Predicting Medical Nonadherence Using Natural Language Processing*

Hedyeh Elahinia[1,2], Jesse D. Raffa[2], Joy Tzung-yu Wu[3], and Mohammad M. Ghassemi[2]

Abstract— **Patient nonadherence is a multi-billion dollar problem in the United States healthcare system[1], accounting for over 93 million people and over 10% of American healthcare spending[2]. This study makes use of techniques in unsupervised natural language processing and human annotation to generate a set of predictive words that detect medical nonadherence. We defined nonadherence with more nuance than just patient medication practices by taking into account various psychosocial factors including adherence to dietary and therapeutic advice. Because of the multifaceted nature of the problem, our study analyzed the most multifaceted element of healthcare data: physician notes. We used natural language processing to extract meaningful keywords that predict nonadherence. We constructed three contextual categories of keywords that were statistically significant *($p<0.05$)*predictors of nonadherence. Using our extracted key features, we made a nuanced contribution to the detection of nonadherence. These findings may be used to facilitate reduction of nonadherence in our healthcare system.**

I. INTRODUCTION

Patient nonadherence is a multi-billion dollar problem in the United States (US) healthcare system[1], accounting for over 93 million people and over 10% of healthcare spending[2]. Typically, nonadherence is measured using structured pharmaceutical data, such as pill counts and insurance claims. However, medication nonadherence accounts for a small subset of all medical nonadherence, which is comprised of many diverse psychosocial factors[3]. Such factors are reflected only in unstructured healthcare data, in the format of physician notes (discharge summaries, nurse's notes, and social work notes). These notes contain a wealth of information about factors that interfere with adherence—social, psychological, economic, and otherwise[4].

Medication nonadherence can be measured using structured healthcare data with such calculations as the medical possession ratio (MPR) and the proportion of days covered (PDC)[5]. For this reason, many previous studies of nonadherence use structured data analysis to assess nonadherence. A more complete definition, however, should include consequential factors in treatment beyond medication; examples include adherence to therapy, drug and alcohol recommendations, dietary restrictions, appointment follow-ups and in-hospital instructions[6].

Nonadherence is an important contributor to morbidity and mortality[1]. While the detection of nonadherence from medical notes is an easy task for a human, automated assessment is complex and technically challenging. These challenges arise from the highly heterogeneous nature of medical notes, which contain everything from family history to blood glucose levels. Our study approaches the issue of medical nonadherence using unsupervised natural language processing (NLP) and human annotation. We defined nonadherence as the temporary or permanent discontinuation of a medical treatment, including medication, appointments, and medical advice. We expanded on the intuition of human annotators and medical experts by generating a list of diverse keywords predictive of nonadherence, making it possible to understand medical nonadherence in a more nuanced way, and paving the way for prediction and reduction of nonadherence in US healthcare.

*This work was supported by the MIT Laboratory for Computational Physiology and grants NIH-R01-EB017205, NIH-R01-EB001659, and NIH-R01-GM104987

[1]The University of Toledo, Toledo, OH 43606, USA helahin@rockets.utoledo.edu

[2]MIT Critical Data, Cambridge, MA, 02139, USA http://criticaldata.mit.edu/

[3]IBM Research Alamaden, San Jose, CA 95120, USA http://linkedin.com/in/joytywu

978-1-5386-2534-7/17/$31.00 ©2017 IEEE

II. METHODS

This study was comprised of three distinct stages as shown in Fig. 1. The data was selected in stage one. Stage two involved generating keywords from three sources: annotator, physician, and Word2Vec analysis. Stage three involved creating contextual categories, tfidf analysis and logistic regression.

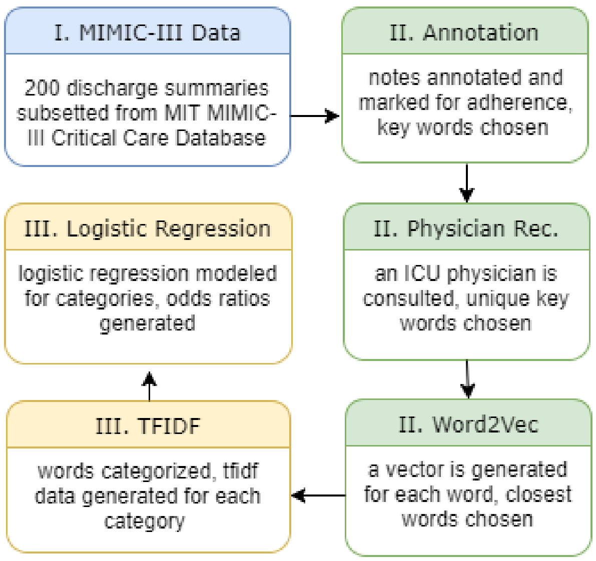

Fig. 1. Our study included three stages: I. Data extraction (blue), II. Keyword selection (green), and III. Analysis (yellow).

