Evaluation of Hypertension Treatment: A Computational Analysis

By

Kimberly Shoenbill, MD, MS

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

Clinical Investigation

at the

UNIVERSITY OF WISCONSIN-MADISON

2018

Date of final oral examination: 12/17/2018

This dissertation presented to members of the Final Oral Committee**:**

Eneida A. Mendonca, Associate Professor; Department of Biostatistics and Medical Informatics; Department of Pediatrics
Mark Craven, Professor; Department of Biostatistics and Medical Informatics; Department of Computer Science
Miguel Leal, Assistant Professor, CHS; Department of Medicine, Division of Cardiovascular Medicine
Maureen Smith, Professor; Department of Population Health Sciences; Department of Family Medicine
Christine Sorkness, Distinguished Professor; Department of Medicine, Division of Allergy and Immunology

ABSTRACT

Evaluation of Hypertension Treatment: A Computational Analysis

Kimberly Shoenbill, MD, MS

The Gap: Over 45% of the 85.7 million US adults with hypertension have uncontrolled blood pressure resulting in increased risks of death, stroke, and myocardial infarction. Guidelines on hypertension management include lifestyle modification and medication initiation as first line treatment. To improve hypertension control, it is important to determine the frequency and inter-relatedness of lifestyle modification, hypertension medication initiation, and variables predictive of treatment. Lifestyle modification data is usually documented in narrative form, making it “invisible” in evaluation of coded data.

Methods: Electronic health record data from 14,860 adult hypertension patients were analyzed using natural language processing, statistical methods, and machine learning algorithms to evaluate documentation of lifestyle modification, hypertension medication initiation, and predictors of treatment.

Results: Combined lifestyle modification (any advice and assessment) recall was 99.27%, precision 94.44%, and correct classification 88.15%. Within one year of hypertension onset over 78% of patients had documented lifestyle modification and 62.2% of patients needing hypertension medication did not have a documented prescription. 69.4% of those “untreated” patients had documented lifestyle modification. Machine learning classification accuracy, as measured with AUROC, was 0.813 for classification of lifestyle modification documentation at any time, 0.685 at ≤ 3 months, and 0.709 for classification of medication initiation within one year of meeting hypertension criteria.

Conclusion: Natural language processing and machine learning analyses provided new information on hypertension treatment and variables predictive of treatment. Knowledge gained from this approach to EHR data evaluation can inform and help improve hypertension treatment, care processes, and metric development.

Acknowledgements

I would like to thank my committee members for their teaching and guidance during my training at UW-Madison. Their insights, guidance, and suggestions have proven invaluable as I completed my PhD dissertation work and will continue to serve me well in my academic career.

Current committee members:

Eneida Mendonca, Associate Professor
Mark Craven, Professor
Miguel Leal, Assistant Professor CHS
Christine Sorkness, Distinguished Professor
Maureen Smith, Professor

Prior committee members:
Patricia Brennan, Emeritus Professor
Patrick McBride, Emeritus Professor

Funding sources:

Clinical and Translational Science Award (CTSA) program, through the NIH National Center for Advancing Translational Sciences (NCATS), grant no. ULITR002373. PI: A. Brasier.

NLM grant no. 5T15LM007359 to the Computation & Informatics in Biology and Medicine Training Program. PI: M. Craven.

UW-Madison Office of the Vice Chancellor for Research and Graduate Education Research – 2014 Fall Research Competition Award. “Predictors of Lifestyle Modification in Hypertension: A Computational Analysis”. PI: E. Mendonca.

University of North Carolina Chapel Hill, start-up funding. PI: K. Shoenbill.

Conflicts of Interest: None

Date of IRB clearance: 01/23/2018

Dedication

This work is dedicated to my family: you have been an endless source of support, encouragement, and inspiration throughout my (many) academic pursuits. Thank you!

Table of Contents

Tables and Figures

1. Introduction

1.1. Hypertension Prevalence and Treatment

Hypertension is aptly named the “silent killer” due to its stealthy attack on millions of people with resultant untimely deaths, heart attacks, strokes, and kidney disease in its victims.[1] Hypertension prevalence continues to rise with an estimated 85.7 million adults in the US and 1.13 billion adults worldwide having hypertension (defined as a systolic blood pressure at or above 140 mmHg or a diastolic blood pressure at or above 90 mmHg).[1,2] In the United States hypertension is a leading cause of disability and is ranked the third highest risk factor related to death.[3] With 1 in 3 US adults having hypertension, the total annual national cost related to this disease is 51 billion dollars.[4] Although a main cause and contributor to morbidity and mortality, hypertension risks can be minimized with implementation of treatment guidelines.[5-8] Many studies have shown multiple health benefits in lowering elevated blood pressure.[9-14] Despite guidelines and well-published information on the benefits of hypertension treatment, 45.6% of US adults with hypertension do not have their blood pressure under control and 15.9% are not aware they have hypertension.[1] Hypertension’s prevalence and its sequelae are expected to rise as the population ages unless successful hypertension prevention practices are instituted. The American Heart Association estimates the 2030 prevalence of hypertension will increase to approximately 41.4% of US adults.[15] First line treatment of hypertension includes lifestyle modification alone or with medication initiation. Increased research into actual use of these proven treatment recommendations in clinical practice is needed to improve hypertension control.[16-19]

This study undertakes this challenging task by evaluating lifestyle modification and hypertension medication initiation documentation within electronic health record notes of incident (not yet diagnosed) hypertension patients.

1.2. Lifestyle Modification

The top 7 US health risks for combined disability and death identified in the Global Burden of Disease Study 2016 were tobacco, dietary risks, high body mass index, alcohol and drug abuse, high blood pressure, high fasting plasma glucose, and high cholesterol.[3,20] These risks can be minimized through the implementation of lifestyle modifications.(Table 1)[5]

Lifestyle Modification Topic	Recommendation
Alcohol	Limit alcohol to no more than 1 oz of ethanol in most men (2 drinks) and 0.5 oz of ethanol in most women (1 drink) per day
DASH Diet (dietary approaches in hypertension)	Consume a diet low in fat (saturated and total) and high in fruits, vegetables, low-fat dairy products, and whole grains
Physical Activity	Increase physical activity to a goal of at least 30 minutes of moderate activity most days per week
Sodium	Reduce dietary sodium to 2.4 gm sodium (6 gm sodium chloride)
Tobacco	Tobacco cessation
Weight management	Maintain normal body weight (body mass index 18.5–24.9 kg/m^2)

Table 1. Lifestyle Modification Recommendations

National guidelines have recommended the use of lifestyle modification in the treatment of multiple prevalent disorders plaguing the US today including hypertension, obesity, coronary artery disease, diabetes and peripheral vascular disease.[21-26] Unfortunately,

how providers counsel and document these low-risk, effective interventions is poorly understood because this information is invisible in studies of coded data (i.e., studies limited to information obtained using diagnosis or procedure codes). Lifestyle modification is recommended by multiple hypertension guidelines as a critical first-step (with or without medication) to achieve hypertension control.[5,6,21,22,27,28]

Lifestyle modification is as effective as single-drug therapy in achieving hypertension control with documented systolic blood pressure reductions of 2-20 mmHg.[1,5-11,29-37] Unfortunately, provider counseling on lifestyle modification is under-utilized.[38-42] Despite proof of tremendous benefits of lifestyle modification on life expectancy and cardiovascular risk reduction, documentation of lifestyle modification assessment and advice has not been adequately examined.[43,44] Further research on how providers can help patients adopt diet and physical activity recommendations was called for in the 2013 American Heart Association/American College of Cardiology guideline.[22] My PhD work helps to fill this knowledge gap by identifying what is currently being recorded as lifestyle modification efforts and missed opportunities for lifestyle modification intervention.

1.3. Natural Language Processing

Overview

With lifestyle modification documentation usually buried within clinical notes, prior studies of lifestyle modification have focused on survey data or smaller chart reviews.[38,40,45-48] Historically, manual chart review has been used to abstract information from patient records, but this has proven to be a time- and labor-intensive process, making large-scale chart abstractions nearly impossible. In order to accomplish this task more

efficiently, natural language processing (NLP) can automatically extract text-based information from narrative notes. NLP tools can process many thousands of notes per hour.[49] This technology makes larger chart abstractions feasible and allows a more comprehensive evaluation of documentation of lifestyle modification. The natural language processing tool used in this study is the Clinical Text Analysis and Knowledge Extraction System (cTAKES). It is an open-source pipeline which processes clinical notes and identifies types of clinical named entities — drugs, diseases/disorders, signs/symptoms, anatomical sites, and procedures. A schematic of an NLP processor workflow is seen in Figure 1.[50]

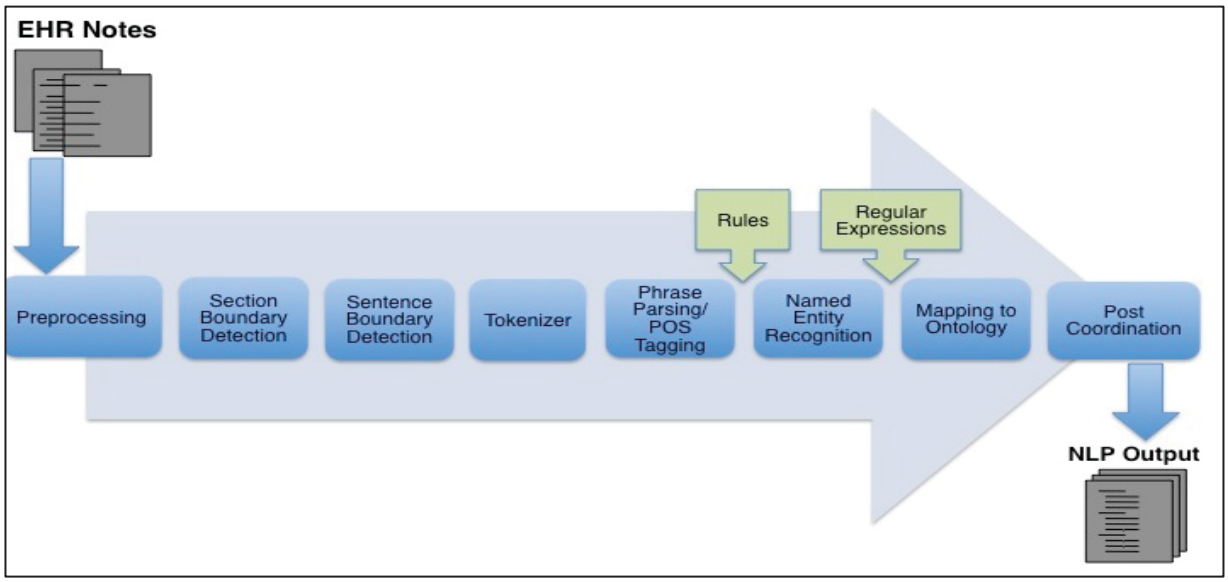


Figure 1. Natural language processing pipeline with enhancements

Each extracted named entity has attributes describing the text span, the ontology mapping code, context (family history of, current, unrelated to patient), and negated/not negated. The typical process starts with detection of each section of the text and then each sentence, followed by the identification of sentence tokens (words, dates, numbers, etc.) in the sentence. Part-of-speech is assigned to tokens (noun, preposition,

noun-phrase, etc.). The Named Entity Recognition component implements a dictionary look-up, so that each entity is mapped to a concept from the tool's dictionary. Post coordination then combines multiple concepts into a single one (e.g., low + cholesterol + diet = low cholesterol diet). To improve relevant term and phrase identification for this study, the process was augmented with additional components of rules and regular expressions (exemplified in Figure 1 by the green down-arrow boxes) and an expanded dictionary, beyond the cTAKES UMLS dictionary module, was used to extract lifestyle modification terms and phrases.

NLP has been employed to successfully extract data from electronic clinical records and has been applied in many fields for efficient and accurate chart abstraction.[51-54] Some studies have explored the identification of smoking status as an isolated finding.[55,56] One study used NLP for extraction of information on weight management counseling in postpartum visits and showed extraction capabilities similar to human abstractors.[57] However, to my knowledge, no prior comprehensive, automated extraction of all lifestyle modification activities recommended for hypertension treatment has been attempted. An example of output information from the natural language processing work is shown in Figure 2.

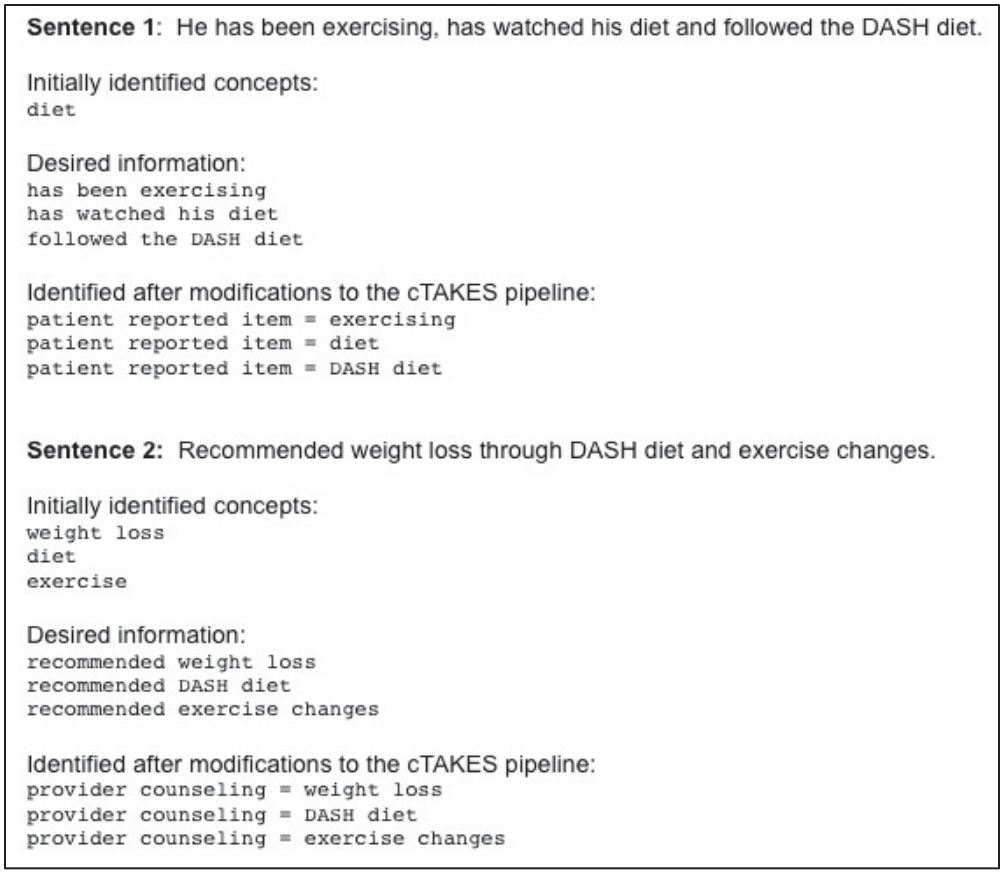


Figure 2. Example of output from natural language processing

Terminology Identification and Dictionary Enrichment

The cTAKES natural language processing tool dictionary could not, without enhancement, identify relevant lifestyle modification terms and phrases. Moreover, existing ontologies (terms and their interconnections) did not adequately describe and map lifestyle modification terms and phrases.[58-62] An empirical approach was used to create and iteratively enrich and refine the lifestyle modification terminology in three stages: (1) literature review to identify relevant terms, (2) identification of relevant terms in current ontologies (terms and their interconnections such as the ontologies in the National Center for Biomedical Ontology and the Consumer Health Vocabulary),[62] and

(3) domain expert input and evaluation of terms and phrases identified within the training set. From this iterative process, the lifestyle modification terminology was created and used as the dictionary for the enhanced NLP process.

