1-1-2015

Inference from Structred and Unstructured Electronic Medical Data for Early Dementia Detection

Joseph Bullard

Rohan Murde

Qi Yu

Cecilia Ovesdotter Alm

http://dx.doi.org/10.1287/ics.2015.XXXX
Creative Commons License

14th INFORMS Computing Society Conference
Richmond, Virginia, January 11–13, 2015
pp. 000–000

Inference from Structured and Unstructured Electronic Medical Data for Early Dementia Detection

Abstract The prevalence of Alzheimer's disease (AD) and other forms of dementia is increasing with the aging population, both in the United States and around the globe. The inability to cure these conditions results in prolonged and expensive medical care. Early detection is critical to potentially postpone symptoms and to prepare both healthcare providers and families for subjects' future needs. Current detection methods are typically costly or unreliable, and much stands to benefit from improved recognition of early AD markers. Electronic patient records provide the potential for computational analysis and prediction of complex diseases like AD. Whereas prior work on this problem has focused on structured data (e.g. test results) alone, this study integrates structured and unstructured (e.g. clinical notes) from the Alzheimer's Disease Neuroimaging Initiative (ADNI)* for classification of subjects' dementia status. Prediction based on unstructured data alone performs with similar accuracy compared to structured data, and integration of the two provides performance improvements over either in isolation. In addition, we provide insights into which structured features were more useful for classification of AD, supporting previously observed trends, while also highlighting the potential for computational methods to discover new early markers.

Keywords electronic medical data; Alzheimer's Disease Neuroimaging Initiative (ADNI); structured and unstructured data; integrating heterogeneous data

1. Introduction

Dementia is an increasing problem for the US and global aging population. Approximately 35 million people worldwide suffer from some form of dementia, and this number is expected to double by the year 2030 [13]. It is the 6th leading cause of death in the US [1]. Alzheimer's Disease in particular has no cure, and treatments are limited, making management of symptoms the main focus of clinical care. Accordingly, there is a high medical cost associated with dementia care, expected to total 214 billion dollars in the US for the year 2014 [1].

* Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. (Per data use agreement, information about the ADNI data are included verbatim from the ADNI project in Section 3.1 and Acknowledgments.) A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

Early detection is critical for potential management of symptoms, and for allowing families to adjust and adequately plan for the future. Despite this importance, current detection methods are costly, invasive, or unreliable. Consequently, most patients are not diagnosed until their symptoms have already progressed. Improved understanding and recognition of early warning signs of dementia would greatly benefit the detection and management of the disease, as well as facilitate appropriate allocation of resources for healthcare organizations' provision of care.

With the advent of electronic health records (EHRs) comes the potential for large-scale computational analysis of patients' clinical data to understand or discover warning signs of the progression of medical conditions. In this context, the data can be considered either *structured* or *unstructured*. Examples of *structured* data include patient demographics, such as age, sex, or ethnic background, as well as test results collected during their visits, such as brain volume measurements or routine blood tests. When a patient is already suspected of having dementia, there is a battery of biological and cognitive tests that can be applied by a clinical professional. However, cognitive tests focus on already observable symptoms. It would be useful to identify early signs that can point towards an individual being at-risk of developing dementia. Intelligent decision-support models that could detect such cases would be very helpful for healthcare providers. Moreover, while certain clinical tests are used in daily operations for dementia diagnosis, other early biophysical or cognitive-social markers may be undiscovered. Here we explore a range of features in a machine learning context to identify useful features for classifying patients' dementia status.

We are also are interested in understanding the roles of different types of medical data in a prediction process, especially since prior work has focused primarily on structured information. Accordingly, we explore what potential role unstructured data can play in data mining for dementia prediction. *Unstructured* data refers to text entries, such as patient histories, impressions, visit summaries, discharge summaries, or other broader or narrower categories. Written, clinical text presents a potentially rich source of information that embeds useful clinical knowledge from the professionals who wrote them. Insights mined from natural language data may be quite straightforwardly interpretable by humans, and written texts may provide opportunities to capture distinct types of information (e.g. as related to behavioral wellness or social lifestyle patterns), compared to structured entries. Additionally, some form of text is present in nearly all patient records, whereas many of the relevant structured data may be absent if the patient has not already been identified as being at-risk for a given condition. Thus, we examine ways in which unstructured data may supplement a diagnostic model based on structured data. Specifically, we integrate models based on each data type to improve performance over either in isolation.

