinery, (2012).

Multi-task Spare Group Lasso for Characterizing Alzheimer's Disease*

Xiaoli Liu[†‡] Peng Cao[†] Dazhe Zhao[†] Arindam Banerjee[‡]

Abstract

Alzheimers disease (AD) is a severe neurodegenerative disorder characterized by loss of memory and reduction in cognitive functions due to progressive degeneration of neurons and their connections, eventually leading to death. AD is the most common cause of dementia in the elderly, and currently affects over 5 million individuals in the US, and over 30 million individuals worldwide. In this paper, we consider the problem of simultaneously predicting several different cognitive scores associated with categorizing subjects as normal, mild cognitive impairment (MCI), or Alzheimer's disease (AD) in a multi-task learning framework using features extracted from brain images obtained from ADNI (Alzheimer's Disease Neuroimaging Initiative). To solve the problem, we present multi-task sparse group lasso (MT-SGL), which does hierarchical sparse feature selection in a multi-task setting, and propose a FISTA-style composite objective accelerated descent method for efficiently learning the model. Through comparisons on a variety of baseline models using multiple evaluation metrics, we illustrate the promising performance of MT-SGL on real imaging data drawn from ADNI.

Keywords: Alzheimer's disease, multi-task learning, sparse group lasso, optimization

1 Introduction

Alzheimer's Disease (AD) is a severe neurodegenerative disorder that results in a loss of mental function due to the deterioration of brain tissue, leading directly to death [10]. It accounts for 60−70% of age related dementia, affecting an estimated 30 million individuals in 2011 and the number is projected to be over 114 million by 2050 [27]. The cause of AD is poorly understood and currently there is no cure for AD. AD has a long preclinical phase, lasting a decade or more. There is increasing research emphasis on detecting AD in the pre-clinical phase, before the onset of the irreversible neuron loss that characterizes the dementia phase of the disease, since therapies/treatment are most likely to be effective in this early phase.

The Alzheimer's Disease Neuroimaging Initiative (ADNI, `http://adni.loni.usc.edu/`) has been facilitating the scientific evaluation of neuroimaging data including magnetic resonance imaging (MRI), positron emission tomography (PET), along with other biomarkers, and clinical and neuropsychological assessments for predicting the onset and progression of MCI (mild cognitive impairment) and AD. Early diagnosis of AD is key to the development, assessment, and monitoring of new treatments for AD. Using datasets from ADNI, predictive models have been developed in recent years to better characterize AD, including classifying a patient's data into categories such as normal, mild cognitive impairment (MCI), and Alzheimer's disease (AD). Accurate classification, especially for early stage MCI in patients, will be key to research and develop cures for AD. From a data analysis perspective, the problem is often posed as a multi-task learning (MTL) problem with a set of tasks derived from salient properties of the different categories of patients.

Existing works in this context has considered sparse models, for example Lasso [23, 14], which does feature selection from high-dimensional imaging data, where the features correspond to specific properties of different regions of the brain. Group Lasso methods, which take group information of the features into account while doing feature selection has also been considered [30, 29, 4, 15, 7]. However, unstructured sparse models neglect potential parameter coupling between tasks. In order to address this problem, multi-task group lasso has been considered in [13, 21]. In recent work, these existing ideas are combined in Group-sparse Multitask Regression and Feature Selection (G-SMuRFS) [28, 25] which takes into account both coupled feature sparsity across tasks using the $L_{2,1}$-norm and coupled group sparsity using the $G_{2,1}$-norm.

In this paper, inspired by the recent success of group sparse methods, we consider a framework of multi-task learning with hierarchical group sparsity. In particular, we consider a specific hierarchical model called multi-task sparse group lasso (MT-SGL) which considers a

*Supported in part by NSF grants by IIS-1447566, IIS-1422557, CCF-1451986, CNS-1314560, IIS-0953274, IIS-1029711, by NASA grant NNX12AQ39A, and by National Natural Science Foundation of China (61502091).

†College of Computer Science and Engineering, Key Laboratory of Medical Image Computing of Ministry of Education, Northeastern University, Shenyang, China.

‡University of Minnesota, Twin Cities.

two-level hierarchy with feature-level and group-level sparsity and parameter coupling across tasks. The regularization for MT-SGL considers both $L_{2,1}$-norm to get feature-level coupled sparsity as well as $L_{G_{2,1}}$-norm to get group-level coupled sparsity across tasks. In the context of AD, the groups correspond to specific regions-of-interest (ROIs) in the brain, and the individual features are specific properties of those regions. While the formulation of MT-SGL follows the recent advances in G-SMuRFS, the key difference between the two formulations is the optimization method. While G-SMuRFS focuses on a sub-gradient approach [25], which can be slow and inaccurate at times, we propose an efficient FISTA-style [3] accelerated descent method for the composite objective under consideration. The key step computational step in the proposed iterative optimization approach is the computation of the proximal operator corresponding to the mixed $L_{2,1}$ and $L_{G_{2,1}}$ regularization. Building on the existing literature, we show that the proximal operator can in fact be computed by simple sequential row-wise and group-wise soft-thresholding operations, making the proposed algorithm really efficient in practice.

Through empirical evaluation and comparison with four different groups of baseline methods on data from ADNI, we illustrate MT-SGL compares favorably with respect to the baselines. In particular, the performance of MT-SGL is better than G-SMuRFS according to different evaluation metrics, and the improvements are statistically significant for most tasks. Also the proposed MT-SGL outperforms all the basic baseline methods, including Ridge regression [18], Lasso [23], Group Lasso [30] applied independently to each task, and multi-task group lasso (MT-GL) based on $L_{2,1}$-norm regularization [13]. Compared with Robust MTL methods, which consider a low-rank and sparse parameter structure with suitable notions of sparsity, MT-SGL outperforms all the three types of Robust MTL methods we considered. Finally, MT-SGL is shown to be competitive with the recently proposed multi-task sparse structure learning (MSSL), which explicitly models task-relationships as a graphical model.

The rest of the paper is organized as follows. Section 2 briefly discusses related work in regression analysis. Section 3 gives the formulation of the proposed MT-SGL. Section 4 discusses an efficient optimization for MT-SGL. Section 5 discusses experimental results on regression using ADNI dataset. We conclude in Section 6.

2 Related Work

Recent studies have shown that regression analysis helps explore the relationship between imaging measures and cognitive outcomes. Many regression models have been used to enhance the AD regression accuracy on the ADNI data.

For example, the standard ridge regression model [18] is a continuous process that shrinks coefficients and hence produces non-sparse results, which is not interpretable for biomarker discovery. Sparse structure learning models tend to produce some coefficients that are exactly zero and hence give sparse results, such as Lasso [23]. Group Lasso [30] is an extension of Lasso, considering the problem of selecting grouped features for accurate prediction in regression. However, these two methods are based on single task learning structure, neglecting the correlations between multiple tasks.

Multi-task sparse matrix learning aims to learn the shared information among related tasks for improved performance. For example, $L_{2,1}$-norm [13] encourages multiple predictors from different tasks to share similar parameter sparsity patterns. Multi-task sparse structure learning (MSSL) [8] considers a joint estimation problem of the task relationship structure and the individual task parameters, which is solved using alternating minimization. FoGLasso [29] considers the efficient optimization of the overlapping group Lasso penalized problem and generalized to tackle the general overlapping group Lasso formulation based on the L_q norm. Robust multi-task learning (RMTL) [5] captures the task relationships using a low-rank structure, and simultaneously identifies the outlier tasks using a group-sparse structure.

Using ADNI data, a sparse inverse covariance estimation technique for identifying the connectivity among different brain regions is investigated by Ye et al. in [22], in which a novel algorithm based on the block coordinate descent approach is proposed for the direct estimation of the inverse covariance matrix. In [24], Ye et al. combined a predictive multi-task machine learning method with novel MR-based multivariate morphometric surface map of the hippocampus to predict future cognitive scores of patients. Recently, some works focus on longitudinal neuroimaging data. In [33], Ye et al. formulate the prediction problem as a multi-task regression problem by considering the prediction at each time point as a task and propose two novel multi-task learning formulations. Unlike the previous methods that explicitly combined the longitudinal information in a feature domain, Shen et al. in [16] propose a multi-task sparse representation classifier to discriminate between MCI-C and MCI-NC utilizing longitudinal neuroimaging data. It can be considered as the combination of the generative and discriminative methods, which are known to be effective in classification enhancement. Most existing works focus on the prediction of target using modalities

of biomarkers. In [31], Shen et al. propose to combine three modalities of biomarkers, i.e., MRI, FDG-PET, and CSF biomarkers, to discriminate between AD (or MCI) and healthy controls, using a kernel combination method. A novel multi-task learning based feature selection method is proposed by Shen et al. in [12] to effectively preserve the complementary information from multi-modal neuroimaging data for AD/MCI identification. In which, the selection of features from each modality is treated as a task and then a new constraint to preserve the inter-modality relationship during the feature selection is proposed. Taking into account both the $L_{2,1}$-norm and group information of the features ($G_{2,1}$-norm), Group-Sparse Multi-task Regression and Feature Selection (G-SMuRFS) is proposed in [25], which is used to identify quantitative trait loci for multiple disease-relevant quantitative traits and applied to a study in mild cognitive impairment and Alzheimers disease. In [28], G-SMuRFS is applied to predict cognitive outcomes using cortical surface biomarkers, in which the objective function is obtained through an iterative optimization procedure.

In this paper, we focus on the problem of multi-task sparse group learning. We devise the formulation with two forms of $L_{2,1}$ norm regularization: a $L_{2,1}$ norm for coupling multiple tasks, and a $L_{2,1}$ norm for grouping relevant MRI features in the same ROI, which close to the model of G-SMuRFS. However, we use a more efficient optimization algorithm FISTA to solve the problem. Besides, the penalty in MSSL is replaced by multi-task group regularization proposed in our work.

3 Multi-task Sparse Group Lasso

The proposed work on multi-task sparse group lasso (MT-SGL) builds on the existing literature on linear regression models with hierarchical sparsity structures over the regression coefficients. With a few recent exceptions [7, 9], most applications of hierarchical sparse regularization based modeling has focused on the regression setting with one task. Our work builds on the literature on sparse multi-task learning [1, 6], which encourages related tasks to have similar sparsity structures.

We start with a basic description of the MT-SGL model [30, 13]. Consider a MTL setting with k tasks. Let p be the number of covariates, shared across all the tasks, and n be the number of samples. Let $X \in \mathbb{R}^{n\times p}$ denote the matrix of covariates, $Y \in \mathbb{R}^{n\times k}$ be the matrix of responses with each row corresponding to a sample, and $\Theta \in \mathbb{R}^{p\times k}$ denote the parameter matrix, with column $\theta_{.h} \in \mathbb{R}^p$ corresponding to task h, $h = 1,\ldots,k$, and row $\theta_{j.} \in \mathbb{R}^k$ corresponding to feature j, $j = 1,\ldots,p$. We assume the p covariates to be divided into ℓ disjoint groups $\mathcal{G}_1,\ldots,\mathcal{G}_\ell$, with each group having $m_1,\ldots,m_\ell$ covariates respectively. In the context of AD, each group corresponds to a region-of-interest (ROI) in the brain, and the covariates in each group correspond to specific features of that region. For AD, the number of features in each group is typically 1 or 4, and the number of groups can be in the hundreds. We discuss specifics for our experimental setting in Section 5.

The MTL problem can be set-up as one of estimating the parameters based on suitable regularized loss function:

$$\min_{\Theta\in\mathbb{R}^{p\times k}} \quad L(Y,X,\Theta) + \lambda R(\Theta)\ , \tag{3.1}$$

where $L(\cdot)$ denotes the loss function and $R(\cdot)$ is the regularizer. In the current context, we assume the loss to be square loss, i.e.,

$$L(Y,X,\Theta) = \|Y - X\Theta\|_F^2 = \sum_{i=1}^{n} \|\mathbf{y}_i - \mathbf{x}_i\Theta\|_2^2\ , \tag{3.2}$$

where $\mathbf{y}_i \in \mathbb{R}^{1\times k}, \mathbf{x}_i \in \mathbb{R}^{1\times p}$ are the i-th rows of Y, X, respectively corresponding to the multi-task response and covariates for the i-th sample. We note that the MTL framework can be easily extended to other loss functions.

For the MTL regularization $R(\Theta)$, different choices encourage different structures in the estimated parameters. The most commonly used regularization is the L_1 norm, $R(\Theta) = \|\Theta\|_1$, which leads to lasso-type problem [23]. The L_1 norm regularization however does not do anything special for the MTL setting, e.g., by enforcing similar sparsity patterns across related tasks. In the MTL setting, a generalization of the lasso, the mutli-task group lasso (MT-GL) [30, 13] based on the $L_{2,1}$ norm, i.e., L_1 norm over the L_2 norm over rows $\theta_{j.}$, has been proposed. In particular, MT-GL considers

$$R(\Theta) = \|\Theta\|_{2,1} = \sum_{j=1}^{p} \|\theta_{j.}\|_2\ , \tag{3.3}$$

and is suitable for simultaneously enforcing sparsity over features for all tasks. In this paper, we consider the multi-task sparse group lasso (MT-SGL) framework [28, 25], which enforces simultaneous hierarchical sparsity by first selecting groups and then selecting some covariates from the selected groups. The difference of MT-SGL and SGL [4, 15, 7] is that SGL considers a similar hierarchical sparsity structure but for one task, whereas MT-SGL enforces the hierarchical sparsity simultaneously for all tasks. For MT-SGL, the regularizer

is given by

$$
\begin{aligned}
R(\theta) &= \lambda_1 \|\Theta\|_{2,1} + \lambda_2 \|\Theta\|_{G_{2,1}} \\
&= \lambda_1 \sum_{j=1}^{p} \|\theta_{j.}\|_2 + \lambda_2 \sum_{h=1}^{k} \|\Theta_{\mathcal{G}_h}\|_F \ ,
\end{aligned} \tag{3.4}
$$

where $\lambda_1 \geq 0, \lambda_2 \geq 0$ are the regularization parameters, and

$$
\|\Theta_{\mathcal{G}_h}\|_F = \sqrt{\sum_{j \in \mathcal{G}_h} \sum_{h=1}^{k} \theta_{jk}^2} \ , \tag{3.5}
$$

is the Frobenius norm of the parameter submatrix corresponding to group $\mathcal{G}_h$. Thus, the overall MT-SGL formulation focuses on the following regularized loss function:

$$
\min_{\Theta \in \mathbb{R}^{p \times k}} \ \|Y - X\Theta\|_F^2 + \lambda_1 \|\Theta\|_{2,1} + \lambda_2 \|\Theta\|_{G_{2,1}} \ . \tag{3.6}
$$

Several existing models can be viewed as special cases of the MT-SGL formulation. When $\lambda_2 = 0$, we get back the MT-GL formulation [21, 13]. For a single task, i.e., $k = 1$, MT-SGL simplifies to SGL [4, 15, 7], and is also related to the composite absolute penalty (CAP) family [32]. For a single task, when $\lambda_1 = 0$ and $\lambda_2 > 0$, MT-SGL reduces to the group lasso (GL) [30]; and if $\lambda_1 > 0$ and $\lambda_2 = 0$, MT-SGL reduces to the lasso [14, 23].

4 Efficient Optimization for MT-SGL

The optimization problem for MT-SGL as in (3.6) is a convex optimization problem with a composite objective with a smooth term corresponding to the square loss and a non-smooth term corresponding to the regularizer. In this section, we present a FISTA-style [3] algorithm for efficiently solving the MT-SGL problem.

Consider a general convex optimization problem with a composite objective given by

$$
\min_{\mathbf{z}} \ f(\mathbf{z}) + g(\mathbf{z}) \ , \tag{4.7}
$$

where $\mathbf{z} \in \mathbb{R}^d$, $f : \mathbb{R}^d \mapsto \mathbb{R}$ is a smooth convex function of the type $C^{1,1}$, i.e., continuously differentiable with Lipschitz continuous gradient so that $\|f(\mathbf{z}) - f(\mathbf{w})\| \leq L\|\mathbf{z} - \mathbf{w}\|$ where L denotes the Lipschitz constant, and $g : \mathbb{R}^d \mapsto \mathbb{R}$ is a continuous convex function which is possibly non-smooth. A well studied idea in efficient optimization of such composite objective functions is to start with a quadratic approximation of the form:

$$
Q_L(\mathbf{z}, \mathbf{z}_t) := f(\mathbf{z}_t) + \langle \mathbf{z} - \mathbf{z}_t, \nabla f(\mathbf{z}_t) \rangle + \frac{L}{2}\|\mathbf{z} - \mathbf{z}_t\|^2 + g(\mathbf{z}) \ . \tag{4.8}
$$

Ignoring constant terms in $\mathbf{z}_t$, the unique minimizer of the above expression can be written as

$$
\pi_L^{f,g}(\mathbf{z}_t) = \operatorname*{argmin}_{\mathbf{z}} \left\{ g(\mathbf{z}) + \frac{L}{2}\|\mathbf{z} - \left(\mathbf{z}_t - \frac{1}{L}\nabla f(\mathbf{z}_t)\right)\|^2 \right\} \ , \tag{4.9}
$$

which can be viewed as a proximal operator corresponding to the non-smooth function $g(\mathbf{z})$. A popular approach to solving problems such as (4.7) is to simply do the following iterative update:

$$
\mathbf{z}_{t+1} = \pi_L^{f,g}(\mathbf{z}_t) \ , \tag{4.10}
$$

which can be shown to have a $O(1/t)$ rate of convergence [19, 20].

For our purposes, we consider a refined version of the iterative algorithm inspired by Nesterov's accelerated gradient descent [19, 20]. The main idea, as studied in the literature as FISTA-style algorithms [3], is to iteratively consider the proximal operator $\pi_L^{f,g}(\cdot)$ at a specific linear combination of the previous two iterates $\{\mathbf{z}_t, \mathbf{z}_{t-1}\}$, in particular at

$$
\zeta_{t+1} = \mathbf{z}_t + \alpha_{t+1}(\mathbf{z}_t - \mathbf{z}_{t+1}) \ , \tag{4.11}
$$

instead of at just the previous iterate $\mathbf{z}_t$. The choice of α_{t+1} follows Nesterov's accelerated gradient descent [19, 20] and is detailed in Algorithm 1. The iterative algorithm simply updates

$$
\mathbf{z}_{t+1} = \pi_L^{f,g}(\zeta_{t+1}) \ . \tag{4.12}
$$

As shown in [3], the algorithm has a rate of convergence of $O(1/t^2)$.