1.4. Machine Learning

Machine learning is an automated approach to data analysis wherein statistical and algorithmic methods are used to build models to classify or predict an output variable using input variables and patterns learned from the data while requiring minimal explicit human instruction in model generation. Machine learning methods have been applied to a variety of medical problems to find new patterns or correlations within existing data.[63-67] This study focused on “supervised” machine learning methods. “Supervised” means the dataset contains the known class value for each patient. For example, if a patient had documented lifestyle modification at < 3 months, her class value is ≤ 3 months. Given a training set consisting of patient, provider, and clinic characteristics along with a known class value for each patient, a “supervised” machine-learning algorithm creates a model that represents the class variable as a function of the patient, provider, and clinic characteristics. Such models can provide insight into the factors that explain the class variable value. They can also be applied to previously unseen patients to predict their class variable values. Supervised machine learning algorithms used are described below in Figure 3 and included logistic regression, decision trees, and random forests. Zero R is a rule-based classifier that served as a “straw man” algorithm – a basic model that predicted the class value of a given instance (patient in this analysis) to be the majority (mode) value for the class of interest in the dataset. This machine learning algorithm is used as a baseline representation of results near random guessing.

Investigators can compare results from other algorithms to Zero R results to determine if the model from a given algorithm is better than randomly guessing the class value based on the dataset's known frequency of class values.

Zero R

- A learned model is a rule that assigns the majority class value to each new instance based on the training data
- The learning algorithm infers a model by predicting the majority class (mode class value) for each instance.
- The model classifies new instances as having the majority class value.
- Provides a baseline understanding of almost random prediction based solely on the most frequent class value in the dataset

Logistic regression.

- In a learned model a patient is represented as a vector of variables corresponding to a point in the hypothesis space of all clinic, patient and provider variables.
- The learning algorithm infers a model by transforming the input variables using the logistic function to the output variable. The output variable is a probability between 0 and 1. The coefficient for each variable is determined from the data to most accurately reproduce each patient's class in the training set.
- Class assignment is determined by applying a threshold value to the weighted sum of the variables for each patient. This value separates new instances into membership into one of the identified classes: 1) medication initiation yes; 2) medication initiation no.

Decision tree learner [68]

- A learned model is represented as an upside down tree with the “root” (at the top) containing all the data with a mixture of class variable values down to the “leaves” which contain data with only one (ideal) or a few (acceptable) class variable values
    - Looking at the tree, an observer can determine which characteristics (or combination of characteristics) is/are predictive (found in the branch path) leading out to the particular class variable value at the end of the branching (a leaf)
- A learning algorithm Iteratively splits the training set on input variable values, into subsets with purer collections of data based on the targeted class with the goal being to decrease heterogeneity in each subsequent subset.
- The model classifies new instances using the variables to split the data as per the training set to achieve a tree with pure leaves which contain instances of a single (or mostly single) class value.

Random Forests[69,70]

- A learned model is represented as a collection of decision trees from repeated resampling of the data to achieve the most homogenous leaves (classes) at the end of the branches
- The learning algorithm infers a model as in decision trees, but assigns the majority class of the many trees that were created for a given instance.
- Classification is done with assignment of the majority class value from all the tree models to a new nominal instance.

Figure 3. Classifier descriptions in machine learning analysis

Random forests are ensembles of trees that have demonstrated state-of-the-art predictive accuracy in a wide array of problem domains. Random forests were evaluated to see if a collection of trees could better predict the class value than a single tree.[69,70]

Machine learning methods have been applied to hypertension and cardiovascular domains with investigation of incident hypertension prediction, evaluation of determinants of successful hypertension treatment with medication, and identification of predictors of intermediate outcomes related to cardiovascular disease.[71-73] One article discussing future directions for use of machine learning in hypertension management calls out the need to incorporate lifestyle and environmental factors in establishing models to choose the most appropriate medication treatment for hypertension.[74]

2. Specific Aims

2.1. Overview

Systematic identification of lifestyle modification (including provider counseling on modifications and patient-reported lifestyle changes) is the first step toward improving its use in clinical practice and establishing it as a quality metric. Therefore, I analyzed clinical narratives and structured data from the electronic health records (EHRs) of 14,860 incident hypertension patients (i.e., patients who met criteria for a hypertension diagnosis but had not received documented ICD-9 diagnosis codes or medications for hypertension). Incident hypertension patients present an excellent patient-population to study lifestyle modification for many reasons: (1) lifestyle modification is recommended by multiple hypertension guidelines as a first-line treatment; (2) hypertension is increasingly prevalent and costly; (3) hypertension control is suboptimal in almost 46% of US hypertension patients; (4) lifestyle modification can be as effective as single-drug therapy in achieving hypertension control; (5) the percentage of eligible patients receiving lifestyle modification counseling is incompletely understood; and (6) delays in antihypertensive medication initiation have been identified but are incompletely understood. My study aims are described below.

2.2. Aim One

Refine existing dictionaries and natural language processing (NLP) methods to identify how lifestyle modification is recorded in electronic health records of adult hypertension patients. I iteratively developed and validated a list of term and phrases reflecting lifestyle modification based on literature review, expert domain knowledge, and terms and phrases found within EHR notes. After validation, the dictionary-

enhanced NLP extraction methods were used to extract information on the content and frequency of lifestyle modification from 14,360 patients' EHR notes.

Hypothesis: Lifestyle modification, as treatment for hypertension, can be extract and measured from electronic health records using NLP methods.

2.3. Aim Two

Determine clinical predictors of lifestyle modification assessment and advice as first-line treatment of hypertension. Using statistical hypothesis testing and existing machine learning algorithms, I analyzed data from 14,360 patients' EHR notes to identify clinical predictors of lifestyle modification assessment and advice.

Hypothesis: Using statistical and machine-learning methods, patient, provider, and clinic characteristics associated with lifestyle modification assessment and advice can be identified.

2.4. Aim Three

Determine clinical predictors of delays in medication initiation for hypertension. Using statistical hypothesis testing and existing machine learning algorithms, I analyzed data from 14,360 patients' EHR notes to identify clinical predictors of delays in medication initiation for hypertension. A delay in medication initiation is defined as no hypertension medication initiation within one year of hypertension onset for patients with persistent hypertension.

Hypothesis: Using statistical and machine-learning methods, patient, provider, and clinic characteristics associated with delays in medication initiation for hypertension can be identified

3. Methods

3.1. Methods Overview

This was a large, retrospective study of 14,860 incident hypertension patients' EHR records from UW Health. Each patient meets criteria for a diagnosis of hypertension, but at the start of the study period had not received an EHR-documented diagnosis or medication prescription for hypertension (i.e., no ICD-9 code for any hypertensive disease and no antihypertensive medication is recorded). Note: meeting hypertension criteria was determined by documented blood pressures, not dependent on provider-defined diagnosis as detailed under inclusion criteria below. This method for defining onset of hypertension has been used in prior research.[75] For each patient, extraction methods employed in this study allowed information on lifestyle modification counseling to contain both coded and text-based outpatient encounter notes and associated administrative data. This data was evaluated to identify types, frequencies, timing, and predictors of lifestyle modification. Data was also evaluated to determine frequency and predictors of hypertension medication initiation within one year of hypertension onset and its relation to lifestyle modification documentation.

3.2. Institutional and Clinical Settings

UW Health is the academic health system for the University of Wisconsin-Madison. UW Health serves a broad range of diverse populations, including academic and community clinics concentrated in south central Wisconsin in urban and rural communities. UW Health ambulatory practices include both large, multi-specialty clinics and community-based primary care clinics. They are located primarily within Dane County, which represents two-thirds of UW Health's patient base, but also are located in communities

across central Wisconsin. The affiliated physician group includes approximately 1,600 physicians. Approximately 2.8 million outpatient visits are completed each year.

3.3. Inclusion and Exclusion Criteria

Sample: Day one of study inclusion was defined as the date the patient reached clinical criteria for hypertension, but did not yet receive a coded diagnosis for hypertension or any hypertensive disease within the electronic health record.

Inclusion criteria: (from this retrospective patient population)

- Patients at least 18 years old “managed” at a UW Health practice between January 01, 2008 and December 31, 2011. “Managed” was defined as having at least two billable office encounters in an outpatient, non-urgent care primary care setting, or one primary care encounter and one office encounter in an urgent care setting (regardless of diagnosis code). Patients accrued time in the study with a maximum time of 4 years. Patients were censored due to death or no longer meeting criteria of being managed at a UW Health practice.
- Adult patients (regardless of gender, race or ethnicity) meeting criteria for the diagnosis of hypertension defined as follows: at least three separate elevated blood pressures within a two year period, measured at least 30 days apart, with systolic ≥ 140 mmHg or diastolic ≥ 90 mmHg; or two elevated blood pressures within a two year period, measured at least 30 days apart, with systolic ≥ 160 mmHg or diastolic ≥ 100 mmHg, as defined by guidelines from the Seventh Report of the Joint National Committee on the Prevention, Detection, Evaluation and Treatment of High Blood Pressure.[5] Hospital and emergency room blood pressures were excluded.

Exclusion criteria:

- Patients pregnant during the retrospective study period, excluded 1 year before, during, and 1 year after pregnancy.
- Patients with a pre-existing diagnosis of hypertension including essential hypertension, hypertensive heart disease, hypertensive renal disease, hypertensive heart and renal disease, secondary hypertension, or any anti-hypertensive medication prescription.
- Children under the age of 18 (because guidelines being used are for adult hypertension patients).
- Prisoners (as a vulnerable population) were excluded: IRB approval was not obtained for inclusion of this population given their limited freedom to control lifestyle modification interventions.

3.4. Data Sources

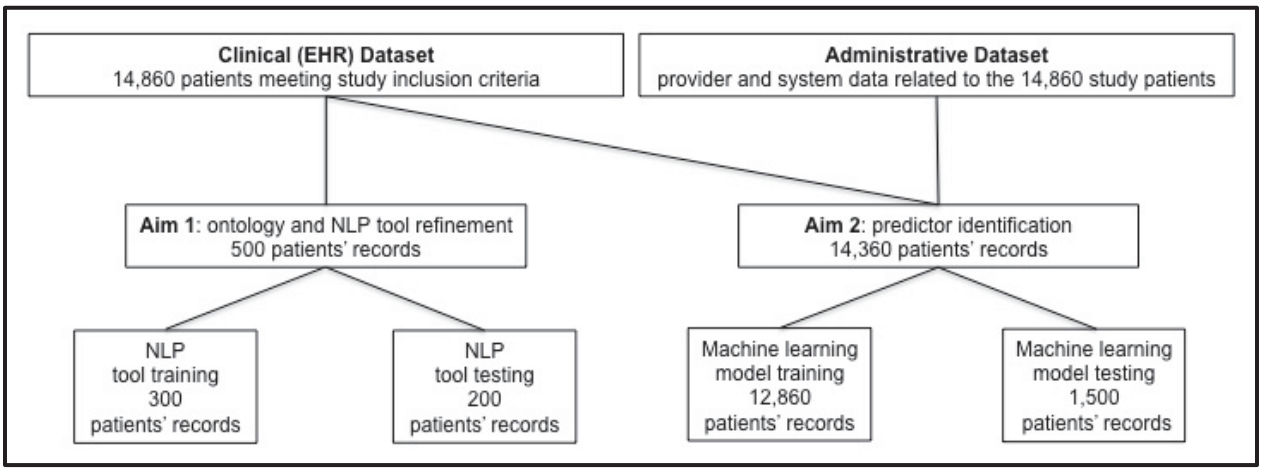


Figure 4. Graphical representation of use of the study datasets

For the clinical data set, a sample of 14,860 patients met inclusion criteria. The data set contained longitudinal information about the patients, including coded instances (e.g.,

demographics, laboratory tests, diagnoses) and textual data (e.g., clinical narratives and reports). All relevant full-text documents from patients meeting inclusion criteria were extracted from the UW Health electronic health record along with coded data. The administrative data set was comprised of UW Health provider and systems data related to patients in the study population. Figure 4 illustrates how the data sets were used in each phase of this proposal. The subsets of patients (as shown in the figure) were randomly selected. IRB approval was obtained for use of the datasets in this study. A copy of the IRB approval document has been included in Appendix E.

Full details of methods and further data divisions for each specific aim are provided in the manuscripts (Chapter 4 sections 4.1 - 4.3). For specific Aims 2 and 3, class variables of time to lifestyle documentation or medication initiation was calculated from day one of study entry (which was when a patient fulfilled inclusion criteria above) until the first time lifestyle modification was documented (Aim 2) or the first time a prescription for medication was documented (Aim 3).

4. Manuscripts

4.1. Natural Language Processing of Lifestyle Modification Documentation

4.2. Identifying Variables Associated with Lifestyle Modification Documentation for Hypertension

4.3. Hypertension Medication Initiation: Statistical and Machine Learning Analyse

4.1. Natural Language Processing of Lifestyle Modification Documentation

Coauthors
Kimberly Shoenbill; Department of Family Medicine; Program on Health and Clinical Informatics; University of North Carolina-Chapel Hill; Chapel Hill, NC, USA

Yiqiang Song; Department of Biostatistics and Medical Informatics; University of Wisconsin-Madison; Madison, WI, USA

Lisa Gress; Department of Biostatistics and Medical Informatics; University of Wisconsin-Madison; Madison, WI, USA

Heather Johnson; Department of Medicine, Division of Cardiovascular Medicine; University of Wisconsin-Madison; Madison, WI, USA

Maureen Smith; Department of Population Health Sciences; Department of Family Medicine; University of Wisconsin-Madison; Madison, WI, USA

Eneida A. Mendonca; Department of Biostatistics and Medical Informatics; Department of Pediatrics; University of Wisconsin-Madison; Madison, WI, USA

Keywords - MeSH Terms
Electronic Health Records, Health Behavior, Hypertension, Life Style, Natural Language Processing

Accepted for publication 11/03/2018, Health Informatics Journal

ABSTRACT

Lifestyle modification, including diet, exercise, and tobacco cessation, is the first-line treatment of many disorders including hypertension, obesity, and diabetes. Lifestyle modification data are not easily extracted or used due to their textual nature. This study addresses this knowledge gap by using natural language processing to automatically identify lifestyle modification documentation from electronic health records.

Electronic health record notes from hypertension patients were analyzed using an open-source natural language processing tool to retrieve assessment and advice regarding lifestyle modification. These data were classified as lifestyle modification assessment or advice and mapped to a coded standard ontology. Combined lifestyle modification (advice and assessment) recall was 99.27%, precision 94.44%, and correct classification 88.15%.

Through extraction and transformation of narrative lifestyle modification data to coded data, this critical information can be used in research, metric development, and quality improvement efforts regarding care delivery for multiple medical conditions that benefit from lifestyle modification.

Key words

Electronic Health Records, Health Behavior, Hypertension, Life Style, Natural Language Processing

BACKGROUND AND SIGNIFICANCE

National guidelines have recommended the use of lifestyle modification in the treatment and prevention of prevalent disorders plaguing the US today including hypertension, obesity, coronary artery disease, diabetes, peripheral vascular disease, and cancer.[1-9] The top 7 US health risks for combined disability and death identified in the Global Burden of Disease Study 2016 were tobacco, dietary risks, high body mass index, alcohol and drug abuse, high blood pressure, high fasting plasma glucose, and high cholesterol. [10,11] These risks can be minimized through the implementation of lifestyle modifications (e.g., tobacco cessation, dietary changes, alcohol moderation, illicit drug use abstinence, increased aerobic exercise, and weight loss to achieve a healthy weight). The efficacy of addressing lifestyle in just one counseling session has been shown.[12] For example, behavior change can be elicited by making patients aware of their elevated blood pressure readings after a provider encounter.[13] Despite the effectiveness and importance of lifestyle modification in treating hypertension and many chronic illnesses, lifestyle modification is under-utilized and not easily measured, due to it being buried in narrative text rather than in consistent coded form (e.g., diagnosis or procedure codes).[14-21] This limits accurate evaluation of the efficacy of these interventions, tracking of quality care metrics, and reimbursement. Evaluation of lifestyle modification includes both the assessment of lifestyle modification activities as reported by a patient or observed by a provider (e.g., “patient started running and weight is down”) and advice on lifestyle modification activities given by a provider to a patient (e.g., “recommend patient lose 30 pounds”). Systematic identification of lifestyle modification is the first step toward improving its use in clinical practice and establishing it as a quality metric. While some clinical information systems code limited individual

behaviors (e.g., smoking history), much of this information continues to be recorded primarily in narrative form.