A computational model based on either or both these data types could be used in an automated screening system to identify and flag potentially problematic patients for further assessment by clinicians, making operations more efficient. Additionally, identification of useful features from classification experiments may improve understanding of important markers in early dementia and Alzheimer's Disease detection.

2. Related Work

The potential of electronic medical records for data mining has been recognized for some time. Importantly, deducing data-driven patterns based on structured data has been the basic approach in the prediction of dementia. Biomarkers from cerebrospinal fluid (CSF), as well as brain volume measurements from magnetic Resonance Imaging (MRI) and Positron Emission Tomography (PET), have been useful in predicting conversion from Mild Cognitive Impairment (MCI) to Alzheimer's Disease (AD) within the ADNI dataset [15]. Cognitive tests may be useful to this end as well, in the absence of other biophysical tests [7] or in combination with them [4]. However, such tests are not typically applied until after symptoms of MCI or dementia have already been observed.

Although the use of structured data has been successful in dementia and AD prediction, unstructured text data may provide new or additional benefits for modeling purposes. Natural language processing (NLP) and statistical text mining (STM) techniques that have been applied to medical texts in the past have focused on extracting known disease markers obtained from medical knowledge sources for machine learning purposes. Examples include mapping terms to a medical ontology to predict post-operative complications [10], and using tools like MedLEE or SymText to extract and codify terms for identifying cases of colorectal cancer [17], suspicious mammogram findings [9], adverse events related to central venous catheters [12], and bacterial pneumonia from chest X-rays [5]. For the scope of this paper, we examine the utility of bag-of-words modeling for capturing important linguistic units in unstructured text data, with an eye towards integrating more sophisticated topic modeling later.

3. Dataset

3.1. Subjects

The dataset used here was obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (`adni.loni.usc.edu`). The ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organizations, as a $60 million, 5-year public-private partnership. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials. The methods used in this paper represent a secondary use of the data for a purpose that is in line with general goal of identifying dementia markers.

The Principal Investigator of this initiative is Michael W. Weiner, MD, VA Medical Center and University of California San Francisco. ADNI is the result of efforts of many co-investigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the U.S. and Canada. The initial goal of ADNI was to recruit 800 subjects but ADNI has been followed by ADNI-GO and ADNI-2. To date these three protocols have recruited over 1500 adults, ages 55 to 90, to participate in the research, consisting of cognitively normal older individuals, people with early or late MCI, and people with early AD. The follow up duration of each group is specified in the protocols for ADNI-1, ADNI-2 and ADNI-GO. Subjects originally recruited for ADNI-1 and ADNI-GO had the option to be followed in ADNI-2. For up-to-date information, see `www.adni-info.org`.

3.2. Dataset Preparation and Processing

Data from the ADNI are split across multiple files, each containing different related clinical and biomarker data. The structured data used here are obtained from a subset of the available files in the `adnimerge` package, listed and described in Table 1. Each of the over 11,000 entries contained in these files represents one visit for one subject, with many subjects having multiple visits. We aggregate entries by subject ID such that all data fields for a given subject are contained in one vector, resulting in a collection of 1,736 subject vectors each containing 22 structured data fields. Each data field of a given subject's vector was equal to the mean of the available values for that field over all of the subject's visits. While none of the files listed above contain any text notes, a number of other ADNI data files do contain an optional text notes field. The list of subject IDs is used to extract text notes from

Table 1. ADNI files used to obtain structured and unstructured data.

File	Description	Data type
arm	Diagnosis at screening	Structured
baimrinmrc	Brain volume atrophy	Structured
cdr	Clinical Dementia Rating (CDR) scores	Structured
upennbiomk5	Cerebrospinal fluid (CSF) biomarkers	Structured
upennbiomk6	Cerebrospinal fluid (CSF) biomarkers	Structured
upennplasma	Plasma biomarkers	Structured
ucberkeleyav45	PET scan with florbetapir	Structured
BLLOG	Baseline symptoms log	Unstructured
BLCHANGE	Changes since baseline visit	Unstructured
RECMHIST	Recent medical history	Unstructured
RECADV	Recent adverse events / hospitalizations	Unstructured

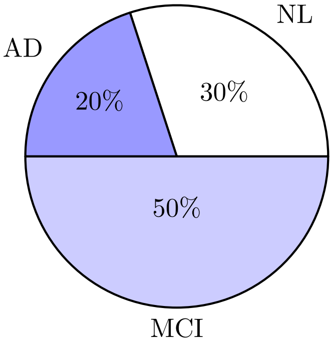
four other files (see Table 1), which were selected both because they contained substantial amounts of text entries and because their content seemed potentially useful for this study. The notes from all visits for a given subject are concatenated and treated as one document for that subject.