A key building block in MT-SGL is the computation of the proximal operator in (4.12) when $g(\cdot) \equiv R_{\lambda_2}^{\lambda_1}(\cdot)$ is the multi-task sparse group lasso regularizer given by

$$
R_{\lambda_2}^{\lambda_1}(\Theta) = \lambda_1 \|\Theta\|_{2,1} + \lambda_2 \|\Theta\|_{G_{2,1}} \ . \tag{4.13}
$$

For MT-SGL, the iterates $\mathbf{z}_t \equiv \Theta_t$ are matrices, and the proximal operator is computed at $\zeta_{t+1} \equiv Z_{t+1} = \Theta_t + \mu_{t+1}(\Theta_t - \Theta_{t-1})$. With $f(Z_{t+1}) = \|Y - XZ_{t+1}\|_F^2$ and $V_{t+1} = (Z_{t+1} - \frac{1}{L}\nabla f(Z_{t+1}))$, the problem of computing the proximal operator $\pi_L^{f,g}(Z_{t+1}) := T_{\lambda_2/L}^{\lambda_1/L}(V_{t+1})$ is given by

$$
\begin{aligned}
T_{\lambda_2/L}^{\lambda_1/L}(V_{t+1}) &= \operatorname*{argmin}_{\Theta \in \mathbb{R}^{p \times k}} \{R_{\lambda_2/L}^{\lambda_1/L}(\Theta) + \frac{1}{2}\|\Theta - V_{t+1}\|^2\} \\
&= \operatorname*{argmin}_{\Theta \in \mathbb{R}^{p \times k}} \{R_{\lambda_2}^{\lambda_1}(\Theta) + \frac{L}{2}\|\Theta - V_{t+1}\|^2\} \ .
\end{aligned} \tag{4.14}
$$

The goal is to be able to compute $\Theta_{t+1} = T_{\lambda_2/L}^{\lambda_1/L}(V_{t+1})$ efficiently.

It can be shown that the proximal operator can be computed efficiently in two steps, as outlined below:

$$U_{t+1} = T_0^{\lambda_1/L}(V_{t+1}) \ , \tag{4.15}$$

$$\Theta_{t+1} = T_{\lambda_2/L}^0(U_{t+1}) = T_{\lambda_2/L}^{\lambda_1/L}(V_{t+1}) \ . \tag{4.16}$$

Next we show that both of these steps can be executed efficiently using suitable extensions of soft-thresholding. The update in (4.15) can be written as

$$U_{t+1} = \underset{U \in \mathbb{R}^{p\times k}}{\operatorname{argmin}} \left\{ \frac{\lambda_1}{L}\|U\|_{2,1} + \frac{1}{2}\|U - V_{t+1}\|_F^2 \right\} \ . \tag{4.17}$$

Following [13], the row-wise updates can be done by soft-thresholding as

$$\mathbf{u}_i = \frac{\max\{\|\mathbf{v}_i\|_2 - \frac{\lambda_1}{L}, 0\}}{\|\mathbf{v}_i\|_2}\mathbf{v}_i \ , \tag{4.18}$$

where $\mathbf{u}_i, \mathbf{v}_i$ are the i-th rows of U_{t+1}, V_{t+1} respectively.

Next we focus on the update (4.16), which can be written as

$$\Theta_{t+1} = \underset{\Theta \in \mathbb{R}^{p\times k}}{\operatorname{argmin}} \left\{ \frac{\lambda_2}{L}\|\Theta\|_{G_{2,1}} + \frac{1}{2}\|\Theta - U_{t+1}\|^2 \right\} \ . \tag{4.19}$$

Following [29], the group specific row-wise updates can be done by soft-thresholding as

$$\Theta_{\mathcal{G}_h} = \frac{\max\{\|\Theta_{\mathcal{G}_h}\|_F - \frac{\lambda_2}{L}, 0\}}{\|\Theta_{\mathcal{G}_h}\|_F} U_{\mathcal{G}_h} \ , \tag{4.20}$$

where $\Theta_{\mathcal{G}_h}, U_{\mathcal{G}_h}$ are group specific $m_h \times k$ sub-matrices correspond to group $\mathcal{G}_h$ in Θ_{t+1}, U_{t+1} respectively. Thus, both the steps (4.15) and (4.16) can be efficiently computed.

In practice, since the Lipschitz constant L may be unknown, we follow the adaptive strategy suggested in [3] to make sure we make progress. The pseudocode of MT-SGL is summarized in Algorithm 1, where $F(\Theta)$ denotes the objective function of MT-SGl as in (3.6), $Q_L(\Theta_1, \Theta_2)$ denotes the quadratic approximation as in (4.8) for the MT-SGL objective, and $\pi_L^{f,g}(\Theta)$ denotes the proximal operator for the MT-SGL regularization. The iterations can be terminated if the change of the function values corresponding to adjacent iterations is within a small value, say 10^{-3}.

Algorithm 1. The MT-SGL Algorithm
Input: $L_0 > 0$, $x_0 \in \mathbb{R}^n$
Step 0. Set $y_1 = x_0, \ t = 1$
Step t. ($t \geq 1$) Find the smallest nonnegative integers i_t such that with $L = 2^{i_t} L_{t-1}$

$$F(\pi_L^{f,g}(\Theta_{t-1})) \leq Q_L(\pi_L^{f,g}(\Theta_{t-1}), \Theta_{t-1})$$

Set $L_t = 2^{i_t} L_{t-1}$ and compute

$$V_t = Z_t - \frac{1}{L}\nabla f(Z_t)$$
$$\Theta_t = T_{\lambda_2/L_k}^{\lambda_1/L_k}(V_t)$$
$$\beta_{t+1} = \frac{1 + \sqrt{1 + 4\beta_t^2}}{2}$$
$$Z_{t+1} = \Theta_t + \frac{\beta_t - 1}{\beta_{t+1}}(\Theta_t - \Theta_{t-1})$$

As shown in [3], such an algorithm is guaranteed to converge at a rate $O(1/t^2)$.

5 Experimental Results

In this section, we present experimental results to demonstrate the effectiveness of the proposed MT-SGL framework on characterizing AD progression using a dataset from the ALzheimer's Disease Neuroimaging Initiative (ADNI) [26].

5.1 Experimental Setting. We start by giving a description of the datasets including features of imaging data and the cognitive scores corresponding to the tasks, followed by the statements of methods compared with MT-SGL and the comparison results.

Data: The data used in this paper were obtained from the ADNI database (adni.loni.usc.edu) [26]. The primary goal of ADNI has been to test whether serial MRI, PET, other biological markers, and clinical and neuropsychological assessments can be combined to measure the progression of MCI and early AD. Approaches to characterize AD progression will help researchers and clinicians develop new treatments and monitor their effectiveness. Further, being able to understand disease progression will increase the safety and efficacy of drug development and potentially decrease the time and cost of clinical trails. The current work focuses on MRI data. The MRI features used in our experiments are based on the imaging data from the ADNI database processed by a team from UCSF (University of California at San Francisco), who performed cortical reconstruction and volumetric segmentations with the FreeSurfer image analysis suite (http://surfer.nmr.mgh.harvard.edu/). There were $p = 327$ MRI features (covariates) in total, including the cortical thickness average (TA), standard deviation of thickness (TS), surface area (SA) cortical volume (CV) and subcortical volume (SV) for a variety of regions-of-interest (ROIs). In this work, only ADNI1 subjects with no missing feature and cognitive outcome information baseline data are included. This yields a total of $n = 816$ subjects, who are categorized into 3 baseline diagnostic groups: Cognitively Normal (CN, $n_1 = 228$), Mild Cog-

Table 1
Summary of ADNI dataset and subject information.

Category	CN	MCI	AD
Number	228	399	189
Gender (M/F)	119/109	257/142	99/90
Age (y, ag ± sd)	75.8 ± 5.0	74.7 ± 7.4	75.2 ± 7.5
Edu (y, ag ± sd)	16.1 ± 2.8	15.6 ± 3.0	14.7 ± 3.2

CN, Cognitively Normal; MCI, Mild Cognitive Impairment; AD, Alzheimer's Disease; M, male; F, female; Edu, Education; y, years; ag, average; sd, standard deviation.

nitive Impairment (MCI, $n_2 = 399$), and Alzheimer's Disease (AD, $n_3 = 189$). Table 1 lists the demographics information of all these subjects, including age, gender and education.

Tasks: For predictive modeling, we focus on 5 widely used cognitive measures [28, 11], which to the $k = 5$ tasks in our setting. In particular, the cognitive scores used in our analysis are: Alzheimers Disease Assessment Scale - cognitive total score (ADAS), Mini Mental State Exam score (MMSE), Rey Auditory Verbal Learning Test (RAVLT) total score (TOTAL), RAVLT 30 minutes delay score (T30), and RAVLT recognition score (RECOG).

Comparisons: In our experiments, we have compared MT-SGL with 4 different groups of methods: (1) Group-sparse Multitask Regression and Feature Selection (G-SMuRFS) [28], which is one of the state-of-the-art methods for characterizing AD progression; (2) Baseline methods, including Ridge regression [18], Lasso [23], Group Lasso [30] applied independently to each task, and multi-task group lasso (MT-GL) based on $L_{2,1}$-norm regularization [13]; (3) Robust Multi-Task Learning (Robust MTL) [5], which considers low-rank and group-sparse parameters; (4) Multi-task Sparse Structure Learning (MSSL) [8], a recent MTL approach which explicitly learns a sparse task dependency structure. In fact, the task dependency learning in MSSL is independent of the multi-task regularization used, and we run experiments combining MSSL with Lasso, Group Lasso, and the proposed MT-SGL.

Methodology: For all experiments, we do a 5-fold cross-validation. In particular, the AD, MCI and CN samples are randomly partitioned into 5 subsamples, respectively. We select one single subsample from AD, MCI and CN, which are combined and retained as the validation data for testing the model, and the remaining 4 subsamples of AD, MCI and CN are used as training data. Performance of different methods was evaluated by two metrics: root mean square error (RMSE, lower is better) and Pearson correlation coefficient (CC, higher is better) of the predicted values with respect to the test-set true values for each task individually. The average (avg) and standard deviation (std) of performance measures across 5 cross-validation trials are shown as avg ± std for each experiment.

In MT-SGL, we initialize the parameter $\lambda_1 = \gamma \times \lambda_1^{\max}$ and $\lambda_2 = \gamma \times \lambda_2^{\max}$ [2, 17], where $\lambda_1^{\max}$ and $\lambda_2^{\max}$ are computed as follows,

$$\lambda_1^{\max} = \|X^T Y\|_\infty ,$$
$$\lambda_2^{\max} = \operatorname*{argmax}_{h,j \in \mathcal{G}_h} \frac{1}{m_h^{1/2}} \| \max\{(|X_j^T Y| - \gamma \lambda_1^{\max}), 0\} \|_2 ,$$

where m_h denotes the number of features in group $\mathcal{G}_h$. The choices follow from the current understanding in the literature of the correct form these parameters, in particular, in terms of the dual norm of the gradient of the objective [2, 15]. Thus, the only parameter to be empirically chosen in MT-SGL is the scaling γ. As we describe later, this is done using cross-validation for each model considered.

5.2 Comparison with G-SMuRFS. Regularization parameters for G-SMuRFS and MT-SGL are chosen using a nested cross-validation strategy on the training data, with search grid in the range of 5×10^{-3} to 5×10^3 using a log-scale. Prediction performance results, measured by RMSE, CC and corresponding p-values of t-test between MT-SGL and G-SMuRFS under 5 cognitive scores are shown in Table 2 and Table 3 respectively. The prediction performance using those features selected by MT-SGL is better (i.e., lower RMSE and higher CC) than those of G-SMuRFS. In particular, MT-SGL demonstrates clear performance improvement over G-SMuRFS on predicting all 5 scores. From the results of t-test, we can see that MT-SGL is significantly better ($p \leq 0.05$) than G-SMuRFS under the score of TOTAL, RECOG and MMSE. Interestingly, one of the key differences between MT-SGL and G-SMuRFS is the optimization method, and this seems to lead to substantial differences in performance on both evaluation metrics and across all tasks.

5.3 Comparison with baseline methods. Regularization parameters for baseline methods are also chosen using a nested cross-validation strategy on the training data, with search grid in the range of 10^{-4} to 10^4 using a log-scale [13]. Prediction performance results, measured by RMSE and CC of MT-SGL and baseline methods under 5 cognitive scores are shown in Table 4 and Table 5, respectively. The sparse learning methods (Lasso, Group Lasso, MT-GL and MT-SGL) is more effective than ridge regression on most of the cognitive scores. Lasso and Group Lasso are single-task learning methods being applied independently on each task, whereas MT-GL and MT-SGL are multi-task learning

Table 2
Comparison of root mean squared error (RMSE) and p-values of t-test using RMSE between G-SMuRFS and MT-SGL across all tasks.

Method	RAVLT			MMSE	ADAS
	TOTAL	T30	RECOG		
G-SMuRFS	0.8508 ± 0.0564	0.8559 ± 0.0483	0.9152 ± 0.0375	0.8258 ± 0.0409	0.7822 ± 0.0433
MT-SGL	**0.8389 ± 0.0536**	**0.8494 ± 0.0440**	**0.9049 ± 0.0395**	**0.8170 ± 0.0392**	**0.7762 ± 0.0458**
p-value	0.0093	0.1882	0.0111	0.0087	0.2120

Table 3
Comparison of correlation coefficient (CC) of G-SMuRFS and MT-SGL across all tasks.

Method	RAVLT			MMSE	ADAS
	TOTAL	T30	RECOG		
G-SMuRFS	0.5242 ± 0.0410	0.5189 ± 0.0956	0.4029 ± 0.0173	0.5650 ± 0.0313	0.6259 ± 0.0356
MT-SGL	**0.5436 ± 0.0364**	**0.5310 ± 0.0970**	**0.4237 ± 0.0248**	**0.5778 ± 0.0349**	**0.6345 ± 0.0392**
p-value	0.0086	0.1449	0.0123	0.0115	0.2194

Table 4
Comparison of root mean squared error (RMSE) of baseline methods and MT-SGL across all tasks.

Method	RAVLT			MMSE	ADAS
	TOTAL	T30	RECOG		
Ridge	0.8566 ± 0.0564	0.8727 ± 0.0439	0.9272 ± 0.0414	0.8246 ± 0.0350	0.7844 ± 0.0532
Lasso	0.8409 ± 0.0548	0.8558 ± 0.0451	0.9194 ± 0.0337	0.8273 ± 0.0435	0.7862 ± 0.0395
Group Lasso	0.8480 ± 0.0534	0.8557 ± 0.0426	0.9203 ± 0.0366	0.8297 ± 0.0390	0.7897 ± 0.0447
MT-GL	0.8510 ± 0.0568	0.8576 ± 0.0428	0.9110 ± 0.0404	0.8301 ± 0.0442	0.7853 ± 0.0410
MT-SGL	**0.8389 ± 0.0536**	**0.8494 ± 0.0440**	**0.9049 ± 0.0395**	**0.8170 ± 0.0392**	**0.7762 ± 0.0458**

Table 5
Comparison of correlation coefficient (CC) of Baseline methods and MT-SGL across all tasks.

Method	RAVLT			MMSE	ADAS
	TOTAL	T30	RECOG		
Ridge	0.5139 ± 0.0389	0.4900 ± 0.0940	0.3756 ± 0.0226	0.5653 ± 0.0319	0.6214 ± 0.0398
Lasso	0.5411 ± 0.0342	0.5203 ± 0.1015	0.3931 ± 0.0207	0.5633 ± 0.0410	0.6216 ± 0.0353
Group Lasso	0.5293 ± 0.0361	0.5213 ± 0.0896	0.3894 ± 0.0245	0.5605 ± 0.0332	0.6184 ± 0.0313
MT-GL	0.5247 ± 0.0365	0.5173 ± 0.0948	0.4108 ± 0.0165	0.5598 ± 0.0398	0.6233 ± 0.0323
MT-SGL	**0.5436 ± 0.0364**	**0.5310 ± 0.0970**	**0.4237 ± 0.0248**	**0.5778 ± 0.0349**	**0.6345 ± 0.0392**

Table 6
Comparison of the root mean squared error (RMSE) of variants of Robust MTL and MT-SGL.

Method	RAVLT			MMSE	ADAS
	TOTAL	T30	RECOG		
Robust MTL	0.8531 ± 0.0637	0.8662 ± 0.0483	0.9130 ± 0.0413	0.8212 ± 0.0441	0.7853 ± 0.0526
RMTL-L_1	0.8528 ± 0.0637	0.8653 ± 0.0484	0.9115 ± 0.0412	0.8211 ± 0.0445	0.7850 ± 0.0527
RMTL-$G_{2,1}$	0.8540 ± 0.0622	0.8669 ± 0.0477	0.9127 ± 0.0412	0.8213 ± 0.0442	0.7855 ± 0.0533
MT-SGL	**0.8389 ± 0.0536**	**0.8494 ± 0.0440**	**0.9049 ± 0.0395**	**0.8170 ± 0.0392**	**0.7762 ± 0.0458**

Table 7
Comparison of the correlation coefficient (CC) of variants of Robust MTL and MT-SGL.

Method	RAVLT			MMSE	ADAS
	TOTAL	T30	RECOG		
Robust MTL	0.5216 ± 0.0444	0.5019 ± 0.0773	0.4072 ± 0.0310	0.5743 ± 0.0342	0.6237 ± 0.0345
RMTL-L_1	0.5214 ± 0.0444	0.5022 ± 0.0794	0.4099 ± 0.0305	0.5745 ± 0.0345	0.6241 ± 0.0348
RMTL-$G_{2,1}$	0.5205 ± 0.0439	0.5007 ± 0.0791	0.4084 ± 0.0312	0.5738 ± 0.0340	0.6238 ± 0.0333
MT-SGL	**0.5436 ± 0.0364**	**0.5310 ± 0.0970**	**0.4237 ± 0.0248**	**0.5778 ± 0.0349**	**0.6345 ± 0.0392**

methods. Prediction performances of Lasso and Group Lasso are similar, and MT-GL is not better than Lasso and Group Lasso. The proposed MT-SGL outperforms all the other four techniques on both metrics and across all tasks. Due to lack of space, we could not conclude the corresponding p-values of t-test between the methods in Sections 5.3 and 5.4.

5.4 Comparison with Robust MTL [5]. The Robust MTL integrates Low-Rank and Group-Sparse Structures to multi-task learning. The group sparse regularization used in [5] is the $L_{2,1}$-norm, and we create variants of that model by using L_1-norm and $G_{2,1}$-norm, respectively calling them RMTL-L_1 and RMTL-$G_{2,1}$ respectively. Parameters for these three methods are set following the same approach as the baseline methods, i.e., chosen using a nested cross-validation strategy on the training data, with search grid in the range of 10^{-4} to 10^4 using a log-scale. Table 6 and Table 7 report the prediction performance results of three Robust MTL methods and MT-SGL in terms of RMSE and CC for all 5 cognitive scores. Prediction performances of the variants of Robust MTL with different regularizations were similar, and MT-SGL outperformed the Robust MTL methods for both metrics and across all tasks. Interestingly, the results suggest that the popular low-rank and suitably sparse structure may not be as effective for characterizing AD progression.