There is a need for automated methods that can facilitate the extraction and integration of lifestyle behavior factors for use in research. Historically, manual chart review has been used to abstract information from patient records, but this has proven to be a time and labor-intensive process, making large-scale chart abstractions nearly impossible. In order to accomplish this task more efficiently, this study used natural language processing (NLP) software tools and processes that can automatically extract text-based information. NLP tools can process many thousands of notes per hour.[22] This technology makes larger chart abstractions feasible and allows a more comprehensive evaluation of documentation of lifestyle modification.

NLP has been used to successfully extract data from electronic clinical records and applied in many fields for efficient and accurate chart abstraction.[23-27] Some studies have explored the automated extraction of information on smoking status as an isolated finding.[28-31] One study looked at NLP tool augmentation to extract cardiovascular risk identification.[32] Another study used the MediClass NLP tool for extraction of information on weight management counseling in postpartum visits and showed extraction capabilities similar to human abstractors.[33] A separate study extracted limited lifestyle modification documentation in evaluation of diabetes management.[34] And multiple prior studies of lifestyle modification have used survey data, with potential bias concerns, or small numbers of recorded patient encounters, with potential lack of generalizability and power. To our knowledge, this work is the first to evaluate lifestyle modification for hypertension in a large population using automated methods.

The primary objective of this study was to use an existing open-source natural language processing tool, cTAKES, along with rules and regular expressions on existing electronic health records to make previously invisible data on lifestyle modification documentation visible and ready for analysis. We used an existing data set that had been used in prior evaluations of hypertension treatment and clinical inertia.[35] This data set was chosen because it could accomplish two goals: (1) evaluate the feasibility of automatic extraction of lifestyle modification (LM) using an open source NLP tool allowing use and application to multiple chronic disease evaluations at different institutions, and (2) facilitate extension of prior work in this patient population to improve care of hypertension patients early in their disease process. Methods used in this study are designed so that they can be applied to many other patient populations including those with other chronic diseases such as obesity, coronary artery disease, diabetes, and peripheral vascular disease.

METHODS AND MATERIALS

Institutional and clinical settings: School of Medicine and Public Health of the University of Wisconsin-Madison caring from more than 600 thousand patients each year. UW Health adopted the Epic Systems Corporation electronic health record in the early 2000s. The data from the Epic system is also stored in an enterprise data warehouse, which supports the use of electronic health records data for clinical operations and research. All relevant full-text documents from patients meeting inclusion criteria were extracted from the UW Health electronic health record. An overview diagram of this study's steps is provided in Figure 1 with more details of the steps and methods following.

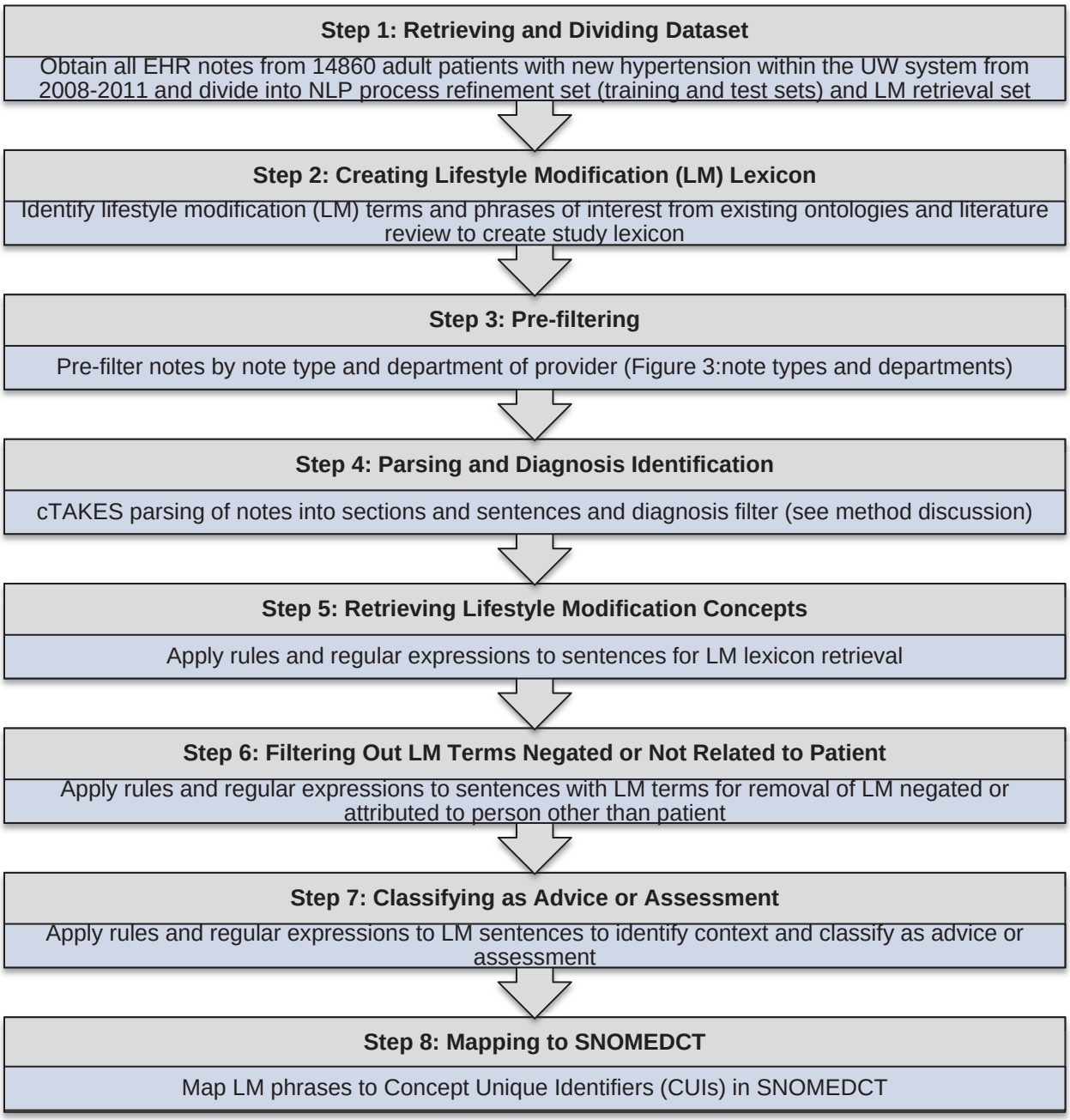


Figure 1: Overview of study steps

Step 1: Extracting and dividing dataset

Study inclusion and exclusion criteria are detailed in Table 1.[35]

Inclusion criteria: (from this retrospective patient population)

• Adult patients ≥18 years old "managed" at a UW Health practice between January 01, 2008 and December 31, 2011. "Managed" is defined as having at least two billable office encounters in an outpatient, non-urgent care primary care setting, or one primary care encounter and one office encounter in an urgent care setting (regardless of diagnosis code).

• Adult patients (regardless of gender, race, or ethnicity) meeting criteria for the diagnosis of hypertension defined as follows by guidelines from the Seventh Report of the Joint National Committee on the Prevention, Detection, Evaluation and Treatment of High Blood Pressure.[36,37] Hypertension is defined as:
  ◦ ≥ **three** separate elevated blood pressures within a two year period, measured at least 30 days apart with: systolic ≥ 140 mmHg or diastolic ≥ 90 mmHg
  ◦ or ≥ **two** elevated blood pressures within a two year period, measured at least 30 days apart with: systolic ≥ 160 mmHg or diastolic ≥ 100 mmHg

Exclusion criteria:
• Patients pregnant during the retrospective study period were excluded 1 year before, during, and 1 year after pregnancy.
• Patients with a pre-existing diagnosis of hypertension including essential hypertension, hypertensive heart disease, hypertensive renal disease, hypertensive heart and renal disease, secondary hypertension, or any anti-hypertensive medication prescription.
• Children under the age of 18 (because we employed guidelines on lifestyle modification for adults).
• Prisoners (as a vulnerable population) were excluded as IRB approval was not obtained for inclusion of this population due to their limited freedom to choose lifestyle activities.

Table 1: Inclusion and exclusion criteria

The hypertension diagnosis criteria were based on JNC 7 criteria, reflecting the guidelines available during the time of the hypertension dataset creation. However, since this analysis focuses on lifestyle modifications, more recent guidelines including, JNC 8[36-39] and the 2017 American Heart Association's guidelines reflect similar lifestyle modification recommendations. LM concepts reflect all three sources. Blood pressure measurements were extracted from discrete fields within the EHR. Preexisting conditions were identified using ICD9 codes. The 14,860 patient dataset was divided

into a 500 patient NLP tool refinement set and a 14,360 patient lifestyle modification extraction set (Figure 2).

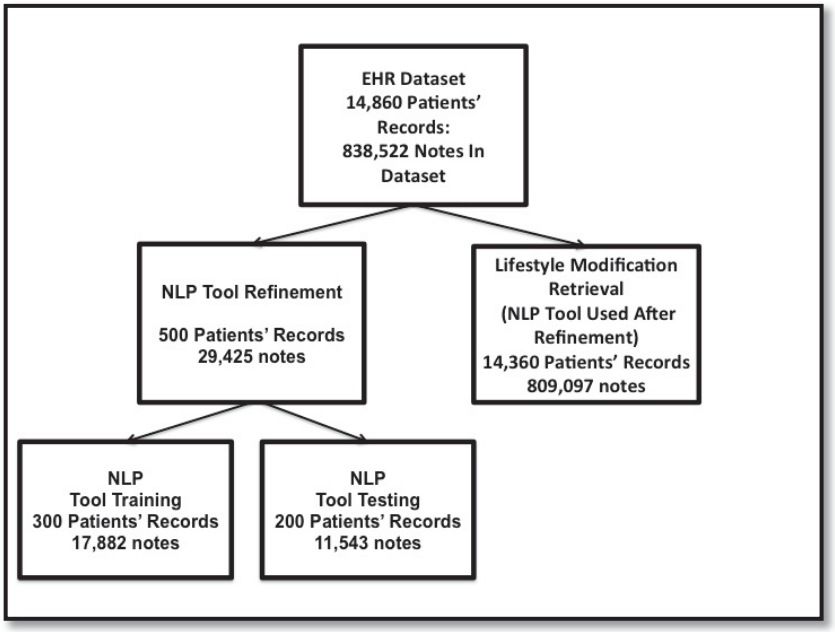


Figure 2: Dataset divisions

The subset of patient notes for NLP tool refinement was randomly selected using Python's random module: "random sample". [40,41] This subset was comprised of notes from throughout the outpatient clinical encounter including nursing notes, provider notes, patient instructions, nutrition consultation, and exercise consultation. The University of Wisconsin IRB approved this study. Baseline characteristics of the study population are listed in Table 2 with the 500 patient subset characteristics listed beside the 14,860 entire data set characteristics.

Baseline Characteristics	**Entire Data Set: 14860 Patients**	**NLP Refinement Subset: 500 Patients**
Age (years, mean, median, standard deviation)	18-80, Mean 49.21, Median 49.00, SD 14.70	18-91, Mean 49.47, Median 50.00, SD 15
Gender	**Number of patients (% of dataset)**	**Number of patients (% of dataset)**

Female	7363 **(49.55%)**	261 **(52.20%)**
Male	7497 **(50.45%)**	239 **(47.80%)**
Race/ethnicity	**Number of patients (% of dataset)**	**Number of patients (% of dataset)**
African American/Black	703 **(4.73%)**	24 **(4.80%)**
Asian	216 **(1.45%)**	7 **(1.40%)**
American Indian/Alaska Native	50 **(0.34%)**	0 **(0%)**
White	13140 **(88.43%)**	438 **(87.60%)**
Hispanic/Latino	281 **(1.89%)**	13 **(2.6%)**
Other/Native Hawaiian/Pacific Islander/Multiple	71 **(0.48%)**	2 **(0.40%)**
Unknown	399 **(2.69%)**	16 **(3.20%)**

Table 2: Study subjects' baseline characteristics

Step 2: Creating lifestyle modification (LM) lexicon

An empirical method was used to create and iteratively enrich and refine the lifestyle modification terminology using four approaches: (1) literature review to identify related terms, including terms and concepts discussed in JNC 7 and 8; (2) existing ontology review (terms and their interconnections) to identify relevant terms including acronyms and abbreviations such as those in the SNOMED CT ontology, the National Center for Biomedical Ontology and the Consumer Health Vocabulary;[42] (3) domain expert collaboration to identify words, acronyms, abbreviations, and phrases relevant to hypertension and lifestyle modification; and (4) electronic health record note training. Identified terms and phrases from this four-fold process generated the initial list of relevant terms and phrases for extraction (Table 3).

Provider Verb Phrase Terms Requiring Any Lifestyle Modification Object: Advice	**Lifestyle Modification Object Terms (May Be Part of Provider or Patient Verb Phrase)**
advise/d/ing avoid begin change congratulated continue counsel/ed/ing covered recommend/recommended refer/referral	alcohol cholesterol diet dietician exercise fat/fats gym health club physical activity salt
Patient Verb Phrase Terms Requiring Any Lifestyle Modification Object: Assessment	**Declaration of Lifestyle Modification - Requiring No LM Object: Assessment**
changed/changing continued/continues/continuing decreased/decreasing eats/eating improved/improving increased/increasing interested in motivated to plans/planning to reports/reported	alcoholic tobacco use change in weight cigarettes: cigars: DASH/dash diet decreased weight down ##/pounds/lb/lbs healthy diet physically active

Table 3: Example of terms and phrases in lifestyle modification lexicon

Our goal was to iteratively extend the NLP tool to have additional capabilities to handle discourse. Terms and phrases were mapped to codes in the UMLS and semantic types based on the UMLS semantic net. (Details available upon request as Appendices A and B).

Step 3: Pre-filtering to identify notes of interest

The total dataset was comprised of 14,860 patients' EHR notes with an average of 56 notes per patient. Each note was comprised of multiple sentences, some with multiple concepts of interest (lifestyle modification, family history). Many notes were not relevant

to lifestyle modification, so pre-filtering for note types and departments likely to have documented lifestyle modification was performed (Figure 3 shows LM-relevant note types and departments agreed upon by three study physicians).

We included pediatric departments because NIH defines children as those persons 18-21 years old and some patients in this age range may continue care in pediatric clinics. Gynecology was also included as a primary care clinic because many women see gynecologists as their primary care provider.

Note Type	**Note Department**
• Alcohol Brief Assessment • Assessment and Plan • Consult • Consult Follow-Up • History and Physical • Letter • Patient Instructions • Progress Note	• Cardiology • Cardio Prevention • Cardio Rehab • Diabetes • Family Medicine • Gynecology • Health Education • Internal Medicine • Nutrition • Pediatric Cardiology • Pediatrics

Figure 3: Note types and departments evaluated

Step 4: Parsing and diagnosis identification

After pre-filtering 14,331 notes were in the NLP tool refinement set and 403,018 notes were in the LM extraction set. These notes were processed using the Clinical Text Analysis and Knowledge Extraction System (cTAKES). This is an open-source NLP pipeline that processes clinical notes and identifies types of clinical named entities — drugs, diseases/disorders, signs/symptoms, anatomical sites and procedures.[43] Each named entity has attributes for the text span, the ontology mapping code, context (e.g., family history of, current, unrelated to patient), and negated/not negated. Figure 4 shows components of a typical NLP pipeline (with study enhancements to improve relevant term and phrase identification depicted by the down-arrow boxes). The typical

process starts with detection of each section of the text and then each sentence, followed by the identification of sentence tokens (e.g., words, dates, numbers) in the sentence. A part-of-speech is assigned to tokens (e.g., noun, preposition, noun-phrase). The Named Entity Recognition component implements a dictionary look-up, so that each entity is mapped to a concept from the tool dictionary. Post-coordination then combines multiple concepts into a single one (e.g., DASH + diet).