Mild Cognitive Impairment (MCI) is a diagnostic category regarded as a precursor to Alzheimer's Disease (AD). The original ADNI dataset encodes three different fine-grained levels of MCI: Early MCI (EMCI), MCI, and Late MCI (LMCI); these were all combined into one class label *MCI* to mitigate data sparseness and class imbalance issues. Additionally, another encoding category used in the ADNI dataset is Significant Memory Complaint (SMC). As SMC is generally thought distinct from MCI and AD, it was grouped with cognitively normal subjects under the class label *NL*. The group of Alzheimer's subjects were left as-is, with the class label *AD*. Each subject was assigned a class label from this 3-class scheme based on their most recent diagnostic state, as some had changed over the course of the ADNI collection period. As seen in in Figure 1, this labeling scheme results in a 50% majority for the MCI class, which we consider as a baseline for comparison.

3.2.1. Handling Missing Values In general, visit entries in the ADNI dataset contain many missing values for data fields, since not all tests are administered during every visit. These missing values present a problem for the classification experiments we wish to perform. Before subject aggregation, a total of 3,329 visit entries have missing values for all tests; we exclude these in our dataset. Figure 2 shows a histogram of the fraction of the 22 structured data fields that were missing for the remaining 8,036 visit entries (across the 1,736 subjects). As expected, most visits are missing most available tests. To deal with this, missing values are replaced by imputation using Amelia II [8], a package in the R programming language. The missing values are obtained by calculating log likelihood and using the EMB algorithm which combines the EM algorithm with bootstrapping on a subset of observed values. Amelia II performs multiple imputations, which reduces bias and increases efficiency of the missing data. A set of five different imputed datasets is generated, the mean of which is used to create one dataset without missing values. As for the unstructured data, only one of the subjects had an empty text field, which was replaced with a special *empty* token.

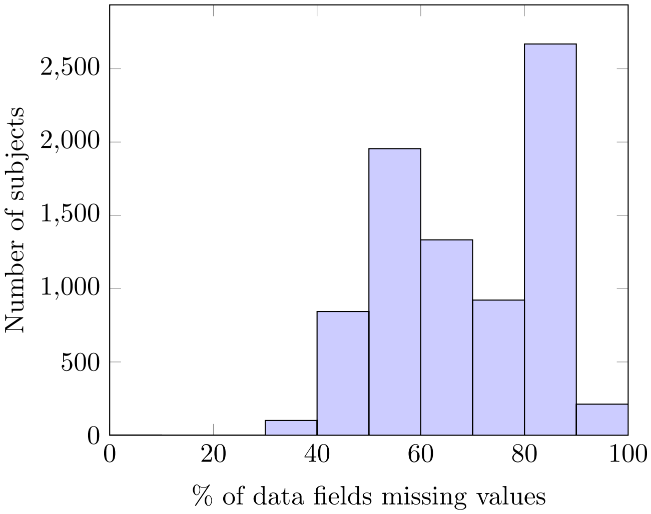
3.2.2. Text Processing and Normalization Text pre-processing included lowercasing, punctuation removal, and stop-listing of frequent tokens such as grammatical function words (e.g. *the*). Text normalization procedures were also performed to deal with variation in linguistic and orthographical representations of numbers, dates, ages, abbreviations, multiword expressions, and range expressions. In particular, dates and ages were represented in many different forms, which were dealt with by matching and tagging with a uniform replacement string. For example, the strings *70-year old*, *70 yo*, *70 years-old*, and *70 y/o*, etc. would

FIGURE 1. Class distribution in final dataset (see Section 3.2).



Note. Alzheimer's Disease (AD), Mild Cognitive Impairment (MCI), Normal (NL)

FIGURE 2. Fraction of structured data fields containing missing values, per visit ($n = 8,036$) over all subjects.



all be replaced with the tag *AGE_70*. This helps with the problem of different representations of the same semantic unit not being recognized in the model, and can also allow for generalizations such as *AGE_senior* to be explored. The techniques applied here appear to cover the vast majority of ambiguities and variations present in the data. Additionally, the most frequent lexical content bigrams and trigrams (i.e. after stop-listing) in the dataset were extracted and treated as multiword expressions (MWEs) to be replaced as unigrams in the texts by concatenating the words with underscores (e.g. *alzheimers disease* became *alzheimers_disease*, *depressed mood* became *depressed_mood*, etc.). The list was inspected to remove any potential errors. This latter text preprocessing step was done because of the expected high number of MWE expressions in a domain such as medicine, and to help disambiguate meaning and aid interpretation of useful text features.