5.5 Comparison and Integration with MSSL [8]. MSSL [8] considers a joint estimation problem of the individual task parameters and the task relationship structure, which is solved using alternating minimization. In [8], the individual task parameters are optimized using FISTA, and task relationship structure is optimized by ADMM. Parameters for these two parts are chosen using a nested cross-validation strategy on the training data, with search grid in the range of 5×10^{-3} to 5×10^3 using log-scale, similar to MT-SGL and other baselines. MSSL comes in two variants: p-MSSL, considering dependencies among tasks parameters, and r-MSSL, considering dependencies among task-specific residual errors. We experiment with both variants in our work.

In MSSL, the sparse regularizations used for both task parameters and task dependencies are L_1-norm, we integrate the ideas in MT-SGL and MSSL by using the MT-SGL regularization to capture the individual task parameters. We call the integrated method MSSL-SGL. The results of MSSL-SGL and MT-SGL are shown in Table 8 and Table 9. Prediction performance of the MSSL-SGL and MT-SGL are similar, which suggests that learning the task relationships explicitly has no added advantage here. Interestingly, both MSSL-SGL and MT-SGL are much better than the basic MSSL and Lasso, which illustrates the value of the hierarchical sparse regularization in the multi-task setting. Besides, we fix the parameters and run the MT-SGL, P-MSSL-SGL and R-MSSL-SGL. P-MSSL-SGL and R-MSSL-SGL took 99.75 seconds and 97.23 seconds, whereas our MT-SGL took only 32.62 seconds, which demonstrates MT-SGL is more efficient and it is applicable to the learning of larger scale data sets.

6 Conclusions

Many clinical/cognitive measures have been designed to evaluate the cognitive status of the patients and they have been used as important criteria for clinical diagnosis of probable AD. We consider three types of cognitive measures in our work: Alzheimer's Disease Assessment Scale - cognitive total score (ADAS), Mini Mental State Exam score (MMSE) and Rey Auditory Verbal Learning Test (RAVLT). ADAS is the gold standard in AD drug trial for cognitive function assessment. It is the most popular cognitive testing instrument to measure the severity of the most important symptoms of AD, including the disturbances of memory, language, praxis, attention and other cognitive abilities, which have been referred as the core symptoms of AD. MMSE is widely used to measure cognitive impairment, previous studies have shown the correlation between MMSE and the underlying AD pathology and progressive deterioration of functional ability. RAVLT is a test of episodic memory and sensitive to deficiencies of memory found in many groups of patients, being widely used for the diagnosis of memory disturbances.

In this paper we propose a framework for multi-task learning with hierarchical group sparsity to better characterize Alzheimer's disease. Our proposed framework considers a two-level hierarchy with feature-level and group-level sparsity and parameter coupling across tasks. The objective formulation was solved by an efficient FISTA-style accelerated descent method. The key step computational step in the proposed iterative optimization approach is the computation of the proximal operator corresponding to the mixed $L_{2,1}$ and $L_{G_{2,1}}$ regularization. Extensive experiments on ADNI dataset illustrate that multi-task sparse group lasso not only improves progression performance, but also helps investigate the correlation within the ROIs and captures the relationships between multiple tasks.

References

Table 8
The root mean squared error (RMSE) of MSSL-GL and MT-SGL.

Method	RAVLT			MMSE	ADAS
	TOTAL	T30	RECOG		
P-MSSL	0.8415 ± 0.0512	0.8541 ± 0.0467	0.9114 ± 0.0371	0.8194 ± 0.0420	0.7808 ± 0.0404
R-MSSL	0.8412 ± 0.0511	0.8533 ± 0.0478	0.9120 ± 0.0371	0.8185 ± 0.0428	0.7794 ± 0.0413
Lasso	0.8409 ± 0.0548	0.8558 ± 0.0451	0.9194 ± 0.0337	0.8273 ± 0.0435	0.7862 ± 0.0395
P-MSSL-SGL	**0.8388 ± 0.0532**	**0.8480 ± 0.0434**	0.9052 ± 0.0393	0.8172 ± 0.0392	0.7764 ± 0.0456
R-MSSL-SGL	0.8389 ± 0.0536	0.8494 ± 0.0440	**0.9049 ± 0.0395**	0.8171 ± 0.0392	**0.7762 ± 0.0457**
MT-SGL	0.8389 ± 0.0536	0.8494 ± 0.0440	**0.9049 ± 0.0395**	**0.8170 ± 0.0392**	0.7762 ± 0.0458

Table 9
The correlation coefficient (CC) of MSSL-SGL and MT-SGL.

Method	RAVLT			MMSE	ADAS
	TOTAL	T30	RECOG		
P-MSSL	0.5387 ± 0.0380	0.5224 ± 0.1020	0.4095 ± 0.0189	0.5740 ± 0.0346	0.6276 ± 0.0383
R-MSSL	0.5391 ± 0.0383	0.5250 ± 0.1017	0.4085 ± 0.0193	0.5754 ± 0.0349	0.6290 ± 0.0396
Lasso	0.5411 ± 0.0342	0.5203 ± 0.1015	0.3931 ± 0.0207	0.5633 ± 0.0410	0.6216 ± 0.0353
P-MSSL-SGL	**0.5437 ± 0.0361**	0.5308 ± 0.0970	0.4232 ± 0.0249	0.5777 ± 0.0344	0.6344 ± 0.0388
R-MSSL-SGL	0.5436 ± 0.0364	**0.5310 ± 0.0969**	0.4237 ± 0.0249	**0.5778 ± 0.0348**	**0.6345 ± 0.0392**
MT-SGL	0.5436 ± 0.0364	0.5310 ± 0.0970	**0.4237 ± 0.0248**	0.5778 ± 0.0349	**0.6345 ± 0.0392**

[1] A. Argyriou, T. Evgeniou, and M. Pontil. Multi-task feature learning. *In NIPS*, 2007.

[2] A. Banerjee, S. Chen, F. Fazayeli, and V. Sivakumar. Estimation with norm regularization. *In NIPS*, pages 1556–1564, 2014.

[3] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2(1):183–202, 2009.

[4] S. Chatterjee, K. Steinhaeuser, A. Banerjee, S. Chatterjee, and A. R. Ganguly. Sparse group lasso: Consistency and climate applications. *In SDM*, pages 47–58, 2012.

[5] J. Chen, J. Zhou, and J. Ye. Integrating low-rank and group-sparse structures for robust multi-task learning. *ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 42–50, 2011.

[6] T. Evgeniou and M. Pontil. Regularized multitask learning. *In KDD*, pages 109–117, 2004.

[7] J. Friedman, T. Hastie, and R. Tibshiran. A note on the group lasso and a sparse group lasso. *Preprint*, 2010.

[8] A. R. Goncalves, P. Das, S. Chatterjee, V. Sivakumar, F. J. V. Zuben, and A. Banerjee. Multi-task sparse structure learning. *In CIKM*, pages 451–460, 2014.

[9] R. Jenatton, J. Mairal, G. Obozinski, and F. Bach. Proximal methods for sparse hierarchical dictionary learning. *In ICML*, pages 487–494, June 2010.

[10] Z. Khachaturian. Diagnosis of alzheimers disease. *Archives of Neurology*, 42(11):1097–1105, 1985.

[11] T. Li, J. Wana, Z. Zhang, J. Yan, S. Kim, S. Risacher, S. Fang, M. Beg, L. Wang, A. Saykin, and L. Shen. Hippocampus as a predictor of cognitive performance: comparative evaluation of analytical methods and morphometric measures. *In: MICCAI Workshop on Novel Imaging Biomarkers for Alzheimers Disease and Related Disorders (NIBAD12)*, pages 133–144, 2012.

[12] F. Liu, C. Wee, H. Chen, and D. Shen. Inter-modality relationship constrained multi-modality multi-task feature selection for alzheimer's disease and mild cognitive impairment identification. *NeuroImage*, 84:466–475, 2014.

[13] J. Liu, J. Chen, and J. Ye. Multi-task feature learning via efficient $l_{2,1}$-norm minimization. *Uncertainty in Artificial Intelligence*, pages 339–348, 2009.

[14] J. Liu and J. Ye. Efficient euclidean projections in linear time. *In ICML*, pages 657–664, 2009.

[15] J. Liu and J. Ye. Moreau-yosida regularization for grouped tree structure learning. *In NIPS*, 2010.

[16] M. Liu, H. Suk, and D. Shen. Multi-task sparse classifier for diagnosis of mci conversion to ad with longitudinal mr images. *In: Wu G, Zhang D, Shen D, Yan P, Suzuki K, Wang F (eds) Machine Learning in Medical Imaging. Springer International Publishing*, 8184:243–250, 2013.

[17] L. Meier, S. van de Geer, and P. Bühlmann. The group lasso for logistic regression. *J. Roy. Statist. Soc. Ser. B*, 70:53–71, 2008.

[18] D. N. and H. Smith. *Applied Regression Analysis*. Wiley, 1981.

[19] Y. Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1):127–152, 2005.

[20] N. Parikh and S. Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, pages 1–96, 2013.

[21] K. Puniyani, S. Kim, and E. P. Xing. Multi-population gwa mapping via multi-task regularized regression. *Bioinformatics*, 26(12):i208–i216, 2010.

[22] L. Sun, R. Patel, J. Liu, K. Chen, T. Wu, J. Li,

E. Reiman, and J. Ye. Mining brain region connectivity for alzheimer's disease study via sparse inverse covariance estimation. *In SIGKDD*, pages 1335–1344, 2009.

[23] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B*, 58(1):267–288, 1996.

[24] S. Tsao, N. Gajawelli, J. Zhou, J. Shi, J. Ye, Y. Wang, and N. Lepore. Evaluating the predictive power of multivariate tensor-based morphometry in alzheimers disease progression via convex fused sparse group lasso. *In Soc Photo Opt Instrum Eng*, pages 9034–90342L, 2014.

[25] H. Wang, F. Nie, H. Huang, S. Kim, K. Nho, S. L. Risacher, A. J. Saykin, and L. Shen. Identifying quantitative trait loci via group-sparse multitask regression and feature selection: an imaging genetics study of the adni cohort. *Bioinformatics*, 28(2):229–237, 2012.

[26] M. Weiner, P. Aisen, C. J. Jack, W. Jagust, J. Trojanowski, L. Shaw, A. Saykin, J. Morris, N. Cairns, L. Beckett, A. Toga, R. Green, S. Walter, H. Soares, P. Snyder, E. Siemers, W. Potter, P. Cole, , and M. Schmidt. The alzheimer's disease neuroimaging initiative: progress report and future plans. *Alzheimers Dement*, 6:202–211, 2010.

[27] A. Wimo, B. Winblad, H. Aguero-Torres, and E. von Strauss. The magnitude of dementia occurrence in the world. *Alzheimer Disease and Associated Disorders*, 17(2):63–67, 2003.

[28] J. Yan, T. Li, H. Wang, H. Huang, J. Wan, K. Nho, S. Kim, S. L. Risacher, A. J. Saykin, and L. Shen. Cortical surface biomarkers for predicting cognitive outcomes using group $l_{2,1}$ norm. *Neurobiology of Aging*, 36(1):S185–S193, January 2015.

[29] L. Yuan, J. Liu, and J. Ye. Efficient methods for overlapping group lasso. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(9):2104–2116, September 2013.

[30] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B*, 68(1):49–67, February 2006.

[31] D. Zhang, Y. Wang, L. Zhou, H. Yuan, and D. Shen. Multimodal classification of alzheimer's disease and mild cognitive impairment. *NeuroImage*, 55:856–867, 2011.

[32] P. Zhao, G. Rocha, and B. Yu. The composite absolute penalties family for grouped and hierarchical variable selection. *Annals of Statistics*, 37(6A):3468–3497, December 2009.

[33] J. Zhou, J. Liu, V. Narayan, and J. Ye. Modeling disease progression via multi-task learning. *NeuroImage*, 78:233–248, 2013.

Behavioral Phenotyping of Digital Health Tracker Data

Thomas Quisel* tquisel@evidation.com
Luca Foschini* lfoschini@evidation.com
Alessio Signorini* asignorini@evidation.com

Abstract

With the surge in popularity of wearable technologies a large percentage of the US population is now tracking activities such as sleep, diet and physical exercise. In this study we empirically evaluate the ability to predict metrics (e.g., weekly alcohol consumption) directly related to health outcomes from densely sampled, multi-variate time series of behavioral data. Our predictive models are based on temporal convolutional neural networks and take as input the raw historical time series of daily step counts, sleep duration, and weight scale usage sourced from an online population of thousands of digital trackers. The prediction accuracy achieved outperforms several strong baselines that use hand-engineered features and indicates that tracker data contains valuable information on individuals' lifestyles even for behavioral aspects seemingly unrelated to the measured quantities. We believe that this insight can be applied to the design of new digital interventions and enable future large-scale preventive care strategies.

Keywords Digital health, activity tracking, behavioral phenotyping, mHealth, temporal convolutional neural networks.

1 Introduction

It is estimated that 69% of the U.S. population keeps track of their weight, diet, or exercise routine, and 20% of trackers claim to leverage technology such as digital health devices and apps to perform self-monitoring [9, 28]. With tech giants like Apple and Google entering the arena of wearable technologies, the market for activity trackers and wearable devices is projected to increase to more than $50 billion by 2018 [33].

Not only has the number of digital health trackers surged in recent years, the breadth of measures these devices can quantify has also dramatically expanded. The last Consumer Electronic Conference held every year in Las Vegas [1] featured consumer-grade sensors able to continuously capture hemoglobin, arterial oxygen saturation (SpO2), pulse rate (PR), perfusion index (PI), and Plethysmograph Variability Index (PVI). With these new additions, the digital tracker ecosystem starts resembling the capabilities of the sensor arrays found in ICU rooms [19], and constitutes a significant step forward from pedometers and calorie counters that have become prevalent in smartphones and watches.

While it is disputed whether digital health tracking alone can lead to healthier behavior in the adopter [24], it is clear that the wealth of information provided by the trackers, however inaccurate [20], can be predictive of lifestyle. In our recent study [29] we provided evidence of this fact by showing that changes in an individual's adherence to weight tracking and food logging are predictive of weight change over time.

A large body of empirical evidence demonstrates that lifestyle plays an important role in long term health outcomes [8, 25, 32]. An illustrative example for the case of cardiovascular diseases is the Harvard Healthy Heart Score survey [2], which calculates a Cardiovascular Lifestyle Risk Score based on lifestyle habits such as smoking, physical activity, and diet. Some of the questions on the survey, such as, "During the past year, what was your average time per week spent on walking (slower than 3 miles per hour)," can be immediately answered by the step count reported by a pedometer. Other questions, such as the lifestyle ones pertaining to alcohol consumption habits, cannot be directly inferred from tracker summary statistics.

That said, the temporally dense information recorded by digital trackers contain more complex patterns. For example, a decrease in sleep duration on Friday nights and corresponding lower step counts in the following day may correlate with a weekly habit of partying–a pattern that might go undetected when only looking at summary statistics of the sleep duration or the step counts taken in isolation–and may be a good predictor of increased weekly alcohol consumption.

1.1 Contribution

In this work we extend the analysis of Pourzanjani et al. [29] in the pursuit of closing the gap between behavioral phenotyping and health outcomes. From an outcome perspective, we focused on metrics known to be important predictors of future health:

- Their (measured) Body-Mass Index;
- The (self reported) frequency of weekly alcohol

*Evidation Health - Menlo Park, CA

consumption;

- The (measured) propensity to increase their level of physical exercise as a result of a digital intervention.

The Body-Mass Index (BMI) is strongly correlated with other aspects of an individual's health and abnormalities are estimated to cost 21% ($190.2 billion) of annual medical spending in the United States [6]. Similarly, immoderate alcohol consumption and lack of physical exercise are associated with unfavorable health outcomes [8, 25, 30].

From a methods perspective, we present a model based on a temporal Convolutional Neural Network (CNN) that allows for prediction of the outcome variables from the raw time series recorded by the digital health trackers: daily step count, sleep duration, and weight scale utilization (i.e., whether or not the individual has weighed themselves on a given day).

We show that the CNN approach matches or outperforms several strong baselines that leverage hand-engineered features, in line with the same groundbreaking advances that representation learning and unsupervised feature discovery via deep learning have brought to image processing [16], speech recognition [10], and natural language processing [22].

Finally, we show that the performance of the CNN model is robust to the imputation strategy used for the time series, in line with the hypothesis of Razavian et al. [31] who argue that missing values do not constitute a major concern in temporally dense time series, such as the ones under study.

2 Related work

The task of deriving observable physiological traits from clinical data is generally termed phenotyping [26]. Although phenotyping has become an established practice in medical machine learning, to the best of our knowledge this is the first attempt at extracting phenotypes from behavioral data to predict health-related outcomes. In [29], Pourzanjani et al. showed that frequency of weight tracking and gaps in tracking behavior are predictive of an individual's weight change. The methodologies used in their work only considered temporal summary statistics such as the frequency and gaps between reported measurements, computed separately on a single time series, and predicted a single outcome. On the contrary, the method presented in this paper uses as input the raw multivariate time-series of digital health measurements and considers several diverse health-related outcome variables. From a methods perspective, the present work shares commonalities with the machine learning research focused on phenotyping of medical data, but while in general medical settings observations such as vital signs, lab test results, and subjective assessments are sampled irregularly [21], behavioral data recorded by digital health trackers is *dense* and recorded at least with daily frequency.

In the medical machine learning community, several recent works have addressed the topic of phenotyping of clinical data. In their recent work [31], Razavian et al. use a multi-resolution CNN to perform early detection of multiple diseases from irregularly measured sparse lab values. We benefit from the same ease of interpretability of the learned model brought about by the temporal convolutional approach, however, as Razavian et al. argue in their paper, their method focuses more on devising a highly refined imputation strategy to cope with missing data, a problem far less common on digital health data.

Another very recent work by Lipton et al. [19], uses Long Short-Term Memory (LSTM) networks, a variant of Recurrent Neural Networks (RNNs), to identify patterns and classify 128 diagnoses from multivariate time series of 13 frequently but irregularly sampled clinical measurements. As pointed out in [31], it is not clear whether the long-term dependencies that RNNs very effectively model are necessary in contexts similar to the one under study.

Neural networks in general have a long history of applications in the medical domain [3, 5]. More recently, deep learning has been applied to assess Parkinsons Disease [12] and feed-forward networks have been applied to medical time series for gout, leukemia, and critical illness classification [7, 17]. Finally, non-neural-network based techniques have been leveraged to perform classification of multi-variate time-series in the medical domain. See [23] for a review.