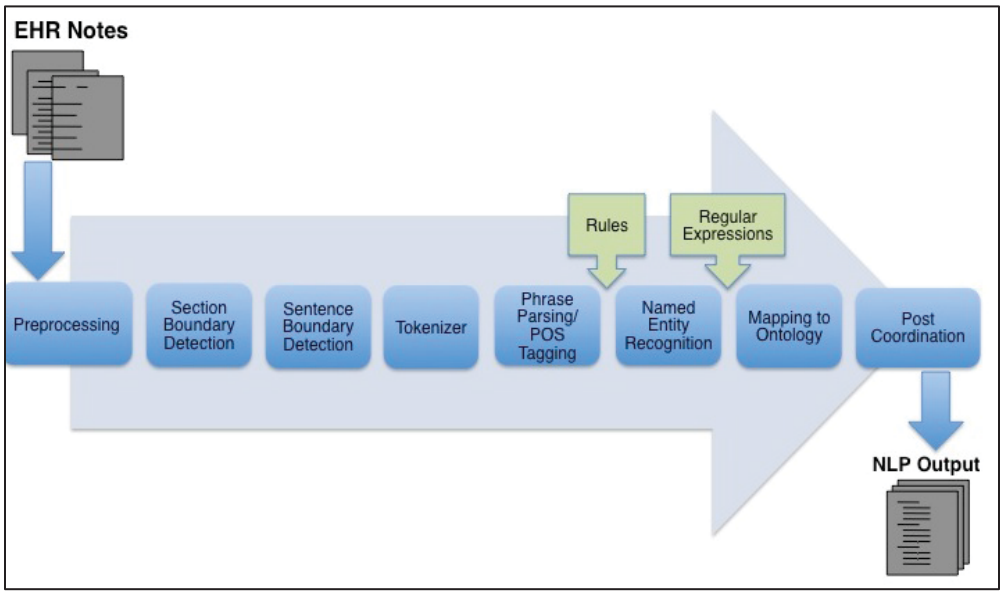


Figure 4: Augmented cTAKES processing steps

Notes were further processed using Python code to retrieve notes with diagnoses likely to have LM (e.g., obesity, diabetes – details of diagnosis filter terms available by request as Appendix C).

Steps 5-8: Extracting LM, filtering out notes with negated or LM terms not related to patient, and mapping

This study initially tried to use cTAKES as the sole method to identify relevant LM, but this approach resulted in poor precision (67.15%) and poor recall (62.48%) because:

1. Searching for semantic types of interest (TUIs) was too broad and extracted many irrelevant items (e.g., “diet counseling” = TUI T033 = “Finding”. Many irrelevant terms are “Findings” such as “sweaty palms” which was pulled as a relevant extraction when searching based on TUI alone).
2. Using concept unique identifiers (CUIs) was too restrictive due to identifying only a specific term and inability to identify phrases with LM (e.g., “exercise” = CUI C0015259. If searching on this CUI, relevant terms such as “swimming” with CUI C0039003 were missed).
3. Using cTAKES’ extraction and dictionary lookup effectively identified nouns, but missed verbs and verb phrases of LM documentation (e.g., extracted “diet”, missed “counseled on diet changes").

Therefore, this study used a combined approach to processing the notes. cTAKES was used for parsing the notes into sections and sentences, but instead of using the parsed parts of speech identified by cTAKES for ontology mapping, Python code with regular expressions and rules was used to process each sentence to identify lifestyle modification terms and phrases from the created lexicon. This enhanced the natural language processing workflow to facilitate extraction of data related to lifestyle modification, including social and behavioral factors that historically have been difficult to extract and verbs. These data were mapped to corresponding terms and codes within the SNOMED CT Ontology using the Unified Medical Language System (UMLS).
For example, a provider recommendation of: “cut down on alcohol intake” was extracted using the study’s enhanced NLP methods resulting in lifestyle modification advice extraction: “alcohol consumption counseling” with SNOMED CT CUI = C1531491 and

TUI = Health Care Activity T058. Without enhancements to NLP, no match for this phrase was extracted or mapped by cTAKES to the SNOMED CT ontology.

In order to capture context-sensitive phrases we used semantic and syntactic rules, as well as regular expressions. Due to many notes containing unusual syntactic expressions impeding extraction of context sensitive lifestyle modification, existing note punctuation was modified. Specifically, when a sentence about LM contained a colon (:), the NLP processor read this as a hard stop and classified concepts after it as belonging to the next sentence. The colon mark was replaced with a comma to allow capture of concepts after the colon mark as contextually related to the first phrase. This improved extraction of the entire lifestyle modification context. For example, prior to modification, cTAKES identified the negated history of “Patient smoking: no.” as a positive history of patient tobacco use: “Patient smoking. No.” After modification, the intended meaning was extracted: “patient history negative for smoking.” Tokenization per cTAKES was not modified, but cTAKES only identifies nouns and noun phrases and this study also needed to identify verbs and verb phrases. Therefore, the algorithm searched the entire sentence using regular expressions for LM object terms and phrases (Figure 5 and details of LM lexicon available upon request as Appendix A).

Example sentence: discussed importance of regular aerobic exercise

The regular expression: 'exercis(e |ing|e\\.|ed|e,|e;)' extracts:
'exercise'
'exercise,'
'exercise.'
'exercising'
'exercised'
'exercise;'

Does not extract 'exercises' as this often referred to particular physical therapy interventions (e.g., “doing shoulder strengthening exercises regularly”).

Figure 5: Example of regular expression to capture exercise

Negation identification was performed using negated regular expressions already employed to extract LM terms and phrases. This approach allowed for more efficient and thorough negation identification of verbs and verb phrases that were not identified by cTAKES’ negation module. Attribution of a concept, as pertaining to the patient or other person for LM terms/phrases and diagnoses, was expanded to include, in addition to mother and father, grandparents, grandmother, grandfather, roommate, spouse, husband, wife, son, daughter, partner, roommate, parent, friend, significant other, coworker, brother, sister, aunt, and uncle. Family history extraction using name of relation (e.g., mom, sister etc.) within rules and regular expressions, along with mapping schemes to SNOMED CT, were created specifically for this project. Mapping of terms to the SNOMED CT hierarchy was restricted to parent and child concepts, and their CUIs, to reduce granularity and facilitate effective use of this coded data in future machine learning projects that required less dimensionality. For example, “brother with non-insulin-dependent diabetes mellitus” was mapped to “FH: diabetes mellitus”, a child of the parent concept “FH: metabolic disorder”, and not mapped to the grandchild concept “FH: diabetes mellitus type 2”. This allowed all diabetics to be binned under one code. Details of family history mapping available upon request as Appendix D.

This study also classified lifestyle modification words and phrases as “assessment” and “advice” as was done in a prior analysis of videotaped encounters involving lifestyle modification.[44] Classification as assessment or advice was based on logic rules that included key words. This allowed for evaluation of LM documentation as either provider documentation of LM activities that the patient reported (e.g., patient reported exercising=assessment) or provider documentation of advice that s/he offered to the patient (e.g. recommended exercise=advice). Matching of LM terms to SNOMED CT was further refined with specific LM terms and phrases grouped under overarching concepts such as: exercise education (advice) or exercise history (assessment). Details of advice/assessment phrases available upon request as Appendix B. This process continued until performance was close to reproducing the manually coded set of terms and phrases, and additional modifications to the system minimally altered NLP tool performance. The training/testing tasks and iterative process details are shown in Table 3.

Annotation Task	**Total Number of Training Notes Manually Annotated and Used in Training (batch counts)**	**Number of Iterations (some batches trained on multiple times)**	**Total Number of Testing Notes Manually Annotated and Used in Testing**	**Annotators’ Initials**
Diagnosis Filter	1,000 (500 x 2)	2	500	LG, KS
Family History Extraction	1,200 (300 x 4)	4	1,500	EM, KS
Family History CUI Mapping	1,200 (300 x 4)	4	1,500	EM, KS
Lifestyle Modification Extraction	1,200 (300 x 4)	12	1,500	EM, KS

Lifestyle Modification CUI Mapping	1,200 (300 x 4)	4	1,500	EM, KS

Table 3: Annotation task iteration details.

Overfitting was discussed by researchers (KS, EM, YS) after each iteration to determine if the extracted terms/phrases from that iteration reflected relevant and generalizable results, or included concepts or terms reflective only of this patient sample. All three members reached consensus in determining if changes made to the tool were to be kept or represented overfitting and should not be retained in the tool development process.

This iterative process resulted in extraction of textual LM documentation and transformation into coded data ready for future statistical and machine learning analyses. (Figure 6)

SENTENCE 1: She has never smoked and drinks 1 alcoholic beverage daily.
Initially identified lifestyle modification relevant concepts:
alcoholic

Desired information:
assessment: nonsmoker
assessment: alcohol use

Information after augmentation of natural language processing steps and cTAKES:
Smoking assessment: C3853073
Assessment of alcohol use: C4076406

SENTENCE 2: Preventive health topics covered, healthy approach to weight loss general issues around healthy diet importance of regular aerobic exercise.
Initially identified lifestyle modification relevant concepts:
weight
diet
exercise

Desired information:
advice: weight management
advice: diet
advice: exercise

Information after augmentation of natural language processing steps and cTAKES:
Patient advised about weight management: C3697318
Dietary management education, guidance, and counseling: C1828150
Exercise education: C0582396

Figure 6: Example of LM extraction and transformation to coded data

Validation of the extended tool

The augmented NLP process was applied to the testing data set as the manually annotated gold standard. Dataset divisions and uses are detailed in Figure 7 and Table 3 describes researchers and numbers of notes employed in each test set evaluation.

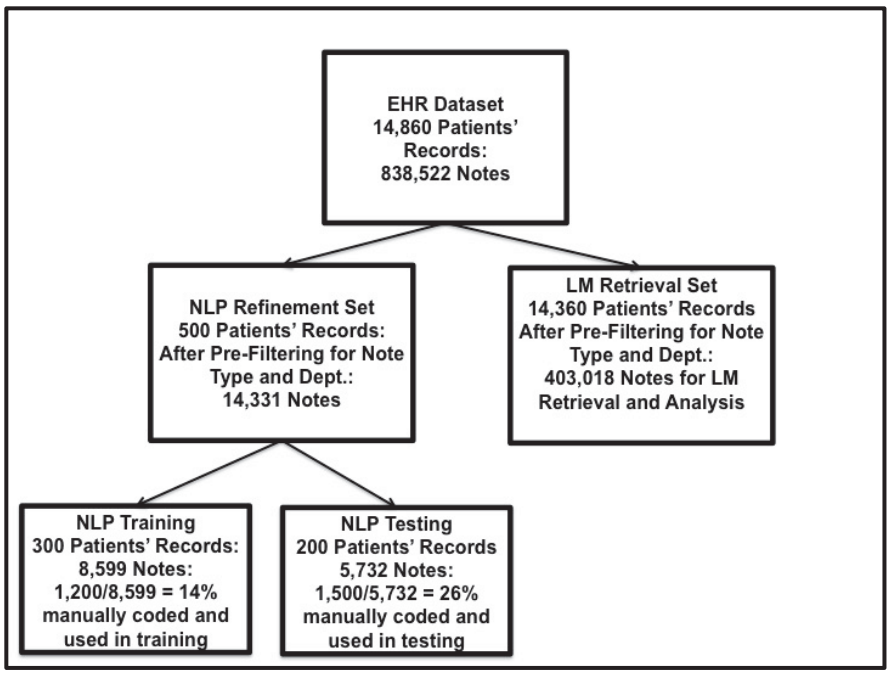


Figure 7: After pre-filtering for note type and department – data divisions and uses

Two sets of domain experts were employed to establish validity in each test task. One team worked on validation of the diagnosis filter and consisted of one clinical data analyst trained in linguistics and healthcare terminologies and one physician (LG, KS). The other team consisted of two physicians who worked on validation of the LM extraction and CUI mapping, and FH extraction and CUI mapping, and LM classification as assessment or advice (EM, KS). Each validation task used the randomly selected test set of patients' records that each team's members manually abstracted independently. From these manual abstractions inter-annotator agreements were calculated using percent agreement and kappa coefficient calculations. Validation of the enhanced NLP process is reported using precision, recall, and F measurements in Table 4. For this study, precision is defined as the number of terms correctly extracted divided by the sum of the number of terms correctly extracted and the number of terms

incorrectly extracted. Formally this is: Precision = TP / (TP + FP) where TP = true positive (a term that was correctly extracted) and FP = false positive (a term that was extracted as relevant but should not have been). Recall is the number of terms correctly extracted divided by the sum of the number of terms correctly extracted and the number of terms that should have been extracted but were missed. Formally this is: Recall = TP / TP + FN where FN = false negative (a term that was not extracted but should have been). The F measure is the harmonic mean between precision and recall with F = 2 [(precision x recall) / (precision + recall)]. In addition to lifestyle modification extraction and CUI mapping, our process was able to identify attribution and mapping to family history CUIs, and LM documentation type (i.e., assessment or advice). Success of classification of LM documentation as advice or assessment was evaluated using percent of LM terms and phrases correctly. Unit of measure for classification as advice or assessment was each LM concept (e.g., if a sentence contained two LM concepts with one being “advised weight loss” and classified correctly and the other being “patient reports gaining weight” and not correctly classified, then one concept would be counted as correct and one counted as incorrect classification).

RESULTS

Results from testing the NLP tool refinement process for combined lifestyle modification extraction were excellent with 99.27% recall and 94.44% precision and an F-Measure of 96.79%. CUI mapping for lifestyle modification was also very good with phrases correctly classified as advice or assessment 88.15%. These results, along with the excellent diagnosis filter testing results and family history extraction and mapping results are shown in detail in Table 4. Inter-annotator agreement was also excellent with

initial agreement percentages (range 96.15%-98.59%) and kappa scores (range 0.847-0.935) (Table 4). In the rare occurrence of reviewers extracting a different number of terms, the consensus-agreed-upon terms were used as the gold standard to calculate inter-rater agreement. After consensus discussions, 100% agreement was reached between annotators for each task.

Measurement: Testing of NLP Tool Refinement							
NLP Task	**Recall**	**Precision**	**F-Measure**	**Initial Annotator % Agreement**	**Initial IAA Kappa *(SE, 95% CI)***	**% Correct Classification as Advice or Assessment**	**Correct CUI Mapping**
Diagnosis Filter Extraction	98.70%	100%	99.35%	97.80%	0.920 *(0.024, 0.874-0.967)*	N/A	N/A
Family History Extraction	95.24%	81.63%	87.90%	96.15%	0.923 *(0.044, 0.838-1.000)*	N/A	N/A
Family History CUI Mapping	N/A	N/A	N/A	97.96%	0.935 *(0.064, 0.808-1.000)*	N/A	81.63%
Combined LM Extraction	99.27%	94.44%	96.79%	96.84%	0.847 *(0.066, 0.717-0.977)*	N/A	N/A
Combined LM CUI Mapping	N/A	N/A	N/A	98.59%	0.868 *(0.092, 0.687-1.00)*	88.15%	93.66%

Table 4: NLP augmentation validation performed on 1,500 manually annotated notes from the 200-patient testing subset. LM = lifestyle modification, SE = standard error, CI = confidence interval, CUI = concept unique identifier, N/A means measurement not applicable to task.

Each specific type of lifestyle modification extraction recall and precision is detailed in Table 5. Overall extraction testing results for each type of lifestyle modification were very good.

Specific LM Type Recall and Precision					
NLP Extraction of LM Advice	Recall	Precision	**NLP Extraction of LM Assessment**	Recall	Precision
Combined LM Advice	97.82%	97.82%	Combined LM Assessment	98.80%	91.21%
Dietary mgmt. education, guidance, and counseling	93.75%	93.75%	Dietary history	100%	100%
Exercise education	100%	100%	Exercise history	100%	100%
Patient advised about weight mgmt.	100%	100%	Weight finding	100%	96.55%
Smoking cessation assistance	100%	100%	Smoking assessment	100%	92.31%
Alcohol counseling	100%	100%	Alcohol use assessment	87.50%	87.50%
Drug addiction counseling	100%	100%	Drug use assessment	100%	100%

Table 5: Recall and precision of specific types of LM concepts

Of the 14,360 patients in the LM extraction set, 11,252 patients (78.36%) had notes documenting lifestyle modification. Each patient had an average of 56 notes in the initial dataset. From the total 809,097 notes in the LM extraction set (NLP refinement set removed), after filters and processes described in Figure 1 were applied, 47,838 notes had at least one documentation of lifestyle modification. Many notes contained more than one documented LM activity. Specific lifestyle modification activities and their CUIs are detailed in Table 6 with patient and note counts for each type of LM.