4. Predictive Modeling Approach

We report on classification models trained on features of the structured and unstructured data in isolation, and additionally provide an integration technique for the two models. In all cases, the goal is to predict a subject's diagnostic state, according to the label assigned in Section 3.2, using features of their structured or unstructured data. All experiments make use of 10-fold cross-validation with a logistic regression classifier using L1 penalty, implemented in Python using the `scikit-learn` machine learning library [11].

4.1. Structured Features

The structured data fields for each subject already contain numerical values that can be used as features without further processing. In addition to the whole collection of structured data features, we also experiment with a subset which excludes all three cognitive tests: Alzheimer's Disease Assessment Scale (ADAS), Clinical Dementia Rating (CDR) memory score, and the Mini Mental State Exam (MMSE). The reason for this is that while cognitive tests are powerful tools for diagnosing dementia and Alzheimer's disease, they typically focus on observable dementia symptoms rather than early detection. It is thus interesting to examine their impact on classification, which we do by performing one experiment using only non-cognitive structured features, and another using all of the structured features.

4.2. Unstructured Features

We construct a standard bag-of-words model on the unstructured dataset and use the word tokens as features for classification. As described earlier in Section 3.2.2, frequent bigrams and trigrams content word sequences were treated as individual tokens in the texts to preserve meaning of common multiword expressions, therefore this is not just a plain unigram model. The features are treated as boolean values indicating the presence or absence of the token in the texts for each subject.

4.3. Integration of Models

We also investigate classification performance when integrating the structured and unstructured models. For each subject, a logistic regression model will compute posterior probabilities for each of the three class labels (diagnostic states) and select the most likely candidate. If we assume that the posterior probabilities from each of two models are independent, then Equation 1, below, can be leveraged to make a final decision in an integrated model.

$$\begin{aligned} p(C_k \mid X_s, X_u) &\propto p(X_s, X_u \mid C_k)\, p(C_k) \\ &\propto p(X_s \mid C_k)\, p(X_u \mid C_k)\, p(C_k) \\ &\propto \frac{p(C_k \mid X_s)\, p(C_k \mid X_u)}{p(C_k)} \end{aligned} \tag{1}$$

Here, X_s and X_u represent the input feature vectors for the structured and unstructured models, respectively, and C_k represents a class label (with $p(C_k)$ being the prior probability of the class within the dataset). As in the original models, the class label with the highest probability is selected as the output. As with the structured features in this experiment, we train two integrated models: one including cognitive features and one excluding them.

5. Results and Discussion

Performance metrics for each experiment are shown in Table 2, averaged over all folds. As shown earlier in Figure 1, the majority class baseline is 50%, represented by the MCI class. All models clearly perform well above this baseline. As expected, the inclusion of the features obtained from cognitive tests markedly improved classification performance for the structured models. Importantly, the unstructured features alone perform almost as well as the structured, when cognitive features are included. In many cases, the latter may not be available and, thus, relying on their inclusion could cause brittleness when translating models into clinical practice. As expected, the probabilistic integration of structured and unstructured models produced additional performance gains over each of the two feature groups in isolation. As before, the inclusion of the cognitive markers results in greater performance in integration, although it is a much smaller gain in this case.

In all three cases, the AD class achieved the lowest precision and recall, possibly due to the small class size. One potential source of confusion may be the decision to combine the three more fine-grained subcategories (EMCI, MCI, and LMCI) into one coarser-grained class label. Thus, it is possible that EMCI and LMCI, which represent the respective peripheries of the MCI continuum, could tend to be confused with NL and AD, respectively; this will be examined in future work.

5.1. Structured Feature Analysis

Logistic regression is a linear classification algorithm whose decision function consists of coefficients on each feature input, and whose output corresponds to a class. The magnitude of a coefficient corresponds to how much influence its feature has on the overall decision,

Table 2. Classification results using structured and unstructured features.