3 Data

The source of our data is AchieveMint[1], a consumer rewards platform for healthy activities powered by Evidation Health[2]. The AchieveMint platform automatically collects data (e.g., step counts) from its users' digital trackers and aggregates it into their accounts rewarding health related activities (e.g., a run) with points. We considered binary classification tasks on three datasets. Each dataset is composed of pairs of multivariate time series and binary labels, each pair associated with a different individual. The multivariate time series for a given individual contained a history (different lengths were used in different datasets) of daily step counts, sleep durations, and interactions with a connected scale (a binary indicator whose value

[1] http://www.achievemint.com
[2] http://www.evidation.com

is 1 if the user weighed themselves through a connected scale, and 0 otherwise). All the time series measurements were passively recorded by the relevant tracker (i.e., pedometer, sleep trackers, scale); none of them was self-reported. A detailed description of each dataset and prediction task is provided below:

UPTAKE The dataset consists of 1,996 users who took part in an IRB-approved study designed to increase the level of physical activity through small monetary incentives. Over the two-week intervention period, the groups were offered the same average incentives for physical activity. We considered a subset of the users in the experimental arms (the control group did not undergo the intervention) that have a history of measurements of at least 147 days. We assigned a positive label to users whose median daily step count during the intervention period had shown an uptake of more than 2,000 steps/day[3] compared to the median pre-intervention. This resulted in 22% positive labels. A visual representation of the daily step count histories for a few hundred users is shown in Figure 1.

BMI The dataset consists of 1,978 AchieveMint users who have shared their BMI measurements (weight reported by a connected scale, height self-reported). We assigned a positive label to users with BMI higher than a chosen clinically relevant threshold [34], which resulted in 44% positive labels.

ALCOHOL The dataset consists of 815 users that agreed to participate in a one-click survey answering the lifestyle question "On average, do you have more than one drink per week?" inspired by the Healthy Heart Survey [2]. We assigned a positive label to the users who answered the question positively, which resulted in 33% positive labels.

4 Methods

We learned a binary classifier to generate estimates $\hat{y_u}$ of the true labels y_u for each user u from the multivariate time series of observations for the time period T, $X_u = x_1, \ldots, x_T$. Each observation x_t is a vector of size K, representing one of the K behavioral biomarkers (in our case, $K = 3$: step count, sleep duration in hours, and binary indicator of weight measurement) recorded for a given day.

Our temporal convolution model is shown in Figure 2. The input to the model can be raw (un-imputed) observations, imputed observations, or the concatenation of imputed data and the binary observation pattern.

Following Zheng et al. [35] each time series is fed separately to a two-stage univariate feature extraction, where each stage is composed of a convolution filter followed by a Max-pooling layer and sigmoid layer (unlike [35], which uses avg-pooling).

The output of the feature extraction layers is flattened and fed to a standard fully connected MLP with hidden layer for classification [18].

The specific architectural choices for the shared part of the prediction network is as follows: we set the number of filters to be 8 for the first convolution layer and 4 for the second, with a kernel length of 7 for the first layer and 5 for the second layer and step size of 1 for all convolutional layers. Each Max-pooling layer has a length of 2 with step size of 2 (i.e. no overlap). Each convolution layer is followed by a Sigmoid nonlinearity ($sigmoid(t) = \frac{1}{1+e^{-t}}$).

We added 1 fully connected hidden layer with 300 nodes after the concatenation of outputs of all convolution layers. After the last Sigmoid layer corresponding to the output of the shared part of the network we added a fully connected layer (of the size of 2 nodes corresponding to binary outcome) and a Log Softmax Layer in this order. We use ADAM [15] instead of SGD for parameter updates.

The loss function for each label is the negative log likelihood of the true label: $L = -\sum_{u \in U} \sum_{c \in 0,1} y_{c,u} \log \hat{y_{c,u}}$. Each gradient is back-propagated throughout the entire prediction network.

We set the momentum to 0.9, use a fixed lr of 0.005, and set a 0.003 weight decay. The ALCOHOL task required a 0.006 weight decay to avoid overfitting,

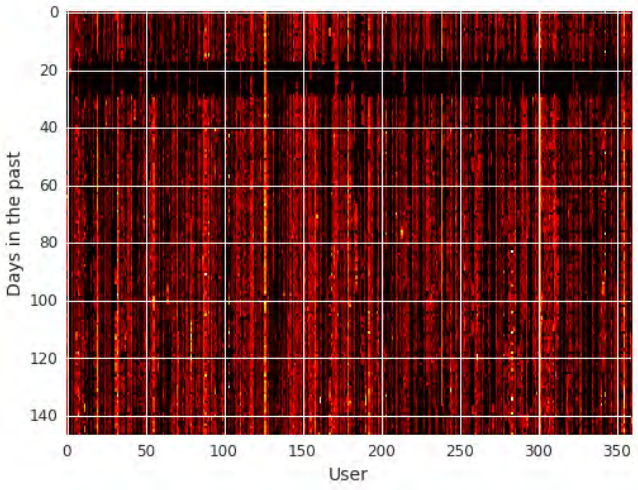


Figure 1: Heatmap for a few hundred users of the UPTAKE dataset. Brighter pixels correspond to higher step counts. The darker band on top represent marks a period of low activity during the winter holidays

[3]Considered an increased in activity that when sustained in the long term can provide health benefits [30]

since it is a smaller dataset.

We implemented our model in the Caffe [14] environment.

5 Results

The CNN model is fed the imputed times series (using linear imputation). We found that mean-centering each day of a time series before imputation, so that the mean across users is zero for each given day, significantly improved the results.

To test the robustness of our model to missing values, we considered a variant of the model, CNN-U, in which each input time series is augmented with its utilization signal: a time series of binary indicators encoding whether the data for a given day was missing and had been imputed. Input time series that are already utilization signals, such as the weight measurement one, are not augmented.

In Table 1 we reported the mean area under the ROC curve (AUC) over 4 cross-validated folds for the three datasets. Given the small size of our datasets, a 4-fold cross-validation mean AUC provides more robust and stable results. We compared the two convolutional neural network approaches with several baseline models (logistic regression, random forest (RF) and SVM classifiers) trained on hand-engineered features. Following [7, 19, 21] the features we computed for each variable are the mean, standard deviation, median, quartiles, minimum, maximum, and a count of non-missing values. Hyperparameters for the baseline models trained on the hand-engineered features were tuned using random search [4]. The SVM hyperparameter search space was derived from [13].

	CNN	CNN-U	logistic	RF	SVM
Uptake	**0.699**	0.698	0.629	0.622	0.611
BMI	0.640	0.639	0.653	**0.654**	0.648
Alcohol	0.549	**0.552**	0.526	0.551	0.526

Table 1: 4-fold cross-validated AUC for the three datasets. CNN is the temporal convolutional model that takes as input the linearly imputed time series. CNN-U takes the step count and sleep utilization time series as additional inputs. Logistic, random forest (RF) and SVM models are trained on hand-engineered features.

We observed that the CNN models significantly outperform the baseline ones on the Uptake dataset and slightly on the Alcohol dataset. We also note that the AUC values reported demonstrate that daily recordings of step counts, sleep duration, and scale usage, however inaccurate, are predictive of an individual's overall behavior, even for health-related properties not directly related to the observed variables.

Unlike other neural network based models, CNNs provide direct interpretability of the learned models. The weekly trends learned by the CNN in the first layer convolutional filters for the step count biomarker are reported in Figure 3.

Figure 3: Convolution weights learned by the CNN on the step count time series of the Uptake dataset. Each graph shows the 7 learned weights for each of the nodes in the first convolutional layer for step counts.

Since our dataset is small when compared to datasets found in common deep learning tasks, regularization heavily affects the results. Figure 4 shows the learning curves for both CNN models and demonstrates that the regularization parameters used successfully avoid overfitting.

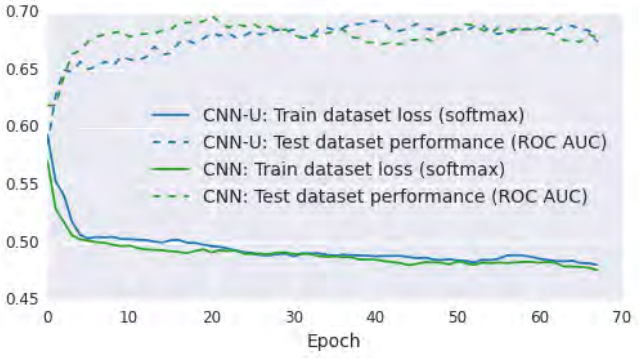


Figure 4: Training set softmax loss and testing set AUC vs. training epocs for the Uptake dataset. The curves demonstrates that the regularization employed successfully prevents overfitting. In addition, the negligible difference between CNN and CNN-U highlights the robustness of the model to imputation.

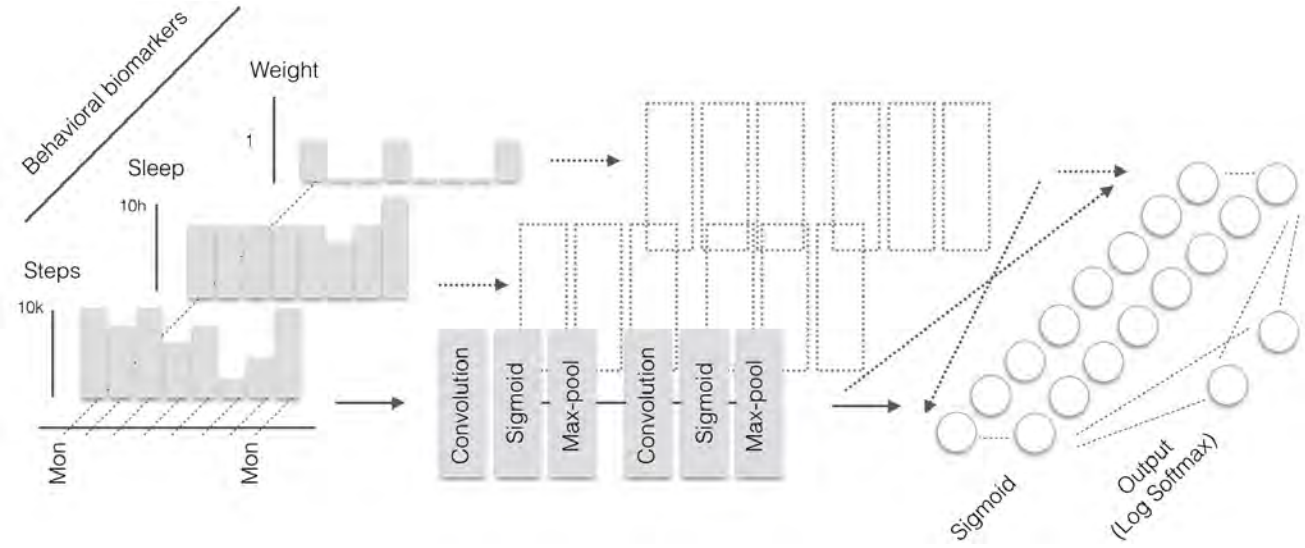


Figure 2: The temporal CNN architecture. Each behavioral biomarker time series is fed through the convolutional layer separately. The output layer, the only one label-specific, is a Log SoftMax classifier.

6 Conclusion and Future work

The results presented in this paper indicate that a temporal convolutional neural network learned on raw time series of data streamed directly from digital health trackers can accurately predict important health-related variables.

Our models' automatic behavioral phenotyping has many potential applications: (1) it can be used to passively infer lifestyle choices of individuals with the goal of complementing, or even replacing, surveys (e.g., [2]) that must actively acquire such data to determine its relationship with disease risk factors; (2) it can be used to augment models based on more traditional medical data sources [31] to further improve medical decision making; and (3) the learned phenotypes can be used to optimize behavior-changing interventions [11, 27] with the goal of proactively addressing high-risk behaviors. The high accuracy achieved on predicting the propensity of individuals to increase their physical activity as a result of a digital intervention can improve targeting decisions for interventions. More-sophisticated, higher-cost interventions (e.g., in-person coaching) can be targeted to individuals identified as less inclined to improve, while simpler and more cost-effective strategies (e.g., an email reminder) can be sufficient for those who display a higher propensity to change.

Possible extensions to our approach include improving the performance of the network by further tuning its architecture and testing it on a larger set of input variables, including fixed-time ones (e.g., demographics). The model could also benefit from more sophisticated imputation strategies (e.g., [31]) and modules that encode longer terms dependencies (e.g., by multiresolution approach as in [31] or using RNN-like techniques, such as in [19]).

References

[1] Ces 2016: Running list of health and wellness devices. `http://mobihealthnews.com/content/ces-2016-running-list-health-and-wellness-devices`. Accessed: 2016-01-13.

[2] Harvard healthy heart score. `https://healthyheartscore.sph.harvard.edu/`. Accessed: 2016-01-17.

[3] W. G. Baxt. Application of artificial neural networks to clinical medicine. *The lancet*, 346(8983):1135–1138, 1995.

[4] J. Bergstra and Y. Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13.

[5] R. Caruana, S. Baluja, T. Mitchell, et al. Using the future to" sort out" the present: Rankprop and multitask learning for medical risk evaluation. *Advances in neural information processing systems*, pages 959–965, 1996.

[6] J. Cawley and C. Meyerhoefer. The medical care costs of obesity: an instrumental variables approach. *Journal of health economics*, 31(1):219–230, 2012.

[7] Z. Che, D. Kale, W. Li, M. T. Bahadori, and Y. Liu. Deep computational phenotyping. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 507–516. ACM, 2015.

[8] S. E. Chiuve, M. L. McCullough, F. M. Sacks, and E. B. Rimm. Healthy lifestyle factors in the primary prevention of coronary heart disease among men benefits among users and nonusers of lipid-lowering and antihypertensive medications. *Circulation*, 114(2):160–167, 2006.

[9] S. Fox and M. Duggan. *Tracking for health*. Pew Research Center's Internet & American Life Project, 2013.

[10] A. Graves and J. Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5):602–610, 2005.

[11] D. Halpern. *Inside the Nudge Unit*. Random House, 2015.

[12] N. Y. Hammerla, J. M. Fisher, P. Andras, L. Rochester, R. Walker, and T. Plötz. Pd disease state assessment in naturalistic environments using deep learning. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[13] C.-W. Hsu, C.-C. Chang, C.-J. Lin, et al. A practical guide to support vector classification, 2003.

[14] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the ACM International Conference on Multimedia*, pages 675–678. ACM, 2014.

[15] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[17] T. A. Lasko, J. C. Denny, and M. A. Levy. Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data. *PloS one*, 8(6):e66341, 2013.

[18] Y. A. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller. Efficient backprop. In *Neural networks: Tricks of the trade*, pages 9–48. Springer, 2012.

[19] Z. C. Lipton, D. C. Kale, C. Elkan, and R. Wetzell. Learning to diagnose with LSTM recurrent neural networks. *CoRR*, abs/1511.03677, 2015.

[20] C. MA, B. HA, V. KG, and P. MS. Accuracy of smartphone applications and wearable devices for tracking physical activity data. *JAMA*, 313(6):625–626, 2015.

[21] B. M. Marlin, D. C. Kale, R. G. Khemani, and R. C. Wetzel. Unsupervised pattern discovery in electronic health care data using probabilistic clustering models. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, pages 389–398. ACM, 2012.

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

[23] R. Moskovitch and Y. Shahar. Classification-driven temporal discretization of multivariate time series. *Data Min. Knowl. Discov.*, 29(4):871–913, July 2015.

[24] P. MS, A. DA, and V. KG. Wearable devices as facilitators, not drivers, of health behavior change. *JAMA*, 313(5):459–460, 2015.

[25] K. J. Mukamal, S. E. Chiuve, and E. B. Rimm. Alcohol consumption and risk for coronary heart disease in men with healthy lifestyles. *Archives of Internal Medicine*, 166(19):2145–2150, 2006.

[26] A. Oellrich, N. Collier, T. Groza, D. Rebholz-Schuhmann, N. Shah, O. Bodenreider, M. R. Boland, I. Georgiev, H. Liu, K. Livingston, et al. The digital revolution in phenotyping. *Briefings in bioinformatics*, page bbv083, 2015.

[27] P. Olson. A massive social experiment on you is under way, and you will love it. `http://www.forbes.com/sites/parmyolson/2015/01/21/jawbone-guinea-pig-economy/`. Accessed: 2016-01-13.

[28] W. Plank. The future of wearables market. `http://www.wsj.com/articles/the-future-of-the-wearables-market-1452736738`. Accessed: 2016-01-20.

[29] A. Pourzanjani, T. Quisel, and L. Foschini. Adherent use of activity trackers is associated with weight loss. *PLOS ONE*, Submitted.

[30] K. E. Powell, A. E. Paluch, and S. N. Blair. Physical activity for health: What kind? how much? how intense? on top of what? *Public Health*, 32(1):349, 2011.

[31] N. Razavian and D. Sontag. Temporal convolutional neural networks for diagnosis from lab tests. *CoRR*, abs/1511.07938, 2015.

[32] A. A. Thorp, N. Owen, M. Neuhaus, and D. W. Dunstan. Sedentary behaviors and subsequent health outcomes in adults: a systematic review of longitudinal studies, 1996–2011. *American journal of preventive medicine*, 41(2):207–215, 2011.

[33] T. Wang. The future of biosensing wearables. rock health. `http://rockhealth.com/2014/06/future-biosensing-wearables`. Accessed: 2016-01-13.

[34] World Health Organization. BMI Classification, Global Database on Body Mass Index, 2006.

[35] Y. Zheng, Q. Liu, E. Chen, Y. Ge, and J. L. Zhao. Time series classification using multi-channels deep convolutional neural networks. In *Web-Age Information Management*, pages 298–310. Springer, 2014.

Predicting frailty in elderly people using socio-clinical databases

Flavio Bertini, Giacomo Bergami, Danilo Montesi
{flavio.bertini2, giacomo.bergami2, danilo.montesi}@unibo.it
Department of Computer Science and Engineering,
University of Bologna, Italy

Paolo Pandolfi
paolo.pandolfi@ausl.bologna.it
Department of Epidemiology, Health Promotion and Risk Communication,
Regional Health Services, Bologna, Italy

Abstract

Life expectancy increases globally and brings out several challenges in almost all developed countries all around the world. Early screening allows to carry out policies to prevent adverse age-related events and to better distribute the associated financial burden on social and healthcare systems. One of the main issues correlated to a longer lifespan is the onset of the frailty condition that typically characterizes who is starting to experience daily life limitations due to cognitive and functional impairment. In literature there are many different definitions of frailty condition, mainly due to the large number of variables that can reduce the autonomy of the elderly. We extend the definition proposed by the Department of Health of the UK National Health Service, with which a frail subject is at risk of hospitalisation or death within a year. In particular, in this paper we propose a predictive model for frailty condition based on 26 variables from 11 socio-clinical databases. The model assigns a frailty risk index in the range 0 to 100 to each subject aged over 65 years old. The risk index is stratified in 5 expected risk classes and allows to carry out different tailored interventions. To evaluate our method, we use a four-year depth dataset of about 100 000 over 65 subjects. The obtained results are compared to a dataset of patients being treated in local health services.