Lifestyle Modification Advice **SNOMED CT Concept**	**Concept Unique Identifier Code**	**Patient Counts**	**Note Counts**
Dietary mgmt. education, guidance, and counseling	C1828150	7589	19963
Exercise education	C0582396	7349	17829
Patient advised about weight management	C3697318	5268	11373
Smoking cessation assistance	C1692317	2139	4463
Alcohol consumption counseling	C1531491	1701	3053
Drug addiction counseling	C0199403	81	106
Lifestyle Modification Assessment **SNOMED CT Concept**	**Concept Unique Identifier Code**	**Patient Counts**	**Note Counts**
Dietary history	C042501	3949	7552
Exercise history	C1287528	5912	12526
Weight finding	C1265588	5358	15749
Smoking assessment	C3853073	4719	10647
Assessment of alcohol use	C4076406	4928	10712
Assessment of drug use	C4075408	915	1544

Table 6: Lifestyle modifications, concept unique identifiers (CUIs), & extraction counts

For advice, diet and exercise advice were most documented. For assessment, exercise and weight assessments were most documented. Counts for advice on exercise were greater than counts for assessment, suggesting that providers are offering advice on exercise regardless of patients' current exercise status. Counts for drug abuse counseling and assessment were low. This was a hypertension population and drug use assessment and counseling is not a recommended lifestyle modification intervention for hypertension. Drug abuse assessment and counseling were included in the design of this NLP tool enhancement for comprehensiveness and facilitation of future training and testing in populations with higher numbers of drug abuse assessment and counseling.

Outcome

Our modified process using an augmented open-source natural language processing tool was successful in identifying lifestyle modification documentation in electronic health records of hypertension patients at an academic medical center. Given the importance of lifestyle modification interventions for multiple medical issues, this is an important and innovative step in care transparency for future comparative effectiveness studies, outcome analyses, and efforts in care improvement. To date, most studies evaluating lifestyle modification as a medical treatment rely on surveys and self-reports which are inherently vulnerable to reporting and recall bias.[45,46] With increasing emphasis placed on the need for LM to treat multiple chronic diseases, our study offers more objective and comprehensive measurement of LM care delivery via EHR analysis than previous studies. It is encouraging that 78% of patients in this study had LM addressed at least once during the study period, but there is room for improvement in how often LM is addressed with each patient. The percentage of notes containing LM was low at 5.9%, suggesting that although providers are addressing LM with many patients, providers are not repeatedly reviewing LM with patients despite the need for behavior modification for treatment of many diseases. As more efforts are made to improve lifestyle modification interventions, our study has shown that LM documentation can be automatically extracted from EHRs, thus offering increased identification of actual use of lifestyle modification in the care of hypertension and multiple chronic disorders in the future (Figure 8).[47]

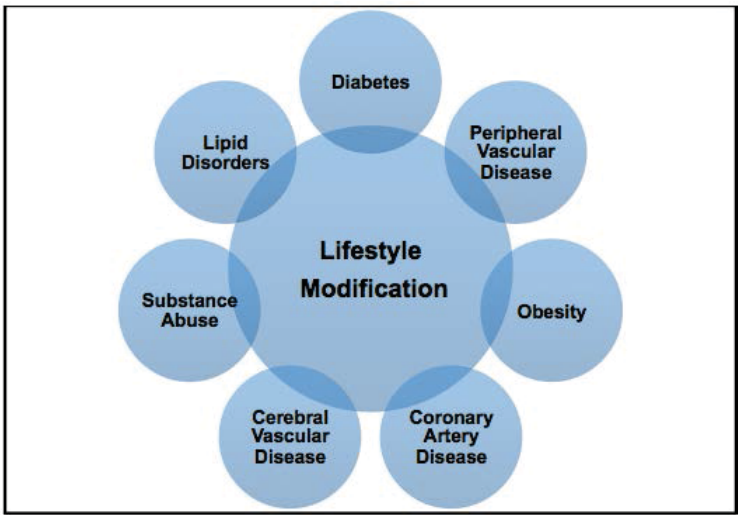


Figure 8: Future related research

Challenges and Limitations

Limitations of this study included the secondary use of EHR data which can present data quality issues (e.g., missing data, reporting bias, recording bias). Caveats and challenges in secondary use of EHR data have been well documented.[48-51] One model of data quality defined completeness as, "the extent to which data are of sufficient breadth, depth, and scope for the task at hand."[52] This study's ability to accurately retrieve and classify LM from this EHR data set confirms that this data set is sufficient for this task. A second challenge was this study's use of an open source NLP tool with requirements for customization. Initial lifestyle modification data extraction attempts using only cTAKES were unsuccessful with poor recall and precision but, after iterative development, the final testing results were successful in LM extraction and classification as assessment or advice. This study overcame a key challenge in NLP evaluation of care delivery: the inability to retrieve verbs and verb phrases. cTAKES alone could not accurately identify verbs, verb phrases, or verb-tense-specific classification of a term as advice or assessment. For example, "patient started walking" is assessment and

documentation of a patient reported LM. This is different from “recommended patient start walking 30 minutes per day” which is provider LM advice. cTAKES alone could not retrieve or distinguish these two different concepts, but using a combination of regular expressions, rules and key words, these critical concepts centered on verb phrases were accurately extracted and classified. However, inherent limitations in generalizability and scalability are present with this current approach.

Future Development

Future research will attempt to improve and augment cTAKES and its dictionary to extract and map verb phrases directly. This approach will minimize use of rules and regular expressions and make this work more generalizable and scalable. We also plan to extend lifestyle modification mapping to ICD10 and other dictionaries of interest. This work will be made available to researchers in related medical areas that use LM as a treatment and could be used to support evaluation of current and future initiatives such as “Exercise Is Medicine” and “Healthy People 2030”.[53-55] Another area for future development could be the assessment of quality of counseling for lifestyle modification with a more granular extraction of lifestyle modification concepts and counseling details to better understand best practices.[56]

CONCLUSION

This study successfully extracted lifestyle modification documentation from EHR notes and its methods and future planned expansion of these methods could be used in studies involving multiple chronic medical conditions. This is an important step in better understanding and quantifying the use of lifestyle modification as a prevention and treatment modality for many disorders. This information can be used in future outcome

and comparative effectiveness research and inform metric development for lifestyle modification documentation and counseling. The automatic identification and mapping of terms, especially verbs, related to care delivery is a major innovation that can allow further evaluation and improvement in care delivery models and treatment approaches to multiple chronic illnesses.

ACKNOWLEDGMENTS

We are grateful to the Health Innovation Program at the University of Wisconsin – Madison for assistance with data acquisition.

COMPETING INTERESTS

Each author claims no competing interests.

FUNDING

Funding for this project will be supplied after review to preserve anonymity per journal peer review policy.

REFERENCES

1. Benjamin EJ, Virani SS, Callaway CW, et al. Heart Disease and Stroke Statistics—2018 Update: A Report From the American Heart Association. *Circulation*; CIR.0000000000000558.

2. Eckel RH, Jakicic JM, Ard JD, et al. 2013 AHA/ACC Guideline on Lifestyle Management to Reduce Cardiovascular Risk: A Report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *Circulation* 2013. Epub ahead of print 12 November 2013. DOI: 10.1161/01.cir.0000437740.48606.d1.

3. Jensen MD, Ryan DH, Apovian CM, et al. *2013 AHA/ACC/TOS guideline for the management of overweight and obesity in adults*. November 2013.

4. Cerezo C, Sequra J, Praga M, et al. Guidelines updates in the treatment of obesity or metabolic syndrome and hypertension. *Current Hypertension Reports*; 15: 196–203.

5. Fleg JL, Forman DE, Berra K, et al. Secondary Prevention of Atherosclerotic Cardiovascular Disease in Older Adults: A Scientific Statement From the American Heart Association. *Circulation*; 128: 2422–2446.

6. American Diabetes Association. Standards of Medical Care in Diabetes--2014. *Diabetes Care*; 37: S14–S80.

7. Anderson JL, Halperin JL, Albert NM, et al. Management of Patients With Peripheral Artery Disease (Compilation of 2005 and 2011 ACCF/AHA Guideline Recommendations): A Report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines. *Circulation*; 127: 1425–1443.

8. Sheehan K. 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA Guideline for the Prevention, Detection, Evaluation, and Management of High Blood Pressure in Adults. 1–481.

9. Islami F, Goding Sauer A, Miller KD, et al. Proportion and number of cancer cases and deaths attributable to potentially modifiable risk factors in the United States. *CA Cancer J Clin*; 67: 7.

10. Institute for Health Metrics. Institute for Health Metrics Country Profile: United States. *www.healthdata.org*; 1–9.

11. The US Burden of Disease Collaborators, Mokdad AH, Ballestros K, et al. The State of US Health, 1990-2016. *JAMA*; 319: 1444.

12. Schoenthaler A, Luerassi L, Silver S, et al. Comparative Effectiveness of a

Practice-Based Comprehensive Lifestyle Intervention vs. Single Session Counseling in Hypertensive Blacks. *American Journal of Hypertension*; 29: 280–287.

13. Pu J, Chewning BA, Johnson HM, et al. Health Behavior Change after Blood Pressure Feedback. *PLoS ONE*; 10: e0141217.

14. Lanier JB, Bury DC, Richardson SW. Diet and Physical Activity for Cardiovascular Disease Prevention. *American Family Physician*; 93: 919–924.

15. Weintraub WS, Daniels SR, Burke LE, et al. Value of primordial and primary prevention for cardiovascular disease: a policy statement from the American Heart Association. *Circulation*; 124: 967–990.

16. Lin JS, O'Connor E, Evans CV, et al. Behavioral counseling to promote a healthy lifestyle in persons with cardiovascular risk factors: a systematic review for the U.S. Preventive Services Task Force. *Annals of Internal Medicine*; 161: 568–578.

17. Patnode CD, Evans CV, Senger CA, et al. Behavioral Counseling to Promote a Healthful Diet and Physical Activity for Cardiovascular Disease Prevention in Adults Without Known Cardiovascular Disease Risk Factors: Updated Systematic Review for the U.S. Preventive Services Task Force. 2017.

18. Lin J, Zhuo X, Bardenheier B, et al. Cost-effectiveness of the 2014 U.S. Preventive Services Task Force (USPSTF) Recommendations for Intensive Behavioral Counseling Interventions for Adults With Cardiovascular Risk Factors. *Diabetes Care*; 40: 640–646.

19. Mozaffarian D. Dietary and Policy Priorities for Cardiovascular Disease, Diabetes, and Obesity: A Comprehensive Review. *Circulation*; 133: 187–225.

20. Tajeu GS, Booth JN III, Colantonio LD, et al. Incident Cardiovascular Disease Among Adults with Blood Pressure. *Circulation*; CIRCULATIONAHA.117.027362.

21. Sacks FM, Lichtenstein AH, Wu JHY, et al. Dietary Fats and Cardiovascular Disease: A Presidential Advisory From the American Heart Association. *Circulation*; CIR.0000000000000510.

22. Turchin A, Kolatkar NS, Grant RW, et al. Using Regular Expressions to Abstract Blood Pressure and Treatment Intensification Information from the Text of Physician Notes. *Journal of the American Medical Association*; 13: 691–695.

23. Mendonca EA, Haas J, Shagina L, et al. Extracting information on pneumonia in infants using natural language processing of radiology reports. *Journal of Biomedical Informatics*; 38: 314–321.

24. Mendonca EA, Cimino JJ, Johnson S. Using Narrative Reports to Support a Digital Library. 2001, pp. 1–5.

25. Murff HJ, FitzHenry F, Matheny ME, et al. Automated identification of postoperative complications within an electronic medical record using natural language processing. *JAMA*; 306: 848–855.

26. Pakhomov S, Weston SA, Jacobsen SJ, et al. Electronic medical records for clinical research: application to the identification of heart failure. *Am J Manag Care*; 13: 281–288.

27. Salmasian H, Freedberg DE, Friedman C. Deriving comorbidities from medical records using natural language processing. *Journal of the American Medical Association*; 20: e239–42.

28. Stevens V, Bailey S, Hazlehurst B, et al. PS1-1d: Use of CER Hub to Evaluate Outcomes of Smoking Cessation Services, a Behavioral Treatment. *Clin Med Res* 2013.

29. Hazlehurst B, Frost R, Sitting D, et al. MediClass: A System for Detecting and Classifying Encounter-based Clinical Events in Any Electronic Medical Record. *Journal of the American Medical Association*; 12: 517–529.

30. Wicentowski R, Sydes MR. Using Implicit Information to Identify Smoking Status in Smoke-blind Medical Discharge Summaries. *Journal of the American Medical Association*; 15: 29–31.

31. Lindholm C, Adsit R, Bain P, et al. A demonstration project for using the electronic health record to identify and treat tobacco users. *WMJ*; 109: 335–340.

32. Khalifa A, Meystre S. Adapting existing natural language processing resources for cardiovascular risk factors identification in clinical notes. *Journal of Biomedical Informatics*; 58: S128–S132.

33. Hazlehurst BL, Lawrence JM, Donahoo WT, et al. Automating Assessment of Lifestyle Counseling inElectronic Health Records. *AMEPRE*; 46: 457–464.

34. Hosomura N, Goldberg SI, Shubina M, et al. Electronic Documentation of Lifestyle Counseling and Glycemic Control in Patients With Diabetes. *Diabetes Care*; 38: 1326–1332.

35. Johnson HM, Thorpe CT, Bartels CM, et al. Undiagnosed hypertension among young adults with regular primary care use. *Journal of Hypertension*; 32: 65–74.

36. James PA, Oparil S, Carter BL, et al. 2014 Evidence-Based Guideline for the Management of High Blood Pressure in Adults. *JAMA* 2013. Epub ahead of print 18 December 2013. DOI: 10.1001/jama.2013.284427.

37. Chobanian AV. Seventh Report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure. *Hypertension*; 42: 1206–1252.

38. Shrout T, Rudy DW, Piascik MT. Hypertension update, JNC8 and beyond. *Current Opinion in Pharmacology*; 33: 41–46.

39. Merai R, Siegel C, Rakotz M, et al. CDC Grand Rounds: A Public Health Approach to Detect and Control Hypertension. *MMWR Morbidity and Mortality Weekly Report*; 65: 1261–1264.

40. Wichmann BA, Hill ID. Algorithm AS 183: An efficient and portable pseudo-random number generator. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*

; 31: 188–190.

41. Matsumoto M, Nishimura T. Mersenne Twister: A 623-Dimensionally Equidistributed Uniform Pseudo-Random Number Generator. *ACM*; 8: 3–30.

42. Zeng QT, Tse T. Exploring and Developing Consumer Health Vocabularies. *Journal of the American Medical Association*; 13: 24–29.

43. Savova GK, Masanz JJ, Ogren PV, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Association*; 17: 507–513.

44. Milder IE, Blokstra A, de Groot J, et al. Lifestyle counseling in hypertension-related visits--analysis of video-taped general practice visits. *BMC Family Practice*; 9: 58.

45. Sreedhara M, Silfee VJ, Rosal MC, et al. Does provider advice to increase physical activity differ by activity level among US adults with cardiovascular disease risk factors? *Fam Pract* 2018. Epub ahead of print 30 January 2018. DOI: 10.1093/fampra/cmx140.

46. Jackson EA, Krishnan S, Meccone N, et al. Perceived quality of care and lifestyle counseling among patients with heart disease. *Clin Cardiol*; 33: 765–769.

47. Bruun Larsen L, Soendergaard J, Halling A, et al. A novel approach to population-based risk stratification, comprising individualized lifestyle intervention in Danish general practice to prevent chronic diseases: Results from a feasibility study. *Health Informatics J*; 23: 249–259.

48. Weiskopf NG, Hripcsak G, Swaminathan S, et al. Defining and measuring completeness of electronic health records for secondary use. *J Biomed Inform*; 46: 830–836.

49. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *Journal of the American Medical Association*; 20: 144–151.

50. Taxiarchis Botsis GHFCCW. Secondary Use of EHR: Data Quality Issues and Informatics Opportunities. *Summit on Translational Bioinformatics*; 2010: 1.

51. Hersh WR, Weiner MG, Embi PJ, et al. Caveats for the Use of Operational Electronic Health Record Data in Comparative Effectiveness Research. *Medical Care*; 51: S30–S37.