		Class *NL*		Class *MCI*		Class *AD*	
Model	Accuracy	Prec.	Recall	Prec.	Recall	Prec.	Recall
Structured (− cognitive)	65%	.67	.59	.63	.73	.69	.54
Structured (+ cognitive)	78%	.83	.82	.76	.81	.75	.64
Unstructured	75%	.77	.76	.78	.80	.65	.63
Integrated (− cognitive)	79%	.80	.81	.80	.79	.75	.74
Integrated (+ cognitive)	82%	.87	.84	.82	.84	.78	.77

Note. Accuracy is over all classes; precision and recall are reported for each of the three classes.

Table 3. Useful structured features for Alzheimer's Disease (AD) (coefficients significantly non-zero).

Feature	Source	
CDMemory	Cognitive test	***
MMSE	Cognitive test	***
VBSI	Boundary shift integral	***
ADAS13	Cognitive test	**
Tau	Cerebrospinal fluid	*

Note. *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$

Table 4. Useful non-cognitive structured features for Alzheimer's Disease (AD) (coefficients significantly non-zero).

Feature	Source
Tau	Cerebrospinal fluid
VentVol	Boundary shift integral
PTau	Cerebrospinal fluid
BrainVol	Boundary shift integral
IPCA	Magnetic resonance imaging

Note. $p < 0.001$ in all cases

and the sign indicates which class a higher value of the feature favors. In this section we inspect these coefficients to gain insights into which features are more influential in deciding a subject's diagnostic state. We are dealing with a multi-class problem: AD vs. MCI vs. NL, but given our primary interest in understanding indicators of Alzheimer's Disease (AD), we focus on the features and coefficients of the AD class alone.[1].

The L1 norm penalty imposed in the algorithm aids in feature analysis by forcing many coefficients to zero, indicating that the corresponding features are not used in the decision. Structured features whose coefficients were significantly different than zero ($p < 0.05$) are shown in Table 3 in descending order of significance. As expected, all three of the cognitive markers - Clinical Dementia Rating memory score (CDMemory), Mini Mental State Exam (MMSE), and Alzheimer's Disease Assessment Scale (ADAS13) - played a significant role in discriminating between subjects who had or did not have Alzheimer's Disease (AD). The Tau biomarker was also significant, verifying results in the literature [16]. As discussed earlier, it would be beneficial to accurately identify at-risk patients who are in need of receiving cognitive tests, therefore we perform the same feature analysis again, excluding the cognitive markers (see Table 4), which resulted in an interesting list of other useful structured features. Again, Tau is among the most significant, along with it's counterpart PTau (phospho-tau), which has been linked to neurodegenerative diseases like AD [6].

6. Conclusions

In this paper we aimed to explore both structured and unstructured data of subject records collected as part of the Alzheimer's Disease Neuroimaging Initiative (ADNI), using machine learning approaches. We experimented with each data type to classify a condensed version of subjects' diagnoses as either normal (NL), mild cognitive impairment (MCI), or Alzheimer's disease (AD). A system like this would be useful as an intelligent early screening system

[1] We provide analysis for structured features, but not for unstructured. The latter would be more interesting in future work with more sophisticated modeling of the text data.

for identifying subjects who may need more attention for dementia testing. The accuracy of our models were well above the majority class baseline held by the MCI subjects, with mostly comparable results for precision and recall. Arguably, high recall (e.g. in the case of AD, the fraction of subjects with AD who were correctly classified as such) is of more importance due to the higher cost of type II errors in this diagnostic context; future work may involve optimizing the modeling for achieving higher recall. Integrated models based on probabilistic output of the structured and unstructured models markedly improved classification performance over either feature group in isolation. This is important due to the fact that many of the tests from the structured data will not be present for patients until their dementia symptoms have progressed, but nearly everyone will have text notes in their medical record. These texts may also allow us to capture or discover other forms of markers (e.g. as related to a patient's behavioral health or social-psychological experiences). In addition to the integration, analyzing the structured features that played a greater role in classification revealed useful features that had been identified in the past, as well as highlighted the potential of new discovery through computational methods.

Ultimately, we hope to use topic modeling [3] to infer more complex linguistic relationships in the unstructured data. Initial topic modeling experiments on the ADNI texts did not produce viable results. We relate this in part due to the heterogeneity of language usage enforced by the specific goals and screening processes of this dataset's collection experiments. In the future, we plan to explore general electronic health records (EHRs) with the expectation that the full scope of typical medical visits will prove more fruitful for topic modeling.