Keywords

Predictive Models, Frailty Condition, Elderly Health, Evolutionary Frailty Models, Healthcare Data Characterization, Health Informatics.

1 Introduction

The population ageing is becoming a real emergency. The phenomenon began more than a century ago in many economically developed countries and more recently, it has involved some others less developed countries as well. In 2030, according to a recent United Nations report [8], the number of people aged 60 years old and over is projected to reach the 16.5% of the total world population (29.2% in more developed regions), equal to 1.4 billion of people. Formally, the population gets older when the fertility rates decrease and/or when the life expectancy increases. The trend is clearly shown in Figure 1. The over 60 population is growing faster than other age groups and in 2030 the number of elderly will more than doubled since 2000 (almost 2.5 times more), as the number of children and adolescents will change relatively little (approximately only 1.2 times more).

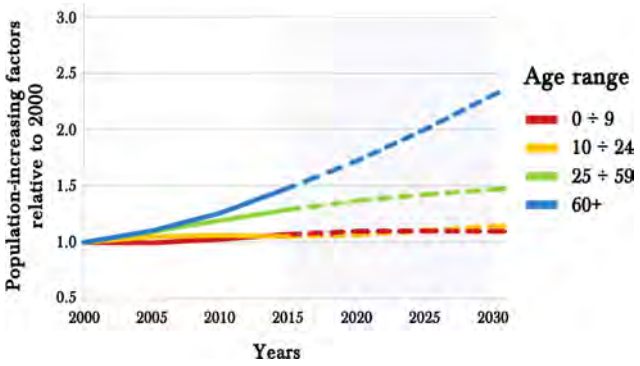


Figure 1: Trend of increment relative to 2000 for each age range of the worldwide population [8].

The ageing of the population has great repercussions on the social and healthcare system, in terms of services and costs. For example, in a comprehensive report [20], Neuman et al. discuss the rising of the health care costs due to a longer lifespans. In 2011, for what concerning the US national social insurance program (Medicare), beneficiaries aged 65 and over covered the

82% of the traditional Medicare population, equivalent to the 78% (292.5 billion dollar) of the total Medicare spending. Moreover, as shown in Figure 2, the higher age classes have higher costs. Thus, the beneficiaries aged 80 years old and over covers 33% (90 billion dollars) of the total spending.

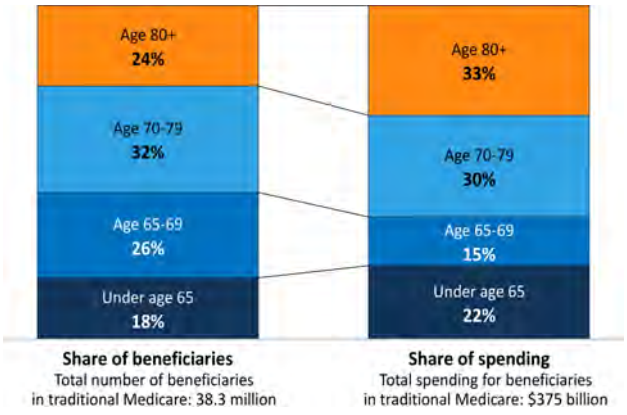


Figure 2: US national social insurance program (Medicare): beneficiaries and spending distribution in 2011, for four different age ranges [20].

In European countries the situation is dramatically similar. In particular, in Italy over 65 represents the 21% of the total population with high health care costs. These numbers identify a quite critical scenario where health care systems necessarily have to adopt countermeasures to meet the growing needs of a greater numbers of elderly. Recently, all health research institutions have agreed on the adoption of early screening and constant monitoring strategies. These policies have a twofold value: mitigate the increasing costs' pressure and prevent adverse events that reduce the older person's autonomy.

The frailty condition represents a limitation of the independence in older subjects and it is characterized by functional and cognitive impairment. Moreover, the major number of frail elderly people has a great impact on the management of the social and health systems. However, frailty condition is quite difficult to be identified, mainly due to the large number of variables from which it depends on. Researchers proposed several definitions of frailty, however there is not yet an overall consensus on it. On which factors, between biomedical and psychosocial, the frailty should depend is the most discussed issue.

In this paper we address the problem of correctly characterization of elderly people according to expected frailty risk. To achieve this goal we extend the definition of frailty condition proposed in [27], with which a frail subject is at risk of hospitalisation or death within a year. In practice, according to this definition of frailty condition, we built a frailty prediction model based on several different socio-clinical databases able to stratify people aged over 65 years old living in Bologna (Emilia-Romagna Region, Italy). We identify 5 expected risk classes that include non-frail, pre-frail and frail condition. The lower is the assigned score the less is the probability that hospitalisation or death occur within a year. Thus, our model can be useful to early detect the frailty condition, to constant monitoring its evolution and to carry out tailored interventions according to expected frailty risk. More generally, the outcome can be exploited to improve the elderly health care and mitigate management costs.

The reminder of the paper is organized as follows. In Section 2, we review literature works related to frailty condition and predictive models for health. In Section 3, we describe the methodology used to build the frailty prediction model. The results achieved by our model are evaluated and discussed in Section 4. Some concluding remarks are made in Section 5.

2 Related works

In this section, we discuss literatures from two different domains, at the intersection of which our work lies. Firstly, in Section 2.1, we present various works concerning frailty condition in older adults. Then, in Section 2.2, we discuss predictive models for health data.

2.1 Frailty condition in elderly people. Typically, older people develop a variety of age-related conditions that contribute to loss autonomy. This phenomenon is well-known as frailty condition, nevertheless its concept has not emerged as a well-defined clinical or social entity. Frailty condition has a great impact on elderly and the lack of a standard definition tends to divide the researchers: those who try to outline the frailty condition with clinical factors and those who also accept to introduce social factors. For example in [13], the author describes the clinical correlates and the biological underpinnings useful to understand the frailty condition. While in [17], they take in account also psychological, social and environmental factors. Recently, almost all researchers agree to consider three domains contributing to frailty condition: physical, cognitive and psychosocial. The link between cognitive decline and frailty condition is widely discussed in [5]. Currently, the main challenge is to define methods and models to detect frailty and measure its severity [6]. In [10], the researchers propose a predictive model for frailty in community-dwelling older adults, based on the following criteria: unintentional weight loss, self-reported exhaustion, weakness, slow walking speed and low physical

activity. An interesting solution has been proposed in [18] by exploiting logistic regression based model. The authors propose a frailty model for cardiac surgery patients in order to decrease the mortality and prolonged institutional care risk. A more general models able to stratify the entire population are described in [27] and [21]. This is important as we identify two critical events to built our model: hospitalization and death. Moreover, according to the most recent literature [26], we exploit a wide range of socio-clinical variables to predict the frailty condition.

2.2 Predictive models in health domain. Reduce mortality rates, detect adverse clinical events and contain healthcare costs represent exciting challenges for researchers. As observed in [4], the current technologies, and especially the predictive model, can produce an imminent revolution in healthcare. For example, the wealth of this research domain is described in [22] and [23]. The authors list the vast array of available data sources: electronic health records, biomedical image, genomic data and biomedical literature, to name a few, and many varieties of aims: personal medicine, healthcare fraud detection, clinical decision support, computer-aided diagnosis. Recently, several interesting studies have been presented. A hybrid approach is proposed in [15], they combine a genetic algorithm to select the features with a logistic regression technique to predict the Alzheimer's disease progress. In [28], the authors compare a top-k stability method to three classic classification methods (support vector machine, logistic regression and random forest) to select features from the electronic health records. The aim is to predict the patient risk in developing a target disease. Currently, data mining are widely use in healthcare. In [16] the authors discuss data mining techniques in major healthcare areas such as evaluation of treatment effectiveness, healthcare management and fraud and abuse detection. The effectiveness of data mining in prediction and decision making in healthcare is discussed in [19]. Different data mining approaches, based on machine learning, are used in [2] in order to characterize metabolic syndrome patients. In [25], the data mining techniques are exploited to diagnose and treat heart disease. Social media provide another fruitful domain, they have gained popularity and can be used for extracting patterns and knowledge. In particular, social media data can be used to inference about population health and public health monitoring. Some interesting attempts in this direction are presented in [1] and [11]. In the first work, the authors try to predict the seasonal influenza epidemics by building a model based on the Google search queries. In the second one, the flu epidemics are monitored through Twitter using a support vector machine. There are several prediction models that can be applied to health data and the choice significantly depends on the outcomes to be predicted. A comprehensive survey is done in [3] and [24]. In particular, logistic regression, binary classification trees and Bayesian models are more adapted to solve a binary classification problem. Moreover, logistic regression is usually preferred in case of noisy data. We exploit these characteristics to build our frailty prediction model for elderly, which is the main focus of our study.

3 Methodology

In this section, we firstly present a brief background in logistic regression. This is important since we used this methodology to build up a predictive model for frailty condition. Then, we describe the databases used to retrieve our data and the variables selected to create our model. Finally, we present the process through which we built our frailty prediction model.

3.1 Logistic regression. Multivariate regression detects the mathematical correlation between the chosen independent variables $X_1, \ldots, X_n$ (the *predictors*) and a dependent variable E (the *expected event*) [12]. This kind of predictive model has some interesting characteristics, that help in understanding the reasons behind our selection. It can mix categorical and continuous predictors, it is more adapted to solve a binary classification problem ("decide if the predictors describe the event E or not") and it is robust enough with respect to noisy data [24]. Moreover, multivariate regression allows to achieve a probability function $P(E|X_1, \ldots, X_n)$ that describes the strength of the whole set of predictors $(X_1, \ldots, X_n)$ in the classification process. The data are intercepted by a sigmoid function that is defined as:

$$P(E|X_1, \ldots, X_n) = \frac{1}{1 + e^{-(\beta_1 X_1 + \cdots + \beta_n X_n)}} \tag{3.1}$$

where the $\beta_1, \ldots, \beta_n$ values represent the regression coefficients. These parameters are tuned during the training process exploiting the maximum likelihood technique. In practice, when $E = 1$ the $\beta_1, \ldots, \beta_n$ combination has to maximize the the likelihood function. Single categorical predictor has a β_i and each continuous predictor is splitted in classes. Each class is represented by some new *dummy variables* with an associated β_i. In the footsteps of [21], we choose the logistic regression model, that is a specific type of multivariate regression. The logistic regression allows to estimate the influence of each single X_i to intercept the event E. In particular, we fix the $X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n$ values and

Database source	Description
Health Registry	It provides the following entries per patient: name, surname, sex, age, Fiscal Code, date of birth, address, residence district, place of birth, nationality. (The Fiscal Code in Italy is similar to the Social Security Number in the US or the National Insurance Number in the UK.)
Hospital Discharge Record (HDR)	This database is build upon the hospitalization records gathered from each public and private hospital in Bologna. Each data record includes the ICD-9-CM codes related to the performed medical interventions and the diagnosed diseases. Also, this dataset provides all the details of acute and post-acute hospital admissions.
Emergency Room Records	Each entrance to either the emergency room or first aid point is collected monthly in this database.
Ambulatory Specialised Assistance (ASA)	This database is updated monthly and contains all the health services provided to the non-patients of the local hospitals. This data source estimates the amount of either clinical exams or specialised assistance services.
Public Home Healthcare (PHH)	This database is updated on a quarterly basis. It provides for each patient demographic, social and healthcare information. In particular, the home health care duration, frequency and costs.
Territorial Pharmaceutical Assistance registry (TPA)	It contains the drug prescriptions presented at drug stores and can be used to trace back some diseases that are not recorded in the HDR database.
Domiciliary Pharmaceutical Dispensing registry (DHD)	In Italy seriously ill patients (e.g. cancer patients) can receive drugs even via a home assistance facility. This database keeps track of the home-delivered medicines and can be used to trace back some diseases that are not recorded in the HDR database.
Death Registry	This database keeps track of the per-year deaths in Bologna.
Mental Health Department registry (MHD)	This database records mental diseases patients.
Care Grants Monitoring System (CGMS)	In this database are recorded senior citizens who receive home care assistance.
Municipal Registry	This database collects social data that characterize each resident, as: income, housing condition, residence district, education level, marital status and emigration date (if any). The residence district determines the area's deprivation.

Table 1: Databases used in the data warehouse reconciliation process.

we can evaluate the influence of X_i as:

$$e^{\beta_i} - 1 = P(E|X_1, \ldots, X_i + 1, \ldots, X_n) - P(E|X_1, \ldots, X_i, \ldots, X_n) \tag{3.2}$$

The equation (3.2) is very important since allows to correctly select the variables that better describe the event E and discard the others.

3.2 Data sources and variables selection. The available data comes from different database belonging to different public bureaus of the Emilia-Romagna Region and Bologna's City Hall. Consequently, we have firstly to carry out a data cleaning and integration phase. In particular, this step is performed through a data warehousing process. The data inconsistencies regarding the different representation and schemas are reconciled, as described in [7]. Then, we carry out a data sources linking phase where similar descriptions are linked together and anomalies are corrected, both manually and automatically via imputation methods (if possible). For example, the inconsistencies of the Fis-

cal Code, similar to a Social Security Number or a National Insurance Number, are corrected in this phase. Also, some pathologies are uniquely derived by the active drug ingredient using TPA ad DHD databases. Table 1 points out the data sources that we have used to create an unique view for our dataset.

The available databases offer a plethora of different variables. Exploiting equation (3.2) we identify the 26 most influential variables among all the other possibilities. The selected variables are described in Table 2. Each single variable weight $(e^{\beta_i} - 1)$, both for categorical and for dummy variables, are plotted in Figure 3. Intuitively, out of all variables the age is the the most predictive. However, our aim is to build a socio-clinical frailty predictive model, for this reason we decide to include both social and clinical parameters to predict the joint event death or hospitalization.

3.3 The predictive model. In this section we describe the process adopted to build the frailty predictive model. In the first phase, we create a two-year depth anonymised cohort of subjects (from January 1st, 2009 to December 31st, 2010) according to these filtering rules:

- we select 63 year old subjects on January 1st, 2009;
- we remove all dead subjects in biennium;
- we remove all emigrated subjects in biennium;
- we remove all non-residents subjects in biennium;
- we remove all subjects without health card.

This filtering process is carried out only to build the model and produces a dataset of 95 368 subjects. For any other evaluations, the frailty predictive model is applied to all over 65 year old subjects.

Then, in a second phase, this cohort is splitted in two well-known sets: the **training set** composed of 63 579 subjects (2/3 of the total) and the **test set** composed of 31 789 subjects (1/3 of the total). This phase is accurately performed in order to create two equivalent and uniformly distributed sets and avoid any overfitting problems. In other words, in each set, the event has the same probability to occur and the predictors are uniformly distributed. We use equation (3.2) to compute the $\beta_1, \ldots, \beta_{26}$ regression coefficients for each variable, on the training set. Then, the regression coefficients are used on the test set to evaluate the accuracy and the calibration of the model. For this reason, we need to define a risk score function as:

$$risk\_score = 100 \cdot P(E|X_1, \ldots, X_{26}) \tag{3.3}$$

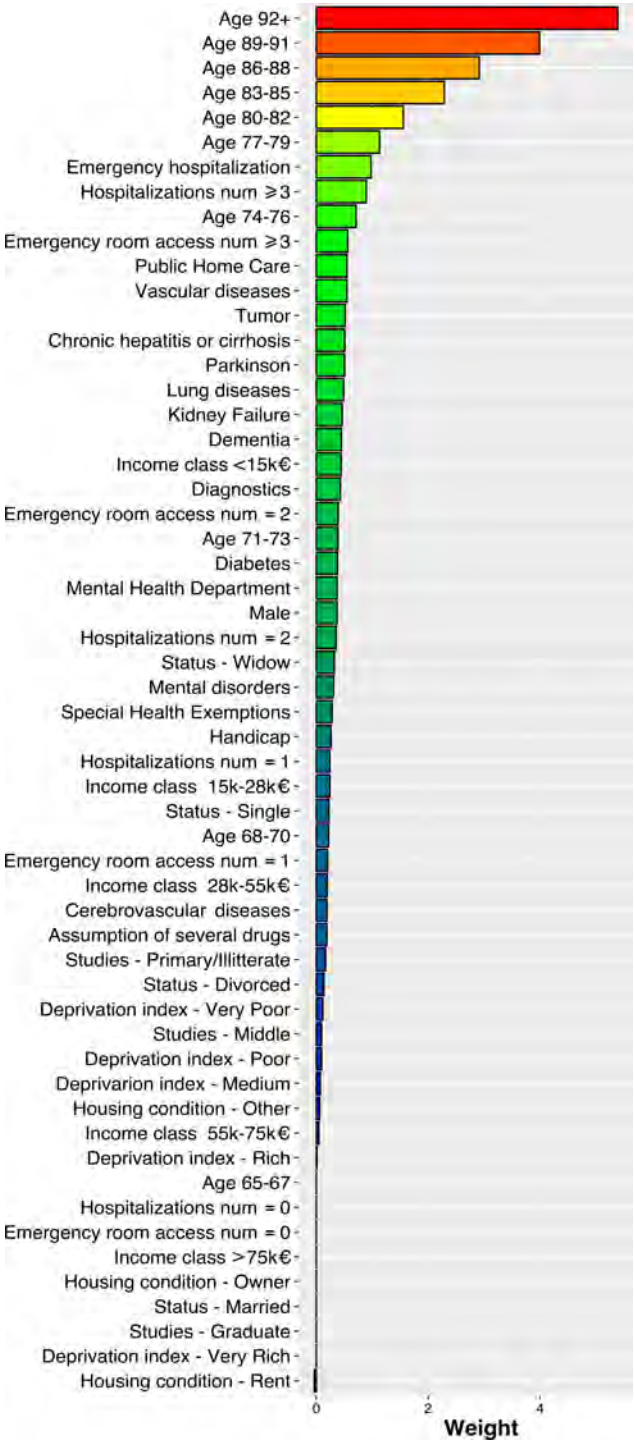


Figure 3: The ordered weights $(e^{\beta_i} - 1)$ of each (categorical and continuous) variable in the final model.

where the $risk\_score$ is a value in the range 0 to 100. The 0 value means a low frailty risk as the 100 value means a high frailty risk of death or hospitalization.