52. Wang RY, Strong DM. Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems*; 12: 5–33.

53. Cowan RE. Exercise Is Medicine Initiative: Physical Activity as a Vital Sign and Prescription in Adult Rehabilitation Practice. *Arch Phys Med Rehabil*; 97: S232–7.

54. Sperling LS, Sandesara PB, Kim JH. Exercise Is Medicine: Proof . . . and Possibilities? *JACC Cardiovasc Imaging*; 10: 1469–1471.

55. *HealthyPeople.gov*. Office of Disease Prevention and Health Promotion.

56. Stonerock GL, Blumenthal JA. Role of Counseling to Promote Adherence in Healthy Lifestyle Medicine: Strategies to Improve Exercise Adherence and Enhance Physical Activity. *Prog Cardiovasc Dis*; 59: 455–462.

4.2. Identifying Variables Associated with Lifestyle Modification Documentation for Hypertension

Coauthors
Kimberly Shoenbill; Department of Family Medicine; Program on Health and Clinical Informatics; University of North Carolina-Chapel Hill; Chapel Hill, NC, USA

Yiqiang Song; Department of Biostatistics and Medical Informatics; University of Wisconsin-Madison; Madison, WI, USA

Mark Craven; Department of Biostatistics and Medical Informatics; Department of Computer Sciences; University of Wisconsin-Madison; Madison, WI, USA

Heather Johnson; Department of Medicine, Division of Cardiovascular Medicine; University of Wisconsin-Madison; Madison, WI, USA

Maureen Smith; Department of Population Health Sciences; Department of Family Medicine; University of Wisconsin-Madison; Madison, WI, USA

Eneida A. Mendonca; Department of Biostatistics and Medical Informatics; Department of Pediatrics; University of Wisconsin-Madison; Madison, WI, USA

Keywords - MeSH Terms
Electronic Health Records, Health Behavior, Hypertension, Machine Learning

Abstract

The Gap: Just under half of the 85.7 million US adults with hypertension have uncontrolled blood pressure resulting in increased risks of death, stroke, heart failure, and myocardial infarction. Guidelines on hypertension management include lifestyle modification such as diet and exercise. In order to improve hypertension control, it is important to identify determinants of lifestyle modification assessment or advice to tailor future interventions.

Methods: Electronic health record data from 14,360 adult hypertension patients at an academic medical center were analyzed using statistical and machine learning methods to identify determinants and timing of lifestyle modification.

Results: Multiple variables were statistically significant in analysis of lifestyle modification documentation at multiple time points. Logistic regression was the best machine learning method to classify lifestyle modification documentation at any time and at ≤ 3 months with AUROC values of 0.813 and 0.685 respectively.

Conclusion: Analyzing narrative and coded data from EHRs can improve understanding of timing of lifestyle modification and patient, clinic and provider characteristics that are correlated with or predictive of documentation of lifestyle modification for hypertension.

INTRODUCTION

Rightly deemed the “silent killer”, hypertension is associated with markedly increased risks for death, stroke, coronary artery disease, and kidney failure.[1] An estimated 85.7 million adults in the United States have hypertension with 46% lacking adequate blood pressure control, even based on the prior hypertension diagnosis guidelines of ≥140/90 mmHg.[2] This inadequate control is present despite established evidence-based guidelines to achieve hypertension control.[3-7] According to these guidelines, as first-line therapy, all adult hypertensive patients should maximize lifestyle modification interventions which include: weight reduction to achieve a healthy body weight; a “Dietary Approaches to Stop Hypertension diet” which is high in fruits, vegetables and low in fat and sodium; regular aerobic exercise; moderation of alcohol consumption; and smoking cessation.[2-4]

The 2015 Global Burden of Disease Study highlighted the need for lifestyle modification when it identified the top five US health risks as: dietary risks, smoking, high blood pressure, high body mass index and physical inactivity.[8] Further evidence from this ongoing study brought to light the increasing prevalence, morbidity and mortality of hypertension worldwide.[9,10]

Multiple studies have shown lifestyle modification to be as effective as single-drug therapy in achieving hypertension control with documented systolic blood pressure reductions of 2-20 mmHg.[4] Further research on how providers can help patients adopt diet and physical activity recommendations was called for in the 2013 American Heart Association/American College of Cardiology guideline.[6] To evaluate lifestyle modification assessment or advice, several prior studies have used survey data, with

potential bias concerns, or small numbers of recorded patient encounters, with potential lack of generalizability and power. Lack of objective large-scale evidence of use of lifestyle modification is partly due to these data being buried in narrative data within clinical notes. This study works toward filling the knowledge gap of how to better treat hypertension by identifying patient, provider, and clinic characteristics that correlate with documentation of lifestyle modification in the electronic health record. To our knowledge, this work is the first to evaluate lifestyle modification for hypertension in a large population using automated methods. In a prior paper we report on our success in extraction of these data from electronic health record (EHR) notes using natural language processing.[11]

It is encouraging to note that even brief interventions for lifestyle modification can improve patient motivation, confidence, and success in implementing lifestyle changes.[12-15] Studies of visits where lifestyle modification is addressed show improved visit satisfaction for both the patient and provider, which may prompt continued discussions and success in lifestyle modification adoption.[16] Also, information from this study can help elucidate where more concentrated efforts or resources could provide the greatest increase in hypertension control through improved lifestyle modification adoption.

METHODS AND MATERIALS

Institutional and clinical settings: UW Health is the academic health system for the School of Medicine and Public Health of the University of Wisconsin-Madison caring from more than 600 thousand patients each year. UW Health adopted the Epic Systems Corporation electronic health record in the early 2000s. The data from the Epic system

is also stored in an enterprise data warehouse, which supports the use of electronic health records data for clinical operations and research. All relevant full-text documents from patients meeting inclusion criteria were extracted from the UW Health systems. Study inclusion and exclusion criteria are detailed in Table 1.[17]

Inclusion criteria: (from this retrospective patient population)
• Adult patients ≥18 years old "managed" at a UW Health practice between January 01, 2008 and December 31, 2011. "Managed" is defined as having at least two billable office encounters in an outpatient, non-urgent care primary care setting, or one primary care encounter and one office encounter in an urgent care setting (regardless of diagnosis code).

• Adult patients (regardless of gender, race, or ethnicity) meeting criteria for the diagnosis of hypertension defined as follows by guidelines from the Seventh Report of the Joint National Committee on the Prevention, Detection, Evaluation and Treatment of High Blood Pressure, which were the active guidelines during this study.[3,4] Hypertension is defined as:
  ◦ ≥ **three** separate elevated blood pressures within a two year period, measured at least 30 days apart with: systolic ≥ 140 mmHg or diastolic ≥ 90 mmHg
  ◦ Or ≥ **two** elevated blood pressures within a two year period, measured at least 30 days apart with: systolic ≥ 160 mmHg or diastolic ≥ 100 mmHg

Exclusion criteria:
• Patients pregnant during the retrospective study period were excluded 1 year before, during, and 1 year after pregnancy.
• Patients with a pre-existing diagnosis of hypertension including essential hypertension, hypertensive heart disease, hypertensive renal disease, hypertensive heart and renal disease, secondary hypertension, or any anti-hypertensive medication prescription.
• Children under the age of 18 (because we employed guidelines on lifestyle modification for adults).
• Prisoners (as a vulnerable population) were excluded as IRB approval was not obtained for inclusion of this population due to their limited freedom to choose lifestyle activities.

Table 1: Inclusion and exclusion criteria

As noted in Table 1, patients with incident (new) hypertension were the target population. These patients met criteria for hypertension, but at the start of the study did not yet have a diagnosis of hypertension. Hypertension diagnosis criteria were based on JNC 7 criteria, reflecting the guidelines available during the time of the hypertension

dataset creation.[4] Although more recent guidelines, JNC 8 and the 2017 American Heart Association's, are now available, they contain similar lifestyle modification recommendations as JNC 7.[3,18]

Data were obtained from notes throughout the outpatient clinical encounter including nursing notes, provider notes, patient instructions, nutrition consultation, and exercise consultation. The division of the data set for analysis is illustrated in Figure 1. The University of Wisconsin IRB approved this study.

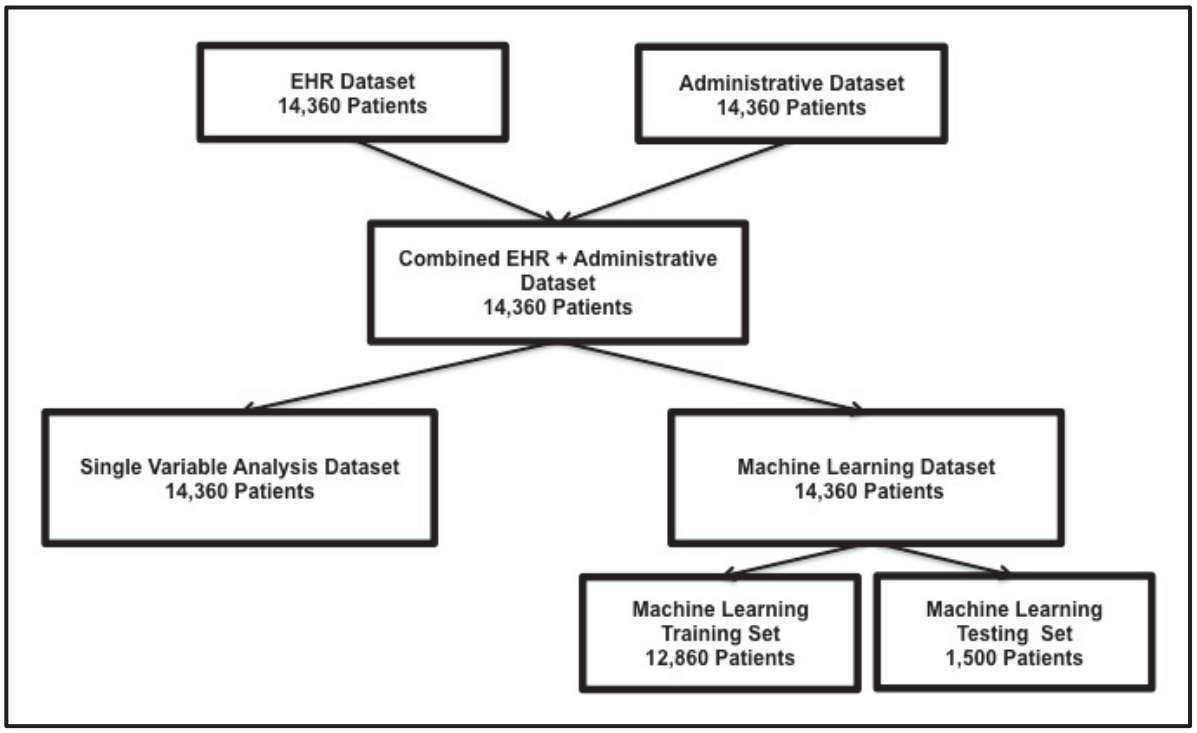


Figure 1. Division of the data set for analysis

Baseline characteristics of the study population are listed in Table 2 with the characteristics of the 1500 patients in the test subset listed beside the 12,860 patients in the machine learning training set.

Baseline Characteristics	Training Data Set: 12,860 Patients	Test Subset: 1,500 Patients

Age (years, mean, median, standard deviation)	18-97, Mean 49.18, Median 49.00, SD 14.68	18-93, Mean 49.38, Median 49.00, SD 14.63
Gender		
Female	6369 (49.5%)	733 (48.9%)
Male	6491 (50.5%)	767 (51.1%)
Race/ethnicity		
African American/Black	607 (4.7%)	72 (4.8%)
Asian	190 (1.5%)	19 (1.3%)
American Indian/Alaska Native	44 (0.3%)	6 (0.4%)
White	11372 (88.4%)	1330 (88.7%)
Hispanic/Latino	240 (1.9%)	28 (1.9%)
Other/Native Hawaiian/Pacific Islander/Multiple	66 (0.5%)	3 (0.2%)
Unknown	341 (2.7%)	42 (2.8%)

Table 2: Study subjects' baseline characteristics

Data Source: Previously coded data (e.g., diagnosis codes) and newly coded data (e.g., output from the NLP system - "father with early MI" = FH-C0455404), along with administrative data on provider and clinic characteristics, were combined for each of the 14,360 patients. These data were used to represent each patient as a vector of characteristics (e.g., comorbid conditions such as diabetes, age, gender, and family risk factors, care received from a primary care doctor, or care received at an urgent care clinic). Data were linked and de-identified for analysis (16 of the 18 HIPAA identifiers removed to create a Limited Data Set).

Class Variables: "Class variables" were used as in chi-square analysis and in building models in machine learning. The main class variables were:

1. Documentation of lifestyle modification counseling anytime (yes/no)
2. Time to documentation of lifestyle modification from time patient meets criteria for hypertension (e.g., ≤ 3 mos., >3 mos. – ≤ 6 mos., > 6 mos. – ≤ 12 mos., > 12 mos.)

Explanatory Variables as Patient Characteristics: Patient-related factors included sociodemographics (age, sex, marital status, Medicaid use during baseline or study period, primary spoken language); behavioral risk factors (baseline tobacco use and body mass index at time of meeting hypertension criteria); and comorbidities. Blood pressure measurements were extracted from structured fields within the EHR. Preexisting conditions were identified using ICD9 codes. Family history diagnoses were extracted using both ICD9 codes and natural language processing, with the natural language processing variables more comprehensive for use in the machine learning tasks.

Measures of utilization included the number of baseline primary care, specialty, and urgent care visits. Primary care visits were divided into two categories: “Primary Care” (comprised of Family Medicine, Internal Medicine, and Pediatrics visits) and a combined category, “Other” for lower prevalence specialties (Obstetrics/Gynecology, Geriatrics). Patients were assigned to the primary care provider they saw most frequently in outpatient face-to-face Evaluation & Management visits, as reported in professional service claims.[19] Provider *c*haracteristics from the administrative database were used as potential explanatory variables. Provider characteristics included provider age, provider gender, and provider specialty.

The variables used in this model were identified using an iterative approach including literature review, clinical domain knowledge, and data from the electronic health record (the top 10 most frequent diagnosis codes within this dataset). Few variables had missing data with the exception of “Provider Age” (58% values missing) and BMI (8.9% values missing). Individual variables were tested for association with the class variables

using chi square analysis. For chi square analysis of each variable, only subjects with data present for these variables were evaluated. For machine learning analysis, a separate variable value of “NA” was created to replace the missing values. Only three other variables had missing data– each less than 1%. These variables’ missing values were imputed using the mode value.

Some variables were combined using Boolean logic and if/then expressions. These combinations were performed to decrease dimensionality of the data in order to improve strength of association with a higher level diagnosis (e.g., all diabetes) compared to potentially weaker correlations when multiple lower level diagnoses (e.g. diabetes mellitus type 1 and diabetes mellitus type 2) were analyzed separately. Combined variables and their component variables are listed in Table 3.