Additionally, we would like to explore a more sophisticated way of combining the structured and unstructured models to improve predictive power. It is possible that one may be more suited than the other under particular circumstances, and a model that more effectively combines both could be useful. To achieve this, we will explore both the boosting technique [2], which combines multiple base classifiers to achieve better overall prediction accuracy, as well as a recent mathematical optimization algorithm that evaluates and ranks multiple alternatives in a group decision-making process [14].

Acknowledgements

We are thankful for being able to use the ADNI project's dataset in this work. Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Alzheimers Association; Alzheimers Drug Discovery Foundation; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; GE Healthcare; Innogenetics, N.V.; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (`www.fnih.org`). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

References

[1] Alzheimer's Association. 2014 Alzheimer's Disease Facts and Figures. *Alzheimer's & Dementia*, 10, 2014.

[2] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[4] D. P. Devanand, X. Liu, M. H. tabert, G. Pradhaban, K. Cuasay, K. Bell, M. J. de Leon, R. L. Doty, Y. Stern, and G. H. Pelton. Combining early markers strongly predicts conversion from mild cognitive impairment to Alzheimer's disease. *Biological Psychiatry*, 64(10):871–879, 2008.

[5] M. Fiszman, W. W. Chapman, D. Aronsky, R. S. Evans, and P. J. Haug. Automatic detection of acute bacterial pneumonia from chest X-ray reports. *Journal of the American Medical Informatics Association*, 7(6):593–604, 2000.

[6] I. Grundeke-Iqbal, K. Iqbal, Y.-C. Tung, M. Quinlan, H. M. Wisniewski, and L. I. Binder. Abnormal phosphorylation of the microtuble-associated protein τ (tau) in Alzheimer cytoskeletal pathology. *Proceedings of the National Academy of Sciences of the United States of America*, 83:4913–4917, 1986.

[7] D. B. Hogan and E. M. Ebly. Predicting who will develop dementia in a cohort of Canadian seniors. *Canadian Journal of Neurological Sciences*, 27(1), 2000.

[8] J. Honaker, G. King, and M. Blackwell. Amelia II: A program for missing data. *Journal of Statistical Software*, 45:1–47, 2011.

[9] N. L. Jain and C. Friedman. Identification of findings suspicious for breast cancer based on natural language processing of mammogram reports. In *Proceedings of the AMIA Annual Fall Symposium: American Medical Informatics Association*, pages 829–833, 1997.

[10] H. J. Murff, F. FitzHenry, M. E. Matheny, N. Gentry, K. L. Kotter, K. Crimin, S. Dittus, Robert, A. K. Rosen, P. L. Elkin, S. H. Brown, and T. Speroff. Automated identification of postoperative complications within an electronic medical record using natural language processing. *The American Journal of Medicine*, 306(8):848–855, August 2011.

[11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[12] J. F. E. Penz, A. B. Wilcox, and J. F. Hurdle. Automated identification of adverse events related to central venous catheters. *Journal of Biomedical Informatics*, 40(2):174–182, April 2007.

[13] M. Prince, M. Prina, and M. Guerchet. *World Alzheimer Report 2013*. Alzheimer's Disease International (ADI), London, 2013. Available (no cost) at `http://www.alz.co.uk/research/world-report-2013`.

[14] T. Schmoke. An optimization-based approach for vaccine prioritization. Master's thesis, Rochester Institute of Technology, Rochester, NY, USA, 2013.

[15] J. L. Shaffer, J. R. Petrella, F. C. Sheldon, K. R. Choudhury, V. D. Calhoun, R. E. Coleman, P. M. Doraiswamy, and Alzheimers Disease Neuroimaging Initiative. Predicting cognitive decline in subjects at risk for Alzheimer disease by using combined cerebrospinal fluid, MR imaging, and PET biomarkers. *Radiology*, 266(2):583–591, 2013.

[16] M. G. Spillantini and M. Goedert. Tau pathology and neurodegeneration. *The Lancet Neurology*, 12(6):609–622, 2013.

[17] H. Xu, Z. Fu, A. Shah, Y. Chen, N. B. Peterson, Q. Chen, S. Mani, M. A. Levy, Q. Dai, and J. C. Denny. Extracting and integrating data from entire electronic health records for detecting colorectal cancer cases. In *AMIA Annual Symposium Proceedings 2011*, pages 1564–1572, 2011.