Variable name	Description
Age	It is represented by 10 dummy variables, each of which groups 3 years from 65 to over 92. The motivation is that the weight of each age step increases linearly.
Sex	It is represented by a binary variable.
Lung diseases	It is extracted from the HDR database and checks if the subject has been hospitalized for acute bronchitis, chronic bronchitis, emphysema, asthma, bronchiectasis, extrinsic allergic alveolitis.
Tumor	It is a binary variable that checks if either the subject has been hospitalized for tumour or if he assumes tumour drugs.
Diabetes	It is a binary variable that checks if the subject has been hospitalized for diabetes. We check the DHD and TPA databases.
Parkinson	It is a binary variable that checks if the subject has been hospitalized for Parkinson. We check the DHD and TPA databases.
Chronic hepatitis or cirrhosis	It is a binary variable extracted from Emergency Room Records. It checks if the subject has been hospitalized for hepatitis or cirrhosis.
Dementia	It is a binary variable that checks if the subject has been hospitalized for Alzheimer's disease or psychotic states. We check the HDR database.
Handicap	It is a binary variable that checks if the subject benefits of special health exemptions.
Cerebrovascular Diseases	It is a binary variable that checks if the subject has been hospitalized for occlusion and stenosis of precerebral and cerebral arteries, or benefits from special health exemptions.
Kidney failure	It is a binary variable that checks if the subject has been hospitalized for kidney failure, or benefits from special health exemptions.
Vascular diseases	It is a binary variable that checks if the subject has been hospitalized for acute myocardial infarction, acute and subacute ischemic heart disease, prior myocardial infarction, angina pectoris and other chronic ischemic heart disease, or benefits from special health exemptions. We check the HDR database.
Assumption of multi drugs	It is extracted from TPA databases and checks if the subject assumes more than three different kind of drugs.
Mental Disorders	It is extracted from DHD and TPA databases and checks if the subject mental disorders, or benefits from special health exemptions.
Special Health Exemptions	It is extracted from CGMS database and checks if the subject benefits from special health exemptions.
Mental Health Department	It is a binary variable and checks if the subject is cured by MHD.
Public Home Care	It is a binary variable and checks if the subject receives public health services at home.
Hospitalizations	It is represented by 4 dummy variables and counts the hospitalizations in the two previous years before the evaluation event. The weight increases linearly up to 3 hospitalization, then is constant.
Emergency Room Accesses	It is represented by 4 dummy variables and counts the hospitalizations in the two previous years before the evaluation event. The weight increases linearly up to 3 accesses, then is constant.
Emergency hospitalization	It is a binary variable and checks if there was an emergency admission in the previous month before the evaluation event.
Diagnostic	It is a binary variable and checks if there was a diagnostic event in the previous three months before the evaluation event.
Income class	It represents the average household income and it is represented by 5 dummy variables according to the Italian personal income tax classes.
Housing condition	It represents the housing condition and it is represented by 3 dummy variables: owned, rented and other.
Marital Status	It represents the marital status and it is represented by 4 dummy variables: married, single, widowed and divorced.
Studies	It is grouped in 3 classes: high school diploma or university degree, compulsory education and primary school or illiterate.
Deprivation Index	It represents the neighbourhood average income and it is represented by 5 dummy variables: very poor, poor, medium, reach and very reach.

Table 2: Variables selected for the frailty prediction model.

Intuitively, a lower *risk_score* means a lower probability that hospitalisation or death occur within a year. The frailty prediction model assigns a *risk_score* to each subject that belongs to the test set, in other words, we predict the joint event in 2011. That is the follow-up year.

4 Validation and Evaluation

The validation of our model is done both with a statistical and empirical approach. In the first one, we use some well-known curves and statistical tests to assess the classification process. In the second one, we compare the most-frail subjects with the information stored in GARCIA, that is a local database of frail patients already being treated in local health services.

The statistical validation is performed both on the training and test set. In particular, we compute the area under the ROC curve and we perform the Hosmer and Lemeshow test. The first one allows to evaluate the discrimination capability of the model [29], while the second tests the calibration of the model [14]. On the training set the area under the ROC curve is 0.7681, that means a good discrimination capability in clinical analysis. Hosmer and Lemeshow test returns a score of 0.1099, that means a good statistical significance of the model (the non-significance threshold is usually fixed to 0.05). On the test set the values are slightly lower 0.7641 for the area under the ROC curve and 0.0768 for the Hosmer and Lemeshow test. However, the are still comparable to the previous ones.

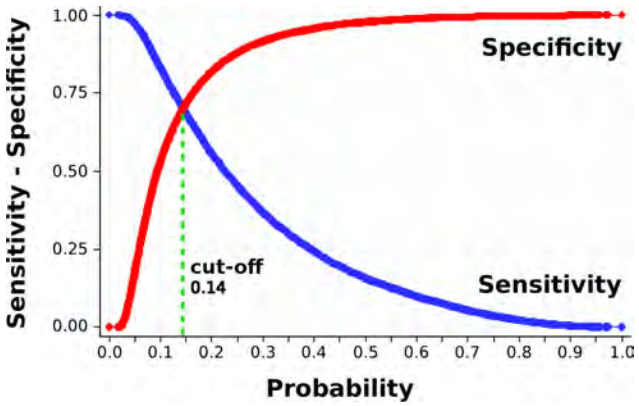


Figure 4: The sensitivity and specificity curves related to the training set values. The intersection point is used to define the first Non-Frail class.

The empirical validation requires the division into classes of the *risk_score*. As shown in Figure 4, the first cut-off value is fixed in the intersection point between the sensitivity and specificity curves (probability $\approx$ 0.14, that is $risk\_score = 14$). This *risk_score* value defines the first Non-Frail class, in other words we discriminate non-frail and frail subjects. Also, we are interested to stratify the others frailty levels. Relying on literature [9], the epidemiology group identifies the others intermediate classes as follows:

- **Non-Frail class**: $risk\_score\ 0-14$
- **Pre-Frail 1 class**: $risk\_score\ 15-29$
- **Pre-Frail 2 class**: $risk\_score\ 30-49$
- **Pre-Frail 3 class**: $risk\_score\ 50-79$
- **Frail class**: $risk\_score\ 80-100$

Epidemiologists noticed that all the subjects belonging to the last frail class completely match the patients in the GARCIA database. Moreover, the classes allow to monitor the evolution over time of the frailty condition and to carry out tailored interventions according to expected risk, which show the usefulness of our model.

The frailty prediction model is applied to the whole cohort of (95 368 subjects). In Table 3 the predicted results are compared to the real data of the follow-up year, that is 2011. We check for each risk class how many events occur. In other words, we empirically assess the prediction capability of the model.

Risk class	Prediction	Event occurred	Event not occurred
Non-Frail (0-14)	60 379	4 380 7.25%	55 999 92.75%
Pre-Frail 1 (15-29)	23 184	5 142 22.18%	18 042 77.82%
Pre-Frail 2 (30-49)	8 000	3 098 38.73%	4 902 61.72%
Pre-Frail 3 (50-79)	3 472	1 931 55.62%	1 541 44.32%
Frail (80-100)	333	261 78.38%	72 21.62%
Total	95 368	14 812 15.53%	80 556 84.47%

Table 3: The comparison between the prediction of our model on the whole cohort (95 368 subjects) and the real data in the follow-up year (2011).

The column "Prediction" (Table 3) represents the distribution of the subjects in each class, that is the outcome of the frailty prediction model. The two columns "Event occurred" and "Event not occurred" (Table 3) represent the number of subjects for which the event occurred or not, that is the real data in the follow-up year. The subjects in the Non-Frail class have from 0% to 14% of probability to be hospitalized or

to die within a year. We observe a real percentage of 7.25%, that represents a good prediction. The situation is similar for each Pre-Frail class. The model slightly overestimate the Frail class with a minimum prediction value of 80% against a real value of 78.38%. This overestimation is preferable, since allows healthcare service to early intercept the weakest elders.

We performed a final evaluation of the frailty prediction model on a four-year depth dataset form 2011 to 2014. In particular, we extracted the data of all over 65 (about 100 000 subjects) and we classified them in the 5 frailty risk class. The numerical results are shown in Table 4. The plotted values, see Figure 5, show an imperceptible fluctuations of the Frail class and a small increase in the Pre-Frail classes. The most significant variation can be observed in the Non-Frail class, this is probably due to demographic fluctuations occurred during 2011. This evaluation remarks the effectiveness of our model in monitoring the evolution over time of the frailty condition.

Risk class	**2011**	**2012**	**2013**	**2014**
Non-Frail (0-14)	70 186 67.40%	61 333 61.24%	62 825 62.93%	60 590 60.63%
Pre-Frail 1 (15-29)	22 782 21.87%	24 106 24.23%	23 364 23.40%	24 126 24.14%
Pre-Frail 2 (30-49)	7 519 7.20%	9 064 9.11%	8 898 8.91%	9 618 9.62%
Pre-Frail 3 (50-79)	3 317 3.18%	4 438 4.46%	4 316 4.32%	5 016 5.02%
Frail (80-100)	324 0.33%	514 0.51%	420 0.42%	570 0.57%
Total	104 128	99 455	99 823	99 920

Table 4: Prediction on four-year depth database. Each row shows the fluctuation (number and percentage) of over 65 in a specific class.

5 Conclusions

Frailty condition in older adults is a crucial research problem and it is particularly meaningful in an ageing society. Frailty does not necessarily coincide with disabilities mostly because, if early detected, it can be reversible. However, if neglected, the frailty condition can lead to complete loss of autonomy in elderly subjects. For these reasons, it is extremely important the early detection and the continuous monitoring of the frailty condition.

In this paper we have proposed a frailty prediction model built on 11 different socio-clinical databases. The

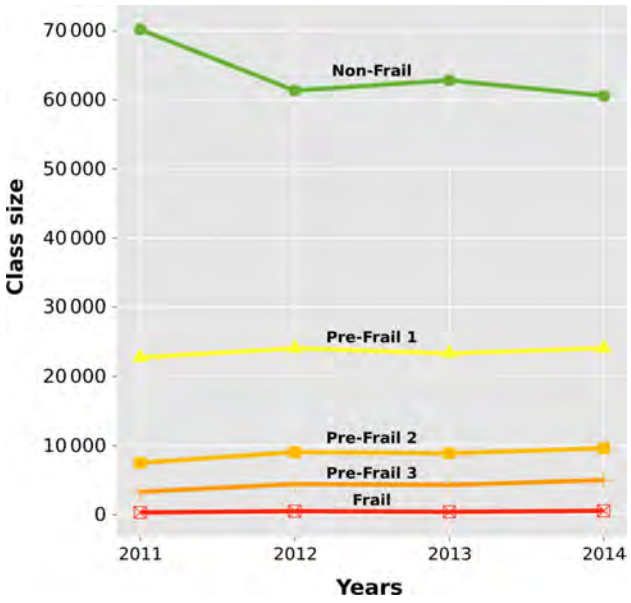


Figure 5: Yearly variation of the frailty classes.

logistic regression based model combines 26 variables and returns a frailty index, according to expected risk of hospitalisation or death within a year. The frailty risk index is used to classify over 65 years old people in 5 different classes. We used a *training set* of 63 579 subjects to tuning the model parameters and a *test set* of 31 789 subjects to evaluate the predictive capability of the model. Then, the model was applied to a four-year depth dataset of about 100 000 subjects aged over 65 years old. The outcome of our frailty predictive model is a characterization of the elderly population in Non-Frail, Pre-Frail and Frail classes. This is useful to carry out tailored health interventions for each of them.

The proposed model is fitted on a specific scenario (Bologna's City Hall) and it uses boolean variables for all clinical pathologies. However, we plan to study an extended version of the model in order to overcome these limits. In particular, we plan to include variables derived from clinical tests and variables able to model the application region of the model. In this way, we intend to address several other problems, for example improve the prediction capability and compare the results from different regions. Also, we plan to investigate other events as claims for durable medical goods or comorbidities.

6 Acknowledgments

This work has been partially supported by the Italian Minister for Instruction, University and Research, OPLON project (OPportunities for active and healthy LONgevity project, SCN_00176).

References

[1] E. Aramaki, S. Maskawa, and M. Morita. Twitter catches the flu: detecting influenza epidemics using twitter. In *Proceedings of the EMNLP Conference*, pages 1568–1576, 2011.

[2] F. Babič, L. Majnarić, A. Lukáčová, J. Paralič, and A. Holzinger. On patients characteristics extraction for metabolic syndrome diagnosis: Predictive modelling based on machine learning. In *ITBAM*, pages 118–132. Springer, 2014.

[3] R. Bellazzi and B. Zupan. Predictive data mining in clinical medicine: current issues and guidelines. *International journal of medical informatics*, 77(2):81–97, 2008.

[4] K. L. Brigham. Predictive health: the imminent revolution in health care. *Journal of the American Geriatrics Society*, 58(s2):S298–S302, 2010.

[5] L. Calzà, D. Beltrami, G. Gagliardi, E. Ghidoni, N. Marcello, R. Rossini-Favretti, and F. Tamburini. Should we screen for cognitive decline and dementia? *Maturitas*, 82(1):28–35, 2015.

[6] A. Clegg, J. Young, S. Iliffe, M. O. Rikkert, and K. Rockwood. Frailty in elderly people. *The Lancet*, 381(9868):752–762, 2013.

[7] M. Dallachiesa, A. Ebaid, A. Eldawy, A. Elmagarmid, I. F. Ilyas, M. Ouzzani, and N. Tang. Nadeef: A commodity data cleaning system. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, SIGMOD '13, pages 541–552, New York, NY, USA, 2013. ACM.

[8] Department of Economic and Social Affairs. World population ageing 2015: Highlights. Technical report, United Nations, 2015.

[9] P. Falasca, A. Berardo, and F. Di Tommaso. Development and validation of predictive MoSaiCo (Modello Statistico Combinato) on emergency admissions: can it also identify patients at high risk of frailty? *Ann. Ist. Super. Sanita*, 47(2):220–228, 2011.

[10] L. P. Fried, C. M. Tangen, J. Walston, A. B. Newman, C. Hirsch, J. Gottdiener, T. Seeman, R. Tracy, W. J. Kop, G. Burke, et al. Frailty in older adults evidence for a phenotype. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, 56(3):M146–M157, 2001.

[11] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014, 2009.

[12] J. F. J. Hair, W. C. Black, B. J. Babin, and R. E. Anderson. *Multivariate Data Analysis*. Pearson, 7th edition edition, 2014.

[13] D. Hamerman. Toward an understanding of frailty. *Annals of internal medicine*, 130(11):945–950, 1999.

[14] D. W. J. Hosmer and S. Lemeshow. *Applied Logistic Regression*. Wiley-Interscience, New York, 1989.

[15] P. Johnson, L. Vandewater, W. Wilson, P. Maruff, G. Savage, P. Graham, L. S. Macaulay, K. A. Ellis, C. Szoeke, R. N. Martins, et al. Genetic algorithm with logistic regression for prediction of progression to alzheimer's disease. *BMC bioinformatics*, 15(Suppl 16):S11, 2014.

[16] H. C. Koh, G. Tan, et al. Data mining applications in healthcare. *Journal of healthcare information management*, 19(2):65, 2011.

[17] F. Lally and P. Crome. Understanding frailty. *Postgraduate medical journal*, 83(975):16–20, 2007.

[18] D. H. Lee, K. J. Buth, B.-J. Martin, A. M. Yip, and G. M. Hirsch. Frail patients are at increased risk for mortality and prolonged institutional care after cardiac surgery. *Circulation*, 121(8):973–978, 2010.

[19] B. Milovic. Prediction and decision making in health care using data mining. *International Journal of Public Health Science (IJPHS)*, 1(2):69–78, 2012.

[20] T. Neuman, J. Cubanski, J. Huang, and A. Damico. The rising cost of living longer: Analysis of medicare spending by age for beneficiaries in traditional medicare. *The Henry J. Kaiser Family Foundation*, 2015.

[21] P. Pandolfi, P. Marzaroli, M. Musti, E. Stivanello, and N. Collina. Un modello statistico previsionale per misurare la fragilità. In G. Cavazza and C. Malvi, editors, *La fragilità degli anziani. Strategie, progetti, srtrumenti per invecchiare bene.*, chapter 2, pages 27–36. Maggioli Editore, 2014.

[22] N. Peek, C. Combi, R. Marin, and R. Bellazzi. Thirty years of artificial intelligence in medicine (aime) conferences: A review of research themes. *Artificial intelligence in medicine*, 65(1):61–73, 2015.

[23] C. Reddy and C. Aggarwal. chapter An Introduction to Healthcare Data Analytics, pages 1–18. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series. Chapman and Hall/CRC, 2015.

[24] C. Reddy and Y. Li. chapter A Review of Clinical Prediction Models, pages 343–378. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series. 2015.

[25] M. Shouman, T. Turner, and R. Stocker. Using data mining techniques in heart disease diagnosis and treatment. In *Electronics, Communications and Computers (JEC-ECC), 2012 Japan-Egypt Conference on*, pages 173–177. IEEE, 2012.

[26] S. A. Sternberg, A. W. Schwartz, S. Karunananthan, H. Bergman, and A. Mark Clarfield. The identification of frailty: a systematic literature review. *Journal of the American Geriatrics Society*, 59(11):2129–2138, 2011.

[27] D. Wennberg, M. Siegel, B. Darin, N. Filipova, R. Russell, L. Kenney, K. Steinort, T.-R. Park, G. Cakmakci, J. Dixon, N. Curry, and J. Billings. Combined predictive model: final report and technical documentation. Technical report, London: Department of Health, The Kings Fund, NYU, Health Dialogue, 2006.

[28] J. Zhou, J. Sun, Y. Liu, J. Hu, and J. Ye. Patient risk prediction model via top-k stability selection. In *SIAM Conference on Data Mining*. SIAM, 2013.

[29] M. H. Zweig and G. Campbell. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. In *Clinical Chemistry*, 1993.

Discriminant Word Embeddings on Clinical Narratives

Paula Lauren* Guangzhi Qu† Feng Zhang‡

Abstract

Clinical narratives are inherently complex due to the medical domain expertise required for content comprehension. In addition, the unstructured nature of these narratives poses a challenge for automatically extracting information. In natural language processing, the use of word embeddings is an effective approach to generate word representations. Word embeddings generated from a neural language model have shown the ability to capture context as well as syntactic and semantic properties. In this work we propose to use a log-linear model along with Linear Discriminant Analysis to acquire lower dimension space features for learning. Experimental results on clinical texts are given that indicate improved performance in accuracy especially for the minority classes.