Combined Variable Name and ICD9 or UMLS Concept Unique Identifier Codes	**Component Variable Names and ICD9 or UMLS Concept Unique Identifier Codes**
Alcohol Abuse-Dependence Combined, 303C	305-305.04 Alcohol abuse 303X Alcohol dependence syndrome
Cerebrovascular Disease Combined, 430C	430-439 Cerebrovascular disease V12.54 History of transient ischemic attack V12.5x History of disease of the circulatory system
Depression Combined, 311C	311 Depression 296.2x Major depressive disorder single episode 296.3x Major depressive disorder recurrent
Glucose Issue: Abnormal glucose without Diabetes Mellitus, 790.R	Evaluated in relation to 250x (if both diagnoses present for one patient, counted only as diabetes mellitus; if only 790.2x present, counted as 790R (revised)
Glucose Issue: Diabetes Mellitus Combined – 250C	250x Diabetes mellitus 249x Secondary diabetes mellitus
Heart disease Combined, 420C	410-414x Ischemic heart disease 420x-427x Other forms of heart disease 428x Heart failure
Long Term Use of Medications Combined, V58C	V58.66 Long term use of aspirin V58.69 Long term use of meds NEC
Menopause Combined, 256.31C	256.31-256.391 Menopause 627.8-627.91 Menopause disorder NEC 256.2 Post-ablative ovarian failure
Rheumatology Disorder Combined, 714C	714.0-714.91 Rheumatoid arthritis and other inflammatory disorders 710 Systemic lupus erythematosus

Combined Variable Name and ICD9 or UMLS Concept Unique Identifier Codes	Component Variable Names and ICD9 or UMLS Concept Unique Identifier Codes
	695.4 Lupus
Tobacco Abuse Combined, 305.1C	305.1 Tobacco use disorder V15.82 History of tobacco use
Family History of Cardiovascular Disease Combined, C0455404	C0455405 Family history of hypertension C0455407 Family history of aneurysm of artery or atherosclerosis C1261367 Family history of stroke C0559128 Family history of cardiac disorder C0475701 Family history of transient ischemic attack
Family History of Metabolic Disorder Combined, C0455367C	C0455367 Family history of metabolic disorder C0455369 Family history of raised lipids C1445950 Family history of polycystic ovaries C1313937 Family history of diabetes mellitus C2317125 Family history of impaired glucose tolerance

Table 3. Combined variables defined and their components

Individual Variable Analysis: This study evaluated each patient, clinic, and provider characteristic in relation to class variables using chi-square tests. Strength of significance of a characteristic on predicting class variables was assessed with p-values. P-values of <0.05 were considered significant. Multiple comparisons were corrected using the Bonferroni method.

Machine Learning Analysis: Machine learning is an automated approach to data analysis wherein statistical and algorithmic methods are used to build models to classify or predict an output variable using input variables and patterns learned from the data while requiring minimal explicit human instruction in model generation. This study used machine learning methods to determine if combinations of patient, clinic, and provider variables were more predictive of the class variable values than the individual variables alone. Machine learning methods have been applied to a variety of medical problems to find and model new patterns or correlations within existing data.[20-23] These models can provide insight into factors that explain the class variable and can be applied to

previously unseen patients to predict their class values. In this study, the input data consisted of patient, clinic, and provider variables and the machine learning methods attempted to identify functions of the given characteristics that were strongly predictive of: (1) receipt of lifestyle modification and (2) receipt of lifestyle modification at the specific time periods listed earlier (≤ 3 mos., >3 mos. - ≤ 6 mos., > 6 mos. – ≤ 12 mos., > 12 mos.).

Models learned by each algorithm were validated using the training data to determine the algorithm with the best classification accuracy. Once the algorithm with the most accuracy was determined, it was used in classifying the test data and its performance is reported below.

Validation using the training data employed a 10-fold cross validation method on the training set of 12,860 patients' data. This required the patient training set to be divided into 10 separate partitions (using the entire data set each time) with 90% of each version used for training and 10% used for validation. Thus, at the end of this validation, all of the 12,860 patients were sequentially used in training and validation sets. (See Figure 1 for a graphical representation of data use). Accuracy was measured by area under the receiver operating characteristic curve (AUROC) with 95% confidence intervals. Machine learning methods (classifiers) used to create the models are explained in Figure 2.

Zero R

- A learned model is a rule that assigns the majority class value to each new instance based on the training data
- The learning algorithm infers a model by predicting the majority class (mode class value) for each instance.
- The model classifies new instances as having the majority class value.
- Provides a baseline understanding of almost random prediction based solely on the most frequent class value in the dataset

Logistic regression.

- In a learned model a patient is represented as a vector of variables corresponding to a point in the hypothesis space of all clinic, patient and provider variables.
- The learning algorithm infers a model by transforming the input variables using the logistic function to the output variable. The output variable is a probability between 0 and 1. The coefficient for each variable is determined from the data to most accurately reproduce each patient's class in the training set.
- Class assignment is determined by applying a threshold value to the weighted sum of the variables for each patient. This value separates new instances into membership into one of the identified classes (e.g., LM Documentation ≤ 3 months yes or no)

Decision tree learner [68]

- A learned model is represented as an upside down tree with the "root" (at the top) containing all the data with a mixture of class variable values down to the "leaves" which contain data with only one (ideal) or a few (acceptable) class variable values
    - Looking at the tree, an observer can determine which characteristics (or combination of characteristics) is/are predictive (found in the branch path) leading out to the particular class variable value at the end of the branching (a leaf)
- A learning algorithm Iteratively splits the training set on input variable values, into subsets with purer collections of data based on the targeted class with the goal being to decrease heterogeneity in each subsequent subset.
- The model classifies new instances using the variables to split the data as per the training set to achieve a tree with pure leaves which contain instances of a single (or mostly single) class value.

Random Forests[69,70]

- A learned model is represented as a collection of decision trees from repeated resampling of the data to achieve the most homogenous leaves (classes) at the end of the branches

- The learning algorithm infers a model as in decision trees, but assigns the majority class of the many trees that were created for a given instance.
- Classification is done with assignment of the majority class value from all the tree models to a new nominal instance.

Figure 2. Machine learning method descriptions

Models were learned and evaluated using different subsets of variables including all variables (excluding redundant variables), the ten most frequent diagnosis variables from the EHR data extraction, only ICD9 variables, only NLP-extracted variables, only demographic + administrative variables, and combinations of each subset. Looking at all variables in a single dataset and using WEKA's attribute selector provided models with the highest predictive accuracy.

Evaluation of the model prediction accuracy included:

- % Correct (with mean absolute error) – the number of instances identified with the correct class value
- AUROC – Area under the receiver operating characteristic curve – the area under the curve formed by plotting the false positive rate (x axis) against the true positive rate (y axis)
- Recall (sensitivity) – the number of true positives identified by the model divided by the total number of true positives in the sample
- Precision (positive predictive value) – the number of true positives identified by the model divided by the number of all positives identified by the model
- F-measure – the harmonic mean of precision and recall:
    - $2 \times (\text{precision} \times \text{recall}) \div (\text{precision} + \text{recall})$

Comparisons of model accuracy were performed to identify whether machine learning could produce a classifier that was significantly better than chance, and whether one machine learning method significantly outperformed others. After determining the best machine learning method in predicting the correct class given a new patient record, its performance was tested on the held-aside test set of 1,500 patients' records.

Results

Counts of lifestyle modification in the dataset, divided by training and test sets, are listed in Table 4. In both the training and test sets almost 78% of patients had lifestyle modification documentation. The highest percentage of first documented lifestyle modification was within three months after the patient met criteria for hypertension. Overall, 22% of patients had no documented lifestyle modification within this study time period.

Time to First Lifestyle Modification Documentation	Training Set		Test Set	
	Yes	% Yes	Yes	% Yes
≤ 3 months of meeting hypertension definition	6996	54.40	815	54.33
> 3 - ≤ 6 months of meeting hypertension criteria	641	4.98	73	4.87
> 6 - ≤ 12 months of meeting hypertension criteria	861	6.70	104	6.93
>12 months of meeting hypertension criteria	1586	12.33	176	11.73
Documented lifestyle modification any time	10084	78.41	1168	77.87

Table 4. Counts of lifestyle modification documentation at time periods after meeting hypertension criteria

Individual Variable Analysis

Independent variables with p-values are shown in Table 5. Multiple variables were statistically significant with p-values <0.05 (bolded), especially in the time periods of ≤ 3 month and “documented any time”.

Chi Square Analysis Results Lifestyle Modification by Time from Hypertension Onset					
Variable Type (bold) & Names	**Bonferroni-Corrected P-Values By Time Period (Months) of First LM**				
	≤ 3	> 3 to ≤ 6	> 6 to ≤ 12	> 12	Any Time
Demographic Variables					
Age (< 40, ≥ 40 years old)	**< 0.01**	0.22	0.08	**< 0.01**	**< 0.01**
Gender (Female, Male)	**0.01**	0.16	0.99	0.36	**< 0.01**
Language (English, Other)	**< 0.01**	0.71	0.21	**< 0.01**	**< 0.01**
Marital Status (Married/Partner, Other)	0.46	1.00	0.76	**< 0.01**	**< 0.01**
Medicaid (Yes, No)	0.19	0.99	0.97	0.40	0.39
Race (Black, White, Other)	0.15	0.57	0.08	0.14	0.33
Clinic and Provider Variables					
Provider Age (< 40, ≥ 40 years)	**< 0.01**	**< 0.01**	**< 0.01**	**< 0.01**	**0.03**
Provider Gender (Female, Male)	**< 0.01**	0.75	0.49	**< 0.01**	**< 0.01**
Provider Specialty (Primary Care, Other)	**< 0.01**	0.75	1.00	**< 0.01**	**< 0.01**
Visits, Primary Care (< 3, ≥ 3 visits)	**0.02**	0.25	0.05	0.06	**< 0.01**
Visits, Specialty Care (< 3, ≥ 3 visits)	**< 0.01**	0.63	0.66	**0.01**	**< 0.01**
Visits, Urgent Care	**< 0.01**	**< 0.01**	0.21	**0.01**	**< 0.01**
Existing Coded and Discrete Data Variables					
Anemia 280-286	**< 0.01**	1.00	0.78	0.82	**< 0.01**
Anxiety 300.0 – 300.91	0.07	0.55	0.88	0.28	**< 0.01**
BMI (< 25 kg/m2, ≥ 25 kg/m2)**	**< 0.01**	0.09	0.70	0.10	**< 0.01**
Chronic Kidney Disease – 585x	0.86	0.59	0.89	**0.01**	**0.01**
Disease of arteries, arterioles, and capillaries – 440 – 450	0.68	0.68	0.08	0.64	0.07
Dysmetabolic Syndrome – 277.7	**0.03**	0.24	0.77	0.35	**< 0.01**
Family History of Cardiovascular Disease – V17.49	0.20	0.38	0.54	0.64	**< 0.01**
Family History of Ischemic Heart Disease – V17.3	**< 0.01**	0.67	0.79	0.68	**< 0.01**
Family History of Stroke – V17.1	0.48	0.20	0.20	0.77	**0.03**
Family History of Sudden Cardiac Death – V17.41	0.31	1.00	1.00	0.77	1.00

Hyperlipidemia – 272.4*	**< 0.01**	0.11	0.19	0.47	**< 0.01**
Hypertension NOS – 401.9*	**< 0.01**	**< 0.01**	0.05	**0.01**	**< 0.01**
Hypertensive Disease– 401 - 406	**< 0.01**	**< 0.01**	0.06	**< 0.01**	**< 0.01**
Lipoid Metabolism Disorders – 272x	**< 0.01**	0.14	0.20	0.31	**< 0.01**
Lumbago – 724.2*	**< 0.01**	**0.03**	0.34	0.18	**< 0.01**
Migraines 346.0 – 346.931	0.54	0.24	0.73	0.51	0.70
Sleep Apnea – 327.2x	**< 0.01**	0.11	0.50	0.57	**< 0.01**
Stress – 308.0 – 308.91	0.34	0.80	0.81	0.41	0.76
Obstructive Sleep Apnea – 327.3*	**< 0.01**	0.10	0.87	0.64	**< 0.01**
Osteoporosis – 733x	0.72	0.71	0.17	0.34	**< 0.01**
Overweight, obesity, other hyperalimentation – 278 – 278.03	**< 0.01**	**< 0.01**	0.95	**< 0.01**	**< 0.01**
Physical Therapy – V57.1*	**< 0.01**	0.25	0.74	0.61	**< 0.01**
Polycystic Ovaries – 256.4	**< 0.01**	0.55	0.58	0.07	**0.01**
Post-procedural State – V45.89	**0.02**	0.29	0.40	0.24	**< 0.01**
Routine Medical Exam – V70.0*	**< 0.01**	0.80	**0.02**	**0.01**	**< 0.01**
Combined Variables					
Alcohol Abuse-Dependence Combined – 303C	**< 0.01**	0.63	0.92	0.89	**< 0.01**
Cerebrovascular Disease in Patient – 430C	0.53	0.33	0.62	0.12	**< 0.01**
Depression Combined – 311C	**< 0.01**	0.47	0.21	0.42	**< 0.01**
Glucose Issue: Abnormal glucose without Diabetes Mellitus – 790.2R	**< 0.01**	0.64	0.41	0.43	**< 0.01**
Glucose Issue: Diabetes Mellitus Combined – 250C	**< 0.01**	0.10	1.00	0.54	**< 0.01**
Heart Disease Combined - 420C	0.46	0.06	0.15	**< 0.01**	**< 0.01**
Long term use of any medication - V58C*	**< 0.01**	0.33	0.84	0.19	**< 0.01**
Menopause Combined - 256.3C	**0.01**	0.50	1.00	0.24	0.23
Rheumatology Disorder Combined - 714C	0.58	0.54	0.59	**0.01**	0.57
Tobacco Abuse Combined 305.1C	**< 0.01**	0.53	0.65	0.76	**< 0.01**
Family History **NLP Extracted Data: Transformed to Coded Variables**					
Family History of Alcohol Abuse/Dependence C2911218 – NLP Extracted	0.81	0.56	0.79	0.81	0.79
Family History of Cardiovascular Disease C0455404 – NLP Extracted	0.54	0.58	0.39	1.00	**< 0.01**
Family History Metabolic Disorder - C0455367 – NLP Extracted	**0.01**	0.14	1.00	0.15	**< 0.01**

Family History of Obesity C0455373 – NLP Extracted	0.26	0.30	**< 0.01**	0.54	1.00

Table 5. Bonferroni-corrected P-values of individual variable analysis using time to first documented lifestyle modification. P-value of <0.05 is statistically significant and bolded.

Machine Learning Results

On both the training and the test sets, only two time periods had models that had predictive accuracy better than baseline prediction using the mode (most frequent class value) as the predicted value for all patients. Those two time periods with classifier results from the test set are listed in Table 6. Variables included in each model are listed in the table and were chosen using WEKA's attribute (variable) selector (CFS, Best First Search).[24,25] Note the improved AUROC on each of these models compared to the Zero R classifier (mode predictor).

Classifier	Class = Documented Lifestyle Modification Any Time				
	% Correct (MAE)	AUROC	Recall	Precision	F-Measure
Baseline Model (Weka's Zero R)	77.87 (0.34)	0.497	0.779	0.606	0.682
Random Forest	81.33 (0.25)	0.831	0.813	0.797	0.801
Variables in model: language, BMI, provider specialty, lipoid metabolism disorders, FH alcohol abuse and dependence, FH obesity, FH metabolic disorder, FH cardiovascular disease					
Classifier	**Class = ≤ 3 months**				
	% Correct (MAE)	AUROC	Recall	Precision	F-Measure
Baseline Model (Weka's Zero R)	54.33 (0.50)	0.497	0.543	0.295	0.383
Logistic Regression	65.73 (0.44)	0.685	0.657	0.657	0.657
Variables in model: BMI, urgent care visits, provider age, provider specialty, FH alcohol abuse/dependence, FH metabolic disorder, FH cardiovascular disease					

Table 6. Significant machine learning model results on test set. FH = family history, MAE = Mean Absolute Error, BMI = body mass index

Coefficients for Variables in Logistic Regression Class = Documented ≤ 3 months – AUROC = 0.685	
BMI = NA	**-0.45**
BMI = < 25	**-0.12**
Urgent_Care_Visits = ≥ 3 visits	**-0.14**
Provider_Age = NA	**-0.2**
Provider_Age = ≥ 40 years	**0.17**
Provider_Age = < 40 years	**-0.05**
PVD_Specialty = other (not primary care)	**-0.26**
Family history of alcohol abuse = Y	**0.34**

Family history of metabolic disorder = Y	**0.27**
Family history of cardiovascular disease = NM (not mentioned)	**-0.47**

Coefficients for Variables in Logistic Regression **Class = Documented Anytime – AUROC = 0.821**	
Language = other (not English)	**-0.22**
BMI = ≥ 25	**0.49**
BMI = NA (not available)	**-0.11**
BMI = < 25	**0.03**
Provider specialty = other (not primary care)	**-0.27**
Lipoid disorder = Y	**0.26**
Family history of alcohol abuse = NM (not mentioned)	**-1.43**
Family history of alcohol abuse = N	**1.01**
Family history of obesity = NM (not mentioned)	**-3.26**
Family history of metabolic disorder = Y	**0.44**
Family history of metabolic disorder = NM (not mentioned)	**-0.56**
Family history of metabolic disorder = N	**-0.04**
Family history of cardiovascular disease = Y	**0.01**
Family history of cardiovascular disease = NM (not mentioned)	**-0.94**
Family history of cardiovascular disease = N	**-0.03**

DISCUSSION

Lifestyle modification documentation was identified in 78.41% of this population. However, lifestyle modification is recommended for all patients with hypertension, so the expectation would be to have nearly 100% lifestyle modification documentation. First

documentation of lifestyle modification was most often identified in the first three months of a patient reaching criteria for hypertension. This is encouraging and consistent with guideline recommendations for lifestyle modification as first-line treatment of hypertension.