Data was collected from the publicly accessible Multi-parameter Intelligent Monitoring in Intensive Care (MIMIC) database[7]. MIMIC catalogs patients who stayed at the Beth Israel Deaconess Medical Center's Intensive Care Units (ICU). We included 198 unique discharge summaries from patients that were in the ICU >3 times per year, 66 non-adherent and 132 adherent cases. This inclusion criteria was selected to maximize the number of instances of nonadherence[1]. A single human annotator then read each discharge summary, annotated it, and coded it as either non-adherent or adherent, using the following definition of nonadherence: temporary or permanent discontinuation of a treatment, including medications or appointments, without consulting a physician prior to doing so. The annotator identified seven keywords that occurred commonly in the non-adherent category (see Table I, column 1). We then applied the Word2Vec tool[8] on all 198 notes. Word2Vec is an unsupervised natural language processing technique that assigns each unique word in the collection of notes to a point in a vector space. Word2Vec places words that are contextually related closer together; for example, "diabetes" and "insulin" are close to each other in the space, while "healthy" and "ill" are far apart. Word2Vec found the words most closely associated with the annotator keyword (see Table I, column 1).

We displayed the vector space[9] in Fig. 2:

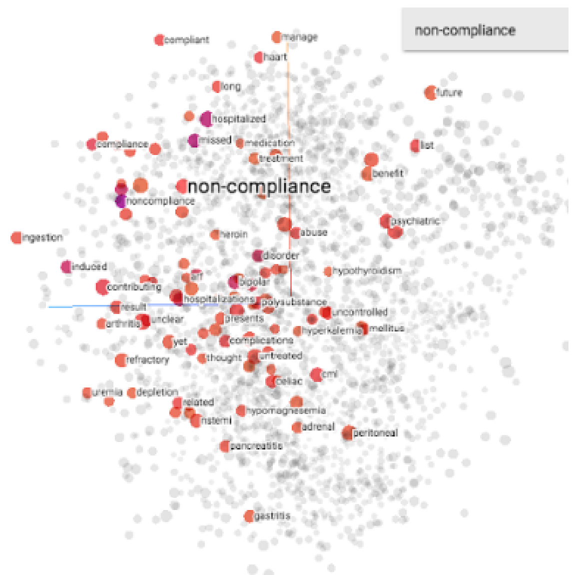

Fig. 2. The word vector space generated using the 198 medical notes. Points in red are associated with "non-compliance." Words close to one another in the vector space are contextually related.

The Word2Vec model represents each words as a point. When using Word2Vec, we discarded terms that appeared less than twenty-five times, removed all punctuation, and ignored filler words like "or" and "is." We discarded keywords that were too common to carry clinical significance, like "patient" and "blood." We selected words that were most closely associated with the annotator kewyords; this provided four additional words (see Table I, column 2). A physician with clinical ICU expertise was consulted, who provided three additional key terms for nonadherence (see Table I, column 3).

Term frequency inverse document frequency (tfidf) analysis was performed on the original discharge summaries for each of the fourteen keywords shown in Table I. Tfidf scores represent, on a

[1]and the amount of annotator exposure to relevant nonadherence terms, because a random sample would not generate enough positive cases for annotator- or model-training.

978-1-5386-2534-7/17/$31.00 ©2017 IEEE

TABLE I

KEYWORDS BY SOURCE

Annotator	Word2Vec	Physician
admits, ama compliance non-compliance non-compliant refused, suspected	dementia abuse uncontrolled untreated	difficult compliant historian

continuous scale, the relative frequency of a given word in the corpus of notes. These scores reflect not only simple frequency but also give weight to more uncommon words; for example, an instance of the more uncommon term "polysubstance" is weighted more heavily than that of the more common term "doctor". Since the tfidf scores ranged from zero to less than 0.1, they were scaled by dividing each value by the maximum tfidf, yielding a score ranging between 0 and 1. We then separated

TABLE II

CONTEXTUAL CATEGORIES FOR TFIDF

C1: Direct Indicators	C2: Implied Indicators	C3: Condition Indicators
compliant non-compliant compliance noncompliance	suspected difficult historian ama refused admits	dementia uncontrolled untreated abuse

the words into three distinct contextual categories as seen in Table II: direct indicators of nonadherence (like non-compliant and compliance), implied indicators of nonadherence (like ama and difficult), and condition-related indicators of nonadherence (like abuse and dementia).

The tfidf scores were added across these categories and a logistic regression analysis was performed for each group using the R programming language. We generated two odds ratios (with confidence intervals and p values) for each category in Table II: one that was unscaled, which compared nonzero tfidf scores to zero scores, and another that was scaled, which compared the maximum tfidf score to the minimum. In this way, we modeled both *whether* and *how well* each category of words could predict nonadherence.

III. RESULTS

In Table III we show the scaled and unscaled odds ratio for the logistic regression analysis using each of the contextual categories set forth in Table II.