1 Introduction

A clinical narrative consists of the written text that documents the encounter between a patient and a healthcare professional (i.e., a physician or nurse). The data captured in clinical narratives can be used for enabling quality-assessment programs [4] for patient care such as patient safety [18], improving surveillance on infectious diseases [28], upholding evidence-based medicine [6], supporting clinical trials [10], and assisting with other clinical research initiatives [29]. The free-form textual nature of these narratives allow clinicians the ease of input. However, the lack of structure of these notes make it difficult to easily extract knowledge from an information system perspective [39]. There are several medical codes such as the *International Classification of Diseases, Tenth Revision, Clinical Modification* (ICD-10-CM)[1] for documenting diagnosis and procedure information and the *International Classification of Primary Care* (ICPC)[2] code for capturing key elements from a healthcare encounter. These codes are typically manually entered and provide the reason for the encounter, diagnosis or problem, and the process of care. The ability to automatically assign these codes using just the clinical narratives would support the initiatives mentioned as well as reduce human error.

In recent years, the use of word embeddings have shown remarkable performance on classification tasks such as sentiment analysis [26], machine translation [44], and determining semantically related words such as in word analogy [24] and word similarity [36] tasks. The classic word embeddings model is Latent Semantic Analysis (LSA) [23], which uses a Singular Value Decomposition (SVD) [15]. Traditionally, classic language models are n-gram models, which essentially count word occurrences from training data. Improvements to the n-gram model can be made with smoothing, caching, skipping and sentence-mixture models and clustering [16]. Word embeddings take into account context for determining meaning based on distributional hypothesis [42]. It is well known that context provides an important source of information in determining meaning. Furthermore, contextual information provides a good approximation to word meaning because similar contextual distributions are a predictor of semantically similar words [32]. Word embeddings can also be based on a Neural Network Language Model (NNLM) [3]. Neural language models use a neural network to generate word vectors in a language modeling task through nonlinear neural networks [26]. The essential idea is that a NNLM represents each word as a d-dimensional vector of real numbers, and the vectors that have a close proximity to each other are shown to be semantically related. A NNLM is typically constructed using a single hidden layer neural network and is regarded as an instance of a vector space model (VSM) [41]. A NNLM learns which words from the vocabulary are more likely to appear after a given word sequence. Essentially, the NNLM learns the next word's probability distribution [42]. Word embeddings represent the meaning of a word with a vector using a numerical representation for keeping track of the number of times that the word occurred in various contexts across a corpus of documents.

Multivariate statistical analysis methods such as Principal Component Analysis (PCA) [19] and Linear Discriminant Analysis (LDA) [14] also known as Fisher's Linear Discriminant, entail the spectral decomposition

*Computer Science and Engineering Department, Oakland University, Rochester MI 48309, Email: palaure2@oakland.edu.

†Computer Science and Engineering Department, Oakland University, Rochester MI 48309, Email: gqu@oakland.edu.

‡School of Computer Science, China University of Geosciences, Wuhan, 430074, China, Email: jeff.f.zhang@gmail.com.

[1] http://www.cdc.gov/nchs/icd/icd10cm.htm

[2] https://www.nlm.nih.gov/research/umls/sourcereleasedocs

of a positive-definite kernel [1]. The application of PCA results in a low-dimensional subspace for explaining most of the variance from the data. LDA determines a separating hyperplane for data classification by seeking to achieve maximum class discrimination by maximizing *between-class* distances and minimizing *within-class* distances simultaneously. LDA defines a projection that makes the *within-class* scatter small and the *between-class* scatter large resulting in compact and well separated clusters. A key distinction between these two methods is that LDA takes the classification aspect into account by using the target labels, PCA does not. Another spectral method is Canonical Correlation Analysis (CCA) [17]. With CCA, two or more views of the data are created, and they are all projected into a lower dimensional space which maximizes the correlation between the views. In the NLP domain, PCA and CCA have been used in the generation of word embeddings for dimensionality reduction, typically these are not based on a NNLM. A non-neural language modeling approach relies on matrix factorization techniques that generate word embeddings based on spectral methods. The CCA spectral method has been used to derive vector representations of words resulting in Eigenwords [12]. The use of PCA in word embeddings relies on the Hellinger distance due to the discrete distributional nature of word co-occurrence statistics then applies SVD to the co-occurrence matrix [25].

In this work, we apply spectral methods to word embeddings generated from a neural language model to help with feature extraction for text classification. Specifically, we apply LDA, PCA, and CCA to the Continuous Skip-gram Log-Linear Model [31] that renders a reduced feature space for classification and report on the results. These techniques are applied to the text classification of a highly imbalanced clinical corpus. The results achieved on the dataset showed an overall improvement especially for the minority classes in comparison to PCA and CCA, methods that have been used with word embeddings in the research literature.

This rest of the paper is organized as follows: Section 2 provides the necessary background for this research describing the log-linear neural language model, specifically the Continuous Skip-gram Model, and Linear Discriminant Analysis. Section 3 describes the process of the methodological approach, Section 4 describes the experiments conducted along with evaluation results. Section 5 is the discussion and section 6 concludes the work.

2 Background

This section will describe the necessary background for understanding this paper which entail language models, n-gram language models, neural language models, word embeddings with a log-linear model, specifically the Skip-gram model, and Linear Discriminant Analysis.

2.1 Language, N-gram and Neural Language Models Statistical language models estimate the probability distribution of various linguistic units such as words, sentences, and entire documents [40]. The objective of language modeling is to estimate the likelihood that a specific sequence of linguistic units will appear in a corpus. The most common language model is the n-gram language model, which is used to compute the statistics of how likely words are to follow each other. N-grams are consecutive word patterns, such as bigrams, trigrams, 4-grams, etc. The goal of a language model is to compute the probability of a sentence or a sequence of words and also includes the related task of predicting an upcoming word [20]. To formulate the n-gram language model, let W be a string of words: $W = w_1, w_2, ..., w_n$. Let the probability of this word sequence be denoted by the joint probability $P(W)$. The estimation of $P(W)$ is determined by the text and how many times W occurs. Typically, long word sequences will not occur in the text, so the computation of $P(W)$ is broken down into smaller steps for which sufficient statistics are collected and probability distributions are estimated [21]. The probability of a sequence of words for an n-gram model can be obtained from computing the probability of each word given the context of words preceding it, using a probabilistic chain rule:

$$P(w_1, w_2, w_3, ..., w_n) = P(w_1)p(w_2|w_1)P(w_3|w_1, w_2)$$
$$...P(w_n|w_1, ..., w_{n-1}) \tag{2.1}$$

The chain rule is applied to determine the joint probability of words in a sequence. In order to estimate these conditional probabilities, the history is limited to $n-1$ words [21], more formally described as:

$$P(w_1, w_2, ..., w_n) = \prod_{i=1}^{n} P(w_i|w_1, w_2, ..., w_{i-1})$$
$$\approx \prod_{i=1}^{n} p(w_i|w_{i-(n-1)}, ..., w_{i-2}, w_{i-1}) \tag{2.2}$$

Estimating these probabilities relies on Maximum Likelihood Estimation (MLE) from relative frequencies [27]. The MLE for the joint probability and the conditional probability are equation 2.3 and equation 2.4, respectively:

$$P_{MLE}(w_1, w_2...w_n) = \frac{count(w_1, w_2...w_n)}{N} \tag{2.3}$$

(2.4)
$$P_{MLE}(w_n|w_1, w_2...w_{n-1}) = \frac{count(w_1, w_2...w_n)}{count(w_1, w_2...w_{n-1})}$$

where *count* refers to the frequency of the N-gram from the training set and N denotes the total number of training instances. There are several limitations of n-grams, such as the problem of representing patterns over more than a few words. In addition, n-grams only consider exact matches despite similar context history. MLE is regarded as unstable due to data sparsity, which is a problem with higher order n-grams [16]. The reason why data sparsity is a major problem in NLP is due to the fact that language is comprised of rare word sequences, which makes it virtually impossible to model all possible string variations of words. Using a large training corpus or assigning non-zero probabilities to unseen n-grams, known as smoothing, is an approach for dealing with data sparsity. There are numerous smoothing techniques, an empirical study on a large number of these smoothing methods have been investigated by Chen and Goodman [7].

In a neural language model, $P(W)$ is generated by a neural network [8]. Word vectors that are learned using a neural language model are trained to predict the next word in the sentence given the preceding words. The use of neural networks for generating language models resolves the discrete property of n-gram language models by providing a continuous representation of text.

2.2 Log-Linear Model using the Continuous Skip-gram for Word Embeddings A log-linear model entails word vectors that are trained with a neural network that has a single hidden layer, based on the inner product between two word vectors. The Skip-gram model, also known as Skip-gram, is a specific type of log-linear model. Its objective is to predict the neighboring words or context of the word, given a word. That is, given a window size of n words around a word w, the Skip-gram model predicts the neighboring words and maximizes the words that appear within n words of w. The Skip-gram model is an unsupervised feature learning algorithm, which uses a log-linear objective function. Referred to as an expanded version of the n-gram model, Skip-gram lacks the consecutive aspect of n-grams by allowing tokens to be skipped, this resolves the data sparsity problem. The Skip-gram model consists of three layers: an input layer, a projection layer, and an output layer as illustrated in Figure 1. A log-linear classifier is used to predict words in a certain range before and after the current word that is used as the input. Unlike other neural network-based language models, the Skip-gram model does not utilize the nonlinear hidden layer which

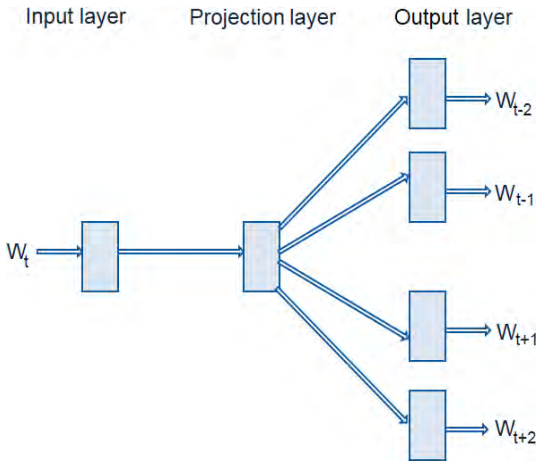


Figure 1: The Continuous Skip-gram model. This model predicts which words will be in the context $w_{t\pm c}$, given the input word w_t.

results in a decrease in computational complexity. The training objective of the Skip-gram model [31] entails predicting the surrounding context, given the current word. To represent formally, given the training words $w_1, w_2, ..., w_T$, the following objective function is maximized:

$$P = \frac{1}{T}\sum_{t=1}^{T}\left(\sum_{-c\leq j\leq c, j\neq 0} \log p(w_{t+j}|w_t)\right) \quad (2.5)$$

where the outer summation covers all the words contained in the training corpus where T is the total number of words. The inner summation spans from $-c$ to c that computes the log probability of predicting the word w_{t+j} given the center word w_t. The basic Skip-gram equation defines $p(w_{t+j}|w_t)$ using the softmax function which ensures that all of the outputs sum to 1. This valid probability distribution is defined as:

$$p(w_{t+j}|w_t) = \frac{\exp\left(vec'^{T}_{w_{t+j}} vec_{w_t}\right)}{\sum_{v=1}^{V}\exp\left(vec'_{w}{}^{\top} vec_{w_t}\right)} \quad (2.6)$$

where V denotes the number of words in the vocabulary, vec_w and vec'_w are the input and output vector representations of w. The Skip-gram model has been made more efficient by approximating softmax by using a normalized hierarchical softmax [34], [30] objective function. This approach approximates the probability distribution with a Huffman Binary Tree for the output layer with the words as its leaves and each node representing the relative probabilities of its child nodes. This

approach significantly improves the training time of the model by resulting in a reduction of computational complexity for $\log p(w_{t+j}|w_t)$ [35]. The Skip-gram model is trained using stochastic gradient descent [38] over the entire training corpus resulting in a distributional representation of words, these are the word embeddings.

2.3 Linear Discriminant Analysis Linear Discriminant Analysis (LDA) is a popular feature extraction technique from statistical pattern recognition that finds a linear combination of features that separates two or more classes [14]. LDA is advantageous in dimensionality reduction, exploratory data analysis and data visualization. As stated in the introduction, LDA works by finding two scatter matrices specified as *between class* and *within class* scatter matrices. Both of these aforementioned scatter matrices seek to separate distinct classes by maximizing their *between class* separability while minimizing their *within class* variability. Let M_j be the total number of samples in class j, so $M = M_1 + M_2 + M_3 + ... + M_c$. Let the mean be denoted as $\bar{x}_j$ for each class j and the mean for M, the entire dataset, as $\bar{x}$. Let $x_{j,k}$ be the m-dimensional pattern k from class c_j, so :

$$\bar{x}_j = \frac{1}{M_j} \sum_{k=1}^{M_j} x_{j,k} \tag{2.7}$$

$$\bar{x} = \frac{1}{M} \sum_{j=1}^{c} M_j \bar{x}_j = \frac{1}{M} \sum_{j=1}^{c} \sum_{k=1}^{N_j} x_{j,k} \tag{2.8}$$

Let $\mathbf{S}_j$ be the sample covariance matrix, $\mathbf{S}_b$ to be the *between class* scatter matrix and $\mathbf{S}_w$ to be the *within class* scatter matrix:

$$\mathbf{S}_j = \frac{1}{M_j - 1} \sum_{k=1}^{M_j} (x_{j,k} - \bar{x}_j)(x_{j,k} - \bar{x}_j)^\top \tag{2.9}$$

$$\mathbf{S}_b = \sum_{j=1}^{c} M_j (\bar{x}_j - \bar{x})(\bar{x}_j - \bar{x})^\top \tag{2.10}$$

$$\mathbf{S}_w = \sum_{j=1}^{c} (M_j - 1)\mathbf{S}_j = \sum_{j=1}^{c} \sum_{k=1}^{M_j} (\bar{x}_{j,k} - \bar{x_j})(\bar{x}_{j,k} - \bar{x})^\top \tag{2.11}$$

The main objective of LDA is to find the projection matrix $\mathbf{W}$ that will maximize the ratio of the determinant for $\mathbf{S}_b$ to the determinant for $\mathbf{S}_w$. This ratio is known as Fisher's criterion [13], which tries to find the projection that maximizes the variance of the class means and minimizes the variance of the individual classes [5].

$$\mathbf{W} = \arg\max_{\mathbf{W}} \frac{|\mathbf{W}^T \mathbf{S}_b \mathbf{W}|}{|\mathbf{W}^T \mathbf{S}_w \mathbf{W}|} \tag{2.12}$$

The optimization problem of equation (2.12) is maximized when the projection matrix $\mathbf{W}$ is composed of the eigenvectors $\mathbf{S}_w^{-1}\mathbf{S}_b$. Furthermore, $\mathbf{W}$ is the solution to the following eigensystem problem [11]:

$$\mathbf{S}_b \mathbf{W} - \mathbf{S}_w \mathbf{W} \Lambda = 0 \tag{2.13}$$

Multiplying $\mathbf{S}_w$ by its inverse $\mathbf{S}_w^{-1}$ in equation (2.13) yields:

$$\mathbf{S}_w^{-1}\mathbf{S}_b \mathbf{W} - \mathbf{S}_w^{-1}\mathbf{S}_w \mathbf{W} \Lambda = 0$$

$$\mathbf{S}_w^{-1}\mathbf{S}_b \mathbf{W} - \mathbf{W} \Lambda = 0 \tag{2.14}$$

$$(\mathbf{S}_w^{-1}\mathbf{S}_b)\mathbf{W} = \mathbf{W}\Lambda$$

where $\mathbf{W}$ and Λ are the eigenvectors and eigenvalues of $\mathbf{S}_w^{-1}\mathbf{S}_b$, respectively. Equation (2.14) states that if $\mathbf{S}_w$ is a non-singular matrix then the Fisher's criterion that is described in equation (2.12) is maximized when the projection matrix $\mathbf{W}$ is composed of the eigenvectors of $\mathbf{S}_w^{-1}\mathbf{S}_b$ with at most $(C-1)$ nonzero eigenvalues [5].

3 Methodology

This section describes the process used from obtaining the clinical narratives for this research, the preprocessing that was done with the clinical texts, generating the word embeddings from the narratives, applying the spectral methods and the classification approach for classifying the narratives according to their respective surgical operation code.

3.1 Dataset Description A total of 2,944 clinical narratives were used for this research that were made available from a partnership with William Beaumont Hospital. The clinical narratives are all related to hip surgery and each record is labeled with an associated surgical operation code for each clinical narrative and there are five total codes, or classes in the clinical corpus. Note that this system was built without any domain knowledge in the medical domain. This includes comprehension of the narrative as well as the explicit meaning of the surgical codes. As illustrated in Fig. 2, this data is highly imbalanced. The highest class count belongs to surgical code *27130* which had a total count of 2,252 records. The lowest class count belongs to code *27138* which has a total of 62 records.

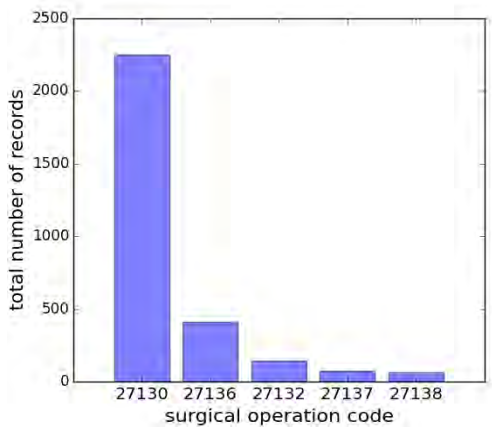


Figure 2: The counts of the clinical narratives for each surgical code class, illustrating skewness of the codes.

3.2 Data Preprocessing The skewed data noted in the previous section pertaining to the classes, also extend to imbalance in terms of character length for the narratives. The average character length was 4,746 characters for individual narratives from the collection of clinical texts. The disparity is especially reflected in the maximum and minimum character lengths for the narratives. The maximum character length was at 22,234 characters and the minimum character length at 169 characters.

Historically, preprocessing text in NLP text mining tasks typically involve various steps such as stemming and stopword removal. These prepocessing steps are not really needed with feature learning utilizing a neural language model such as the Skip-gram model. The reason for this is due to the core objective that context is taken into account when rendering the distributed word representation, or word embeddings. Typically in linguistic processing, multi-words are taken into account in that they are treated as one phrase or term. Multi-words are especially notorious in medical corpora. For example, the medical phrase *glucose metabolism disorders* is represented as: *glucose_ metabolism_disorders* in the vocabulary [33]. No further preprocessing to represent multi-words holistically as one term was done in this research. For our dataset and the methods used, the formation of multi-word terms may have been redundant due to the context-preserving aspect of neural language modeling. We relied on the capability of the model to account for terms based on the contextual usage across the corpus.

During preprocessing, all non-letters were removed from the text and all upper case words converted to lower case. Tokenization was further applied where each narrative was split by the sentences. The outcome was a list of sentences for the entire clinical corpus, a total of 126,585 sentences had been generated to be used for further training. There was also a significant amount of variation in each sentence with each sentence itself containing a list of words. The smallest sentence consisted of one word and the largest sentence consisted of 137 words. The Holdout [22] method with stratification was used for training and testing, with a 75% and 25% respective split using the stratified holdout method. Stratification enabled the respective training and test sets to have a distribution of classes that is representative in Fig. 2. Training had been performed using 2,208 instances and testing done on 736 instances. The training set consisted of the following number of instances for each class (surgical code): 27130: 1678 instances, 27136: 315 instances, 27132: 116 instances, 27137: 55 instances and 27138: 44 instances.

3.3 Generating the Word Embeddings The Skip-gram neural language model was used for generating the word embeddings. The Word2Vec[3] library that has been extended in the Gensim[4] library had been utilized for this research. There are various parameters that can be adjusted such as the feature dimensions, minimum word count, and the window size. The parameters were determined empirically. The window size determines how many words that the training algorithm should use, this represents the window or range of surrounding words that give the word its context. The minimum word count is the minimum threshold that must be met in order for the word to appear in the vocabulary. Various word vector dimensionality values were used to determine the number of features that the model should produce. A feature vector size of 2000, context window of 10 and minimum word count of 30 were used in generating the word embeddings. During preliminary experiments these values gave the best results on the dataset used for this research. The hierarchical softmax training method was used with the Skip-gram model.

The words from the vocabulary (generated from the Skip-gram model) that matched the words in each clinical narrative varied across the clinical corpus. A *representative word vector* for each clinical narrative was used where the word vectors from the vocabulary that matched the words in each clinical narrative were averaged together. That is, the word vectors that matched were added together and then divided by the total number of words from the clinical narrative that

[3] https://code.google.com/p/word2vec/
[4] https://radimrehurek.com/gensim/models/word2vec.html

Table 1: Results from LDA applied to word embeddings matrix using three classifiers.

	kNN			SVM			MLP		
Class	*Precision*	*Recall*	F_1	*Precision*	*Recall*	F_1	*Precision*	*Recall*	F_1
27130	0.98	0.95	0.96	0.97	0.96	0.96	0.94	0.97	0.95
27132	0.67	0.83	0.74	0.69	0.84	0.76	0.78	0.69	0.73
27136	0.89	0.90	0.90	0.93	0.90	0.92	0.89	0.85	0.87
27137	0.74	0.88	0.80	0.74	0.88	0.80	0.87	0.68	0.76
27138	0.56	0.64	0.60	0.56	0.69	0.62	0.75	0.56	0.64

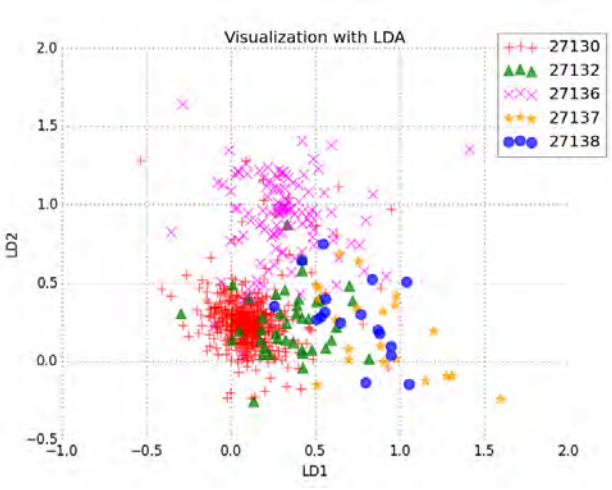


Figure 3: LDA applied to word embeddings matrix.

matched the vocabulary. A total of 2208 *representative word vectors* were used for training and 736 for testing, these are the word embeddings for the training and test sets.

3.4 Applying the Spectral Methods The word embeddings generated for the training set entailed a p x n matrix of 2000 x 2208 with p being the number of features (the word vector) and n being the total number of training instances. Similarly, for the test set but with a p x n matrix of 2000 x 736. Reducing the feature space for the training word embeddings ensures that there are sufficient statistics for estimating the training model. After normalizing the word embeddings, to achieve a reduced representative feature space we applied LDA, PCA and CCA to the training set word embeddings where several classification methods could be utilized effectively.

4 Experiments and Results

This section describes the experiments implemented to reduce the feature space and the classification of the surgical operation codes using the reduced feature space training sets with k-Nearest Neighbor (kNN) [37], Support Vector Machine (SVM) [9] and an artificial neural network, specifically a Multilayer Perceptron (MLP) [43]. Evaluation and the results are also presented in this section. Results are evaluated using standard machine learning evaluation measures using Precision and Recall along with the F_1 score applied to each class. These equations are as follows where TP is *True Positive*, FP is *False Positive* and FN is *False Negative*:

$$Precision = \frac{TP}{TP + FP} \tag{4.15}$$

$$Recall = \frac{TP}{TP + FN} \tag{4.16}$$

$$F_1 = 2 \,.\, \frac{Precision \,.\, Recall}{Precision + Recall} \tag{4.17}$$

4.1 Applying LDA to Word Embeddings In order to apply LDA to the word embeddings matrix requires computing the inverse by multiplying $\mathbf{S}_w$ by its inverse $\mathbf{S}_w^{-1}$ as noted in equation (2.14) from Section 2.3. Due to the nonsingularity of the matrix generated from this particular corpus necessitated the use of a pseudoinverse, the Moore-Penrose pseudoinverse [2] had been used during the computation. The application of LDA to the word embeddings resulted in a reduced feature space of four features. As stated in Section 2.3, LDA results in $C - 1$ feature projections. Figure 3 shows a visualization of LDA which in two dimensions shows clustering for several of the classes. For the classification task, the best parameters were determined empirically and are as follows: kNN with k=3 nearest neighbors, SVM with a Radial Basis Function (RBF) and a single layer neural network using stochastic gradient descent with one hidden layer and 400 neurons, maximum epochs set at 500. Results for the classification methods applied are in Table 1. Using LDA applied to word embeddings shows an overall consistency in accuracy in the F_1 score across all three classification methods with no more than a five percent differential in the F_1 score which measures individual class performance.

Table 2: Results from PCA applied to word embeddings matrix using three classifiers.

	kNN			SVM			MLP		
Class	*Precision*	*Recall*	F_1	*Precision*	*Recall*	F_1	*Precision*	*Recall*	F_1
27130	0.97	0.93	0.95	0.96	0.96	0.96	0.97	0.97	0.97
27132	0.33	0.50	0.40	0.67	0.73	0.70	0.71	0.71	0.71
27136	0.93	0.90	0.91	0.90	0.93	0.92	0.93	0.93	0.93
27137	0.53	0.63	0.57	0.74	0.64	0.68	0.52	0.63	0.57
27138	0.19	0.43	0.26	0.50	0.47	0.48	0.57	0.50	0.53

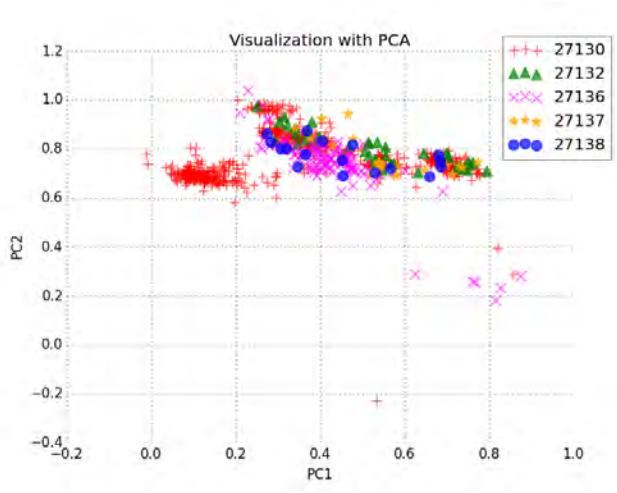


Figure 4: PCA applied to word embeddings matrix.

4.2 Applying PCA to Word Embeddings The application of PCA to the word embeddings matrix resulted in 180 feature projections, which explained 99.07% of the variation in the training set. Classification for the word embeddings with PCA applied was performed on 180 features. Figure 4 shows a visualization of PCA which in two dimensions, the first and second principal components explain 49.14% percent of the variation in the training data. For the classification task, the best parameters were determined empirically and are as follows: kNN with k=4 nearest neighbors, SVM with a polynomial kernel and a single layer neural network using stochastic gradient descent with one hidden layer and 600 neurons, maximum epochs set at 1000. Results for the classification methods applied are in Table 2. With PCA there is more variation specifically with the kNN classifier on the minority classes. It seems that SVM and the neural network have better performance accuracy and are more consistent with each other with an eleven percent variation in classifier accuracy on the minority class F_1 score.

4.3 Applying CCA to Word Embeddings With CCA, the optimal feature space entailed 120 dimensions. Classification for the word embeddings with CCA applied was performed on 120 features. Figure 5 shows a visualization of CCA in two dimensions. For the classification task, the best parameters were determined empirically and are as follows: kNN with k=3 nearest neighbors, SVM with a polynomial kernel and a single layer neural network using stochastic gradient descent with one hidden layer and 600 neurons, maximum epochs set at 1000. Results for the classification methods applied are in Table 3. With CCA there is more variation specifically with the kNN classifier on the minority classes. Just as with applying PCA to word embeddings, it seems that SVM and the neural network have better performance accuracy and are more consistent with each other, showing an eleven percent variation in classifier accuracy performance with the F_1 score on the minority class.

5 Discussion

Applying LDA to word embeddings appears to provide an overall improvement especially with the minority classes. However, using LDA for resolving high-dimensionality is a powerful method but it does have a limitation in that the total number of features that can be extracted are at most $C-1$ features. In this research, the worst F_1 score occurred with the greatest minority class from our datatset where only 44 records were available during training. It would be interesting to see if the LDA feature limitation permitting only $C-1$ feature projections could be addressed, possibly relaxing this constraint could improve the performance for the minority class even greater.

For this research we used a non-supervised approach for the feature learning using the Skip-gram model and applied LDA which is a supervised approach that reduced the dimensions and helped with classification. The clinical data used for this data had been manually annotated with labels, requiring a human effort for adding the correct hip-surgery code to each narrative. We understand that most clinical data is unlabeled so a clustering approach would need to be explored for

Table 3: Results from CCA applied to word embeddings matrix using three classifiers.

	kNN			SVM			MLP		
Class	*Precision*	*Recall*	F_1	*Precision*	*Recall*	F_1	*Precision*	*Recall*	F_1
27130	0.97	0.92	0.94	0.98	0.95	0.96	0.96	0.96	0.96
27132	0.31	0.61	0.41	0.51	0.74	0.61	0.59	0.58	0.58
27136	0.92	0.93	0.92	0.93	0.93	0.93	0.92	0.90	0.91
27137	0.53	0.53	0.53	0.68	0.72	0.70	0.83	0.79	0.81
27138	0.00	0.00	0.00	0.44	0.50	0.47	0.50	0.62	0.56

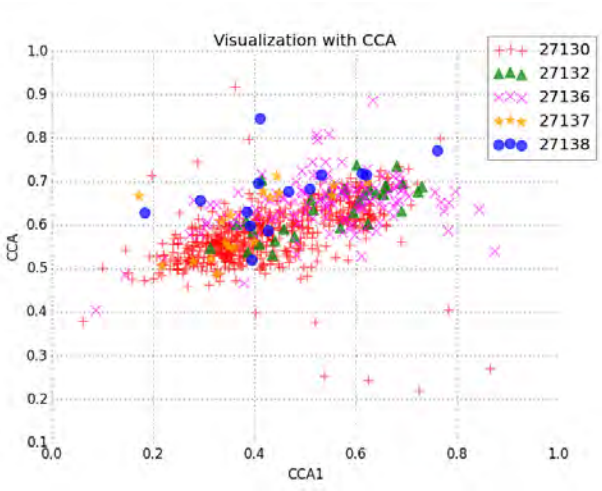


Figure 5: CCA applied to word embeddings matrix.

arriving at a representative cluster for each class. Considering that we have the labels for this clinical corpora, it would be interesting to repeat this experiment using a clustering approach to see how a complete unsupervised approach compares to the semi-supervised approach that we used in this paper. Also, for parameter estimation for the word embeddings and the classification methods we utilized an empirical approach. In future work, we intend to utilize cross validation to systematically arrive at the parameters.

6 Conclusion

Statistical pattern recognition techniques such as LDA, PCA and CCA are capable of significantly reducing dimensionality. Applying these spectral methods to word embeddings are especially advantageous. To the best of our knowledge, LDA has not been applied to word embeddings in the existing research. We have demonstrated with this research that the application of LDA to word embeddings shows an overall improvement with the minority classes in comparison to the more utilized PCA and CCA methods that have been used with word embeddings. In applying LDA to the word embeddings from the clinical corpus, it is especially remarkable that 2000 features from the word embeddings matrix had been reduced to four representative features for classification and for visualization.

References

[1] M.-A. Belabbas and P. J. Wolfe. Spectral methods in machine learning and new strategies for very large datasets. *Proceedings of the National Academy of Sciences*, 106(2):369–374, 2009.

[2] A. Ben-Israel and T. N. Greville. *Generalized inverses: theory and applications*, volume 15. Springer Science & Business Media, 2003.

[3] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin. A Neural Probabilistic Language Model. *Journal of Machine Learning Research*, 3:1137–1155, 2003.

[4] A. L. Benin, G. Vitkauskas, E. Thornquist, E. D. Shapiro, J. Concato, M. Aslan, and H. M. Krumholz. Validity of using an electronic medical record for assessing quality of care in an outpatient setting. *Medical care*, 43(7):691–8, July 2005.

[5] C. M. Bishop. *Pattern recognition and machine learning*. springer, 2006.

[6] T. Borlawsky, C. Friedman, and Y. A. Lussier. Generating executable knowledge for evidence-based medicine using natural language and semantic processing. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, pages 56–60, Jan. 2006.

[7] S. F. Chen and J. Goodman. An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, pages 310–318. Association for Computational Linguistics, 1996.

[8] R. Collobert and J. Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM, 2008.

[9] C. Cortes and V. Vapnik. Support vector machine. *Machine learning*, 20(3):273–297, 1995.

[10] B. de Bruijn, S. Carini, S. Kiritchenko, J. Martin, and I. Sim. Automated information extraction of key trial design elements from clinical trial publications. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, pages 141–5, Jan. 2008.

[11] P. Devyver and J. Kittler. *Pattern Recognition: A Statistical Approach*. Prentice-Hall, 1982.

[12] P. S. Dhillon, D. P. Foster, and L. H. Ungar. Eigenwords: Spectral word embeddings. *Journal of Machine Learning Research (pending)*, 2015.

[13] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern classification*. John Wiley & Sons, 2012.

[14] K. Fukunaga. *Introduction to statistical pattern recognition*. Academic press, 2nd edition, 1990.

[15] G. Golub and C. Reinsch. Singular Value Decomposition and Least Squares Solutions. *Numer. Math.*, 14:403–420, 1970.

[16] J. T. Goodman. A bit of progress in language modeling. *Computer Speech & Language*, 15(4):403–434, 2001.

[17] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664, 2004.

[18] M. Z. Hydari, R. Telang, and W. M. Marella. Electronic health records and patient safety. *Communications of the ACM*, 58(11):30–32, Oct. 2015.

[19] I. Jolliffe. *Principal component analysis*. Wiley Online Library, 2002.

[20] D. Jurafsky and J. H. Martin. *Speech and language processing*. Pearson, 2014.

[21] P. Koehn. *Statistical machine translation*. Cambridge University Press, 2009.

[22] R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. *IJCAI Proceedings of the 14th international joint conference on Artificial intelligence*, pages 1137–1143, 1995.

[23] T. K. Landauer, P. W. Foltz, and D. Laham. An Introduction to Latent Semantic Analysis. *Discourse Processes*, 25:259–284, 1998.

[24] Q. V. Le and T. Mikolov. Distributed representations of sentences and documents. *In Proceedings of ICML*, 2014.

[25] R. Lebret and R. Collobert. Word emdeddings through hellinger pca. *In EACL*, 2014.

[26] A. L. Maas and A. Y. Ng. A Probabilistic Model for Semantic Word Vectors. *In NIPS 2010 Workshop on Deep Learning and Unsupervised Feature Learning*, pages 1–8, 2010.

[27] C. D. Manning and H. Schütze. *Foundations of statistical natural language processing*. MIT press, 1999.

[28] J. Mayer, T. Greene, J. Howell, J. Ying, M. A. Rubin, W. E. Trick, and M. H. Samore. Agreement in classifying bloodstream infections among multiple reviewers conducting surveillance. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, 55(3):364–70, Aug. 2012.

[29] S. Meystre and P. J. Haug. Natural language processing to extract medical problems from electronic clinical documents: performance evaluation. *Journal of biomedical informatics*, 39(6):589–99, Dec. 2006.

[30] T. Mikolov, A. Deoras, D. Povey, L. Burget, and J. Černockỳ. Strategies for training large scale neural network language models. In *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*, pages 196–201. IEEE, 2011.

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

[32] G. A. Miller and W. G. Charles. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28, 1991.

[33] J. A. Miñarro-Giménez, O. Marín-Alonso, and M. Samwald. Applying deep learning techniques on medical corpora from the world wide web: a prototypical system and evaluation. *arXiv preprint arXiv:1502.03682*, 2015.

[34] A. Mnih and G. E. Hinton. A scalable hierarchical distributed language model. In *Advances in neural information processing systems*, pages 1081–1088, 2009.

[35] F. Morin and Y. Bengio. Hierarchical probabilistic neural network language model. *AISTATS*, 2005.

[36] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. *In EMNLP*, 2014.

[37] L. E. Peterson. K-nearest neighbor. *Scholarpedia*, 4(2):1883, 2009.

[38] X. Rong. word2vec parameter learning explained. *arXiv preprint arXiv:1411.2738*, 2014.

[39] S. T. Rosenbloom, J. C. Denny, H. Xu, N. Lorenzi, W. W. Stead, and K. B. Johnson. Data from clinical notes: a perspective on the tension between structure and flexible documentation. *Journal of the American Medical Informatics Association : JAMIA*, 18(2):181–6, 2011.

[40] R. Rosenfield. Two decades of statistical language modeling: Where do we go from here? *presented at the Workshop-2000 Spoken Lang. Reco. Understanding, Summit, NJ*, 2000.

[41] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620, Nov. 1975.

[42] P. D. Turney and P. Pantel. From Frequency to Meaning : Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, 37:141–188, 2010.

[43] G. P. Zhang. Neural networks for classification: a survey. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 30(4):451–462, 2000.

[44] W. Y. Zou, R. Socher, D. M. Cer, and C. D. Manning. Bilingual word embeddings for phrase-based machine translation. In *EMNLP*, pages 1393–1398, 2013.