TABLE III

LOGISTIC REGRESSION

Category	Odds Ratio	95% CI	p-value
C1, unscaled	11.85	4.53-37.20	<0.001
C2, unscaled	2.67	1.44-4.97	1.80E-03
C3, unscaled	3	1.50-6.05	1.97E-03
C1, scaled	338.93	26.32-9227.98	<0.001
C2, scaled	22.45	3.75-173.94	1.34E-03
C3, scaled	3.44	0.73-16.83	0.116

The first category (C1: direct indicators) shows an 11.85-fold increase in the odds of nonadherence for nonzero tfidf scores versus zeros ($p<0.001$) and a 338.93-fold increase in the odds of nonadherence when comparing the highest tfidf score versus the lowest score ($p<0.001$). The second category (C2: condition indicators) showed a 2.67-fold increase in nonadherence odds for nonzero tfidf scores as compared to zeros ($p<0.05$), and a 22.45-fold scaled increase for the highest versus lowest score ($p<0.05$). The third category (C3: condition indicators) showed a three-fold increase in the odds of nonadherence prediction when comparing nonzero tfidf scores to zeros ($p<0.05$), and a 3.44-fold increase in scaled score comparison of highest to lowest tfidf values ($p=0.12$). Each of the three categories measured in our study displayed statistical significance in its unscaled odds ratio; the words contained in these categories can therefore be used in the prediction of nonadherence. Categories one and two were statistically significant in a scaled logistic regression, meaning their frequency is also predictive of nonadherence; the more they appear, the more likely it is that a patient is nonadherent.

IV. DISCUSSION

Our results show the possibility of predicting patient nonadherence using a combination of machine- and human-generated keywords. These keywords, when assembled into contextual categories, were shown to be statistically significant predictors of nonadherence status in the patient population we analyzed. This work expands on our

978-1-5386-2534-7/17/$31.00 ©2017 IEEE

understanding of nonadherence as a multifaceted problem by tackling its detection and prediction with a multifaceted and nuanced approach. This relationship between our key terms and nonadherence is significant because it allows us to ascertain various psychosocial factors not represented in structured data. The capacity to extract features like nonadherence from unstructured medical data can be used in various settings. Patient notes can be analyzed and flagged for nonadherence ”risk factors,” in settings like primary care or hemodialysis. This would allow for prophylactic measures to be taken to prevent poor outcomes. Areas of future study include the processing of different types of unstructured data, like social work and nursing notes. Our categories can also be expanded with findings from future annotations.

Of note, some words within the individual categories were more contributive than others; for example, the word ”ama” (against medical advice) was found exclusively in nonadherent cases and can be considered a strong predictor, whereas some of the more ubiquitous terms, like ”dementia,” were less strongly contributive. While this study focused on single words, it could be expanded by examining bigrams or even longer phrases predictive of nonadherence.

V. CONCLUSIONS

This study makes use of unsupervised natural language processing to expand on the work of a human annotator, and investigates a psychosocially nuanced interpretation of nonadherence in the ICU setting. We found several keyword categories predictive of nonadherence, and made use of unstructured (discharge summary) data to investigate factors predictive of or correlated with nonadherence. Medical nonadherence is not a simple, monolithic problem, and should not be treated as such; moving towards a more comprehensive and holistic analysis of adherence data is the first step to understanding and diminishing its occurrence and impact on our healthcare system.

VI. ACKNOWLEDGEMENTS

The authors would like to acknowledge Dr. Roger Mark for hosting Hedyeh Elahinia at the MIT Laboratory for Computational Physiology, where this research was performed. We also acknowledge Dr. Leo Celi and Dr. Rodrigo Deliberato for their clinical expertise and advice. We acknowledge Edward Moseley for his help in data extraction and annotator GUI development. Finally, we acknowledge Eric Carlson of Philips Research for identifying the 198 cases used in this study.

REFERENCES

[1] R. M. Benjamin, “Medication adherence: helping patients take their medicines as directed,” *Public health reports*, vol. 127, no. 1, p. 2, 2012.

[2] A. O. Iuga and M. J. McGuire, “Adherence and health care costs,” *Risk management and healthcare policy*, vol. 7, p. 35, 2014.

[3] R. Heynes, “Determinants of compliance: the disease and the mechanics of treatment,” 1979.

[4] L.-E. García-Pérez, M. Álvarez, T. Dilla, V. Gil-Guillén, and D. Orozco-Beltrán, “Adherence to therapies in patients with type 2 diabetes,” *Diabetes Therapy*, vol. 4, no. 2, p. 175, 2013.

[5] J. LaFleur and G. M. Oderda, “Methods to measure patient compliance with medication regimens,” *Journal of pain & palliative care pharmacotherapy*, vol. 18, no. 3, pp. 81–87, 2004.

[6] A. D. Association *et al.*, “Standards of medical care in diabetes,” *Diabetes care*, vol. 28, no. suppl 1, pp. s4–s36, 2005.

[7] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, “Mimic-iii, a freely accessible critical care database,” *Scientific data*, vol. 3, 2016.

[8] R. Řehůřek and P. Sojka, “Software Framework for Topic Modelling with Large Corpora,” in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, May 2010, pp. 45–50.

[9] M. A. *et.al*, “TensorFlow: Large-scale machine learning on heterogeneous systems,” 2015, software available from tensorflow.org. [Online]. Available: http://tensorflow.org/