Multiple variables had statistically significant associations with lifestyle modification documentation especially for evaluation of the documentation time period any time and ≤ 3 months. The variables present in both predictive models of the two classes listed above were BMI, provider specialty, family history of alcohol abuse/dependence, family history of metabolic disorder, and family history of cardiovascular disease. These variables' direction of correlation with the classes is consistent with clinical practice. Specifically related to these models, the presence of a higher BMI, lipoid disorder, and family history of alcohol abuse, obesity, diabetes (metabolic disorder), and/or cardiovascular disease often result in providers assessing lifestyle modification and advising on lifestyle changes. Additionally, language other than the primary language of the provider can be a barrier (negatively correlated) with lifestyle modification documentation or delivery.

Other time period models were not much better than "guessing" using the baseline Zero R classifier. The best two time period results had some improvement in the % correctly classified, but the most improvement was in the AUROC showing improved accuracy of predicting the true positives with a reasonable avoidance of false positives (AUROC = 0.831 for any time with random forest and 0.685 for ≤ 3 months with logistic regression). In an effort to improve predictive capability future analysis may benefit from: (1) identifying even more variables that could add to the prediction accuracy, (2) studying

an even larger patient population, or (3) overlapping time periods to account for the possibility that there may exist no distinguishing features to predict LM documentation at < 3 months versus 3-6 months. Explanations for variability in hypertension treatment have been offered in prior studies showing that multiple factors influence whether providers address lifestyle modification. One factor is provider beliefs surrounding lifestyle modification which may include: (1) personal knowledge and comfort with lifestyle modification assessment and counseling, (2) professional and patient expectations of lifestyle modification counseling, (3) perceived effectiveness of lifestyle modification counseling or referrals for interventions, (4) perceived patient motivation for implementing lifestyle changes, and (5) constraints impeding lifestyle modification interventions (e.g., time, cost, availability of community resources, and practice resources to facilitate ongoing lifestyle changes).[26-29] Another factor may be a provider's personal health habits, as provider exercise and nonsmoking status correlated with increased lifestyle modification counseling.[30]

This study found patient demographics, comorbidities, family history, and care details were overall weak predictors of lifestyle modification documentation in machine learning models. Regarding comorbidities, this finding was not consistent with some survey-based studies identifying a specific patient comorbidity (e.g. CVD) as being predictive of more lifestyle modification interventions.[31-34] This study's inability to find a pattern may reflect biases in the data, which may be bias in the EHR data not accurately reflecting what was done at a visit, bias in the survey data due to incomplete or selective recall, or a combination of both resulting in the inconsistency across studies. Prior studies have identified lifestyle modification counseling frequency of 40% in videotaped encounters,

84-90% in survey-based analysis, and 55% in manual review of EHR data.[31,35-37] Lifestyle modification documentation was identified in 78% of patients in this study, which is within the range identified in prior studies. However, this study focused on patients with incident (non-diagnosed) hypertension at the beginning of the study. More participants in this study were younger and may have had fewer concomitant chronic illnesses than in the survey studies, which may have resulted in these comorbidities not being significant in the models. Race was also non-predictive in our models. This is likely due to this predominantly white sample population (88%) having too few participants of other races to be informative in the models. Another potential reason for poor performance of the machine learning models may be that informative variables such as patient complexity, provider practice patterns for other chronic illnesses, or clinic time constraints need to be included to better predict patterns of lifestyle modification assessment and advice.

Limitations: The main limitation of this study is its retrospective design with secondary use of EHR data. Although a prospective study would provide results from a more controlled environment, a similar-sized prospective trial investigating lifestyle modification would not be feasible due to the cost and logistics of conducting such a large clinical trial. Secondary use of data can present data quality issues (e.g., missing data, reporting bias, recording bias).[38] Despite these difficulties, multiple studies have effectively used EHR data for many clinical research purposes if data is found to be sufficient for the task.[39,40] Another limitation of this study is the use of a single health system's data. Future studies could be designed to expand and use these methods at multiple sites.

CONCLUSION

Given the importance of lifestyle modification interventions for multiple medical issues, this study is an important and innovative step in care transparency for future comparative effectiveness studies, outcome analyses, and care improvement efforts. To date, most studies evaluating lifestyle modification as a medical treatment rely on surveys and self-reports that are inherently vulnerable to reporting and recall bias. With increasing emphasis placed on the need for lifestyle modification to treat multiple chronic diseases, our study offers a more objective and comprehensive measurement of lifestyle modification care via EHR analysis. As more efforts are made to improve lifestyle modification interventions, our study has shown that lifestyle modification documentation can be automatically extracted and evaluated from EHR data, thus offering increased identification of actual use of lifestyle modification in the care of hypertension or multiple other chronic disorders. This will be even more valuable if the 2017 AHA guidelines defining hypertension as pressures of ≥130/80mmHg are widely adopted resulting in more patients meeting hypertension criteria and requiring lifestyle modification discussions and interventions.[41,42] Our methods for extracting and evaluating lifestyle modification could also be used to support evaluation of current and future initiatives such as “Exercise Is Medicine” and “Healthy People 2030”.[43,44]

ACKNOWLEDGEMENTS

We are grateful to the Health Innovation Program at the University of Wisconsin – Madison for assistance with data acquisition.

COMPETING INTERESTS

Each author claims no competing interests.

FUNDING

Funding for this project **will be supplied after review to preserve anonymity per journal peer review policy.**

REFERENCES

1. World Health Organization. Hypertension. WHO. http://www.who.int/cardiovascular_diseases/publications/global_brief_hypertension/en/. Published 2013. Accessed January 20, 2018.

2. Benjamin EJ, Virani SS, Callaway CW, et al. Heart Disease and Stroke Statistics—2018 Update: A Report From the American Heart Association. *Circulation*. January 2018:CIR.0000000000000558. doi:10.1161/CIR.0000000000000558.

3. James PA, Oparil S, Carter BL, et al. 2014 Evidence-Based Guideline for the Management of High Blood Pressure in Adults. *JAMA*. December 2013. doi:10.1001/jama.2013.284427.

4. Chobanian AV. Seventh Report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure. *Hypertension*. 2003;42(6):1206–1252. doi:10.1161/01.HYP.0000107251.49515.c2.

5. Go AS, Bauman MA, Coleman King SM, et al. An effective approach to high blood pressure control: a science advisory from the American Heart Association, the American College of Cardiology, and the Centers for Disease Control and Prevention. *Hypertension*. 2014;63(4):878–885. doi:10.1161/HYP.0000000000000003.

6. Eckel RH, Jakicic JM, Ard JD, et al. 2013 AHA/ACC Guideline on Lifestyle Management to Reduce Cardiovascular Risk: A Report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *Circulation*. November 2013. doi:10.1161/01.cir.0000437740.48606.d1.

7. Weber MA, Schiffrin EL, White WB, et al. Clinical Practice Guidelines for the Management of Hypertension in the Community. *The Journal of Clinical Hypertension*. December 2013:n/a–n/a. doi:10.1111/jch.12237.

8. Reddy KS. Global Burden of Disease Study 2015 provides GPS for global health 2030. *Lancet*. 2016;388(10053):1448–1449. doi:10.1016/S0140-6736(16)31743-3.

9. Forouzanfar MH, Liu P, Roth GA, et al. Global Burden of Hypertension and Systolic Blood Pressure of at Least 110 to 115 mm Hg, 1990-2015. *JAMA*. 2017;317(2):165. doi:10.1001/jama.2016.19043.

10. Global Burden of Cardiovascular Diseases Collaboration, Roth GA, Johnson CO, et al. The Burden of Cardiovascular Diseases Among US States, 1990-2016. *JAMA Cardiol*. April 2018. doi:10.1001/jamacardio.2018.0385.

11. Shoenbill K, Song Y, Gress L, Johnson H, Smith M, Mendonca E. Natural

Language Processing of Lifestyle Modification Documentation. *Health Informatics Journal*. (In press.).

12. Frates EP, Bonnet J. Collaboration and Negotiation. *American Journal of Lifestyle Medicine*. 2016;10(5):302–312. doi:10.1001/jama.1984.03350200032016.

13. Kreuter MW, Chheda SG, Bull FC. How does physician advice influence patient behavior? Evidence for a priming effect. *Arch Fam Med*. 2000;9(5):426–433.

14. Halm J, Amoako E. Physical activity recommendation for hypertension management does healthcare provider advice make a difference - 2008. *Ethnicity & Disease*. 2008;18(3):278–282.

15. Viera AJ, Kshirsagar AV, Hinderliter AL. Lifestyle modifications to lower or control high blood pressure: is advice associated with action? The behavioral risk factor surveillance survey. *Journal of Clinical Hypertension (Greenwich, Conn)*. 2008;10(2):105–111.

16. Bell RA, Kravitz RL. Physician counseling for hypertension: What do doctors really do? *Patient Educ Couns*. 2008;72(1):115–121. doi:10.1016/j.pec.2008.01.021.

17. Johnson HM, Thorpe CT, Bartels CM, et al. Undiagnosed hypertension among young adults with regular primary care use. *Journal of Hypertension*. 2014;32(1):65–74. doi:10.1097/HJH.0000000000000008.

18. Whelton PK, Carey RM, Aronow WS, et al. 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA Guideline for the Prevention, Detection, Evaluation, and Management of High Blood Pressure in Adults: Executive Summary: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *Journal of the American College of Cardiology*. 2018;71(19):2199–2269. doi:10.1016/j.jacc.2017.11.005.

19. Thorpe CT, Flood GE, Kraft SA, Everett CM, Smith MA. Effect of patient selection method on provider group performance estimates. *Medical Care*. 2011;49(8):780–785. doi:10.1097/MLR.0b013e31821b3604.

20. Roque FS, Jensen PB, Schmock H, et al. Using Electronic Patient Records to Discover Disease Correlations and Stratify Patient Cohorts. Ritchie MD, ed. *PLoS Comput Biol*. 2011;7(8):e1002141. doi:10.1371/journal.pcbi.1002141.s004.

21. Jensen MD, Ryan DH, Apovian CM, et al. *2013 AHA/ACC/TOS Guideline for the Management of Overweight and Obesity in Adults*. 2013:n/a–n/a. doi:10.1002/oby.20660.

22. Kawaler E, Cobian A, Peissig P, Cross D, Yale S, Craven M. Learning to predict post-hospitalization VTE risk from EHR data. *AMIA Annu Symp Proc*.

2012;2012:436–445.

23. Kohane IS. Using electronic health records to drive discovery in disease genomics. *Nature reviews Genetics*. 2011;12(6):417–428. doi:10.1038/nrg2999.

24. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten I. *The WEKA Data Mining Sofware: an Update*. Vol 11. SIGKDD Explorations; 2009:10–18.

25. Witten IH, Frank ES. *Data Mining: Practical Machine Learning Tools and Techniques*. 3rd ed. San Francisco: Elsevier; 2011.

26. Ampt AJ, Amoroso C, Harris MF, McKenzie SH, Rose VK, Taggart JR. Attitudes, norms and controls influencing lifestyle risk factor management in general practice. *BMC Family Practice*. 2009;10(1):247. doi:10.1016/j.ypmed.2004.07.015.

27. Faria C, Wenzel M, Lee KW, Coderre K, Nichols J, Belletti DA. A narrative review of clinical inertia: focus on hypertension. *J Am Soc Hypertens*. 2009;3(4):267–276. doi:10.1016/j.jash.2009.03.001.

28. Svetkey LP, Pollak KI, Yancy WS, et al. Hypertension Improvement Project: Randomized Trial of Quality Improvement for Physicians and Lifestyle Modification for Patients. *Hypertension*. 2009;54(6):1226–1233. doi:10.1161/HYPERTENSIONAHA.109.134874.

29. Jerdén L, Dalton J, Johansson H, Sorensen J, Jenkins P, Weinehall L. Lifestyle counseling in primary care in the United States and Sweden: a comparison of patients' expectations and experiences. *Global Health Action*. 2018;11(1):1438238. doi:10.2147/JMDH.S14900.

30. Hung OY, Keenan NL, Fang J. Physicians' health habits are associated with lifestyle counseling for hypertensive patients. *American Journal of Hypertension*. 2013;26(2):201–208. doi:10.1093/ajh/hps022.

31. Viera AJ, Kshirsagar AV. Lifestyle modification advice for lowering or controlling high blood pressure: who's getting it? *… of Clinical Hypertension*. 2007.

32. Sinclair J, Lawson B, Burge F. Which patients receive advice on diet and exercise? Do certain characteristics affect whether they receive such advice? *Can Fam Physician*. 2008;54(3):404–412.

33. Sreedhara M, Silfee VJ, Rosal MC, Waring ME, Lemon SC. Does provider advice to increase physical activity differ by activity level among US adults with cardiovascular disease risk factors? *Fam Pract*. January 2018. doi:10.1093/fampra/cmx140.

34. Honda K. Factors underlying variation in receipt of physician advice on diet and exercise: applications of the behavioral model of health care utilization. *Am J*

Health Promot. 2004;18(5):370–377. doi:10.4278/0890-1171-18.5.370.

35. Milder IE, Blokstra A, de Groot J, van Dulmen S, Bemelmans WJ. Lifestyle counseling in hypertension-related visits--analysis of video-taped general practice visits. *BMC Family Practice*. 2008;9(1):58. doi:10.1186/1471-2296-9-58.

36. Lopez L, Cook EF, Horng MS, Hicks LS. Lifestyle Modification Counseling for Hypertensive Patients: Results From the National Health and Nutrition Examination Survey 1999-2004. *American Journal of Hypertension*. 2009;22(3):325–331. doi:10.1001/jama.279.11.839.

37. Johnson HM, Olson AG, LaMantia JN, et al. Documented Lifestyle Education Among Young Adults with Incident Hypertension. *J GEN INTERN MED*. 2014;30(5):556–564. doi:10.1007/s11606-014-3059-7.

38. Hersh WR, Weiner MG, Embi PJ, et al. Caveats for the Use of Operational Electronic Health Record Data in Comparative Effectiveness Research. *Medical Care*. 2013;51:S30–S37. doi:10.1097/MLR.0b013e31829b1dbd.

39. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *Journal of the American Medical Association*. 2013;20(1):144–151. doi:10.1136/amiajnl-2011-000681.

40. Weiskopf NG, Hripcsak G, Swaminathan S, Weng C. Defining and measuring completeness of electronic health records for secondary use. *J Biomed Inform*. 2013;46(5):830–836. doi:10.1016/j.jbi.2013.06.010.

41. Wilt TJ, Kansagara D, Qaseem A, for the Clinical Guidelines Committee of the American College of Physicians. Hypertension Limbo: Balancing Benefits, Harms, and Patient Preferences Before We Lower the Bar on Blood Pressure. *Annals of Internal Medicine*. 2018;168(5):369. doi:10.7326/M17-3293.

42. Gu A, Yue Y, Kim J, Argulian E. The Burden of Modifiable Risk Factors in Newly Defined Categories of Blood Pressure. *Am J Med*. 2018;131(11):1349–1358.e5. doi:10.1016/j.amjmed.2018.06.030.

43. Cowan RE. Exercise Is Medicine Initiative: Physical Activity as a Vital Sign and Prescription in Adult Rehabilitation Practice. *Arch Phys Med Rehabil*. 2016;97(9 Suppl):S232–S237. doi:10.1016/j.apmr.2016.01.040.

44. *HealthyPeople.Gov*. Office of Disease Prevention and Health Promotion https://www.healthypeople.gov/2020/About-Healthy-People/Development-Healthy-People-2030/Framework. Accessed October 28, 2017.

4.3 Hypertension Medication Initiation: Statistical and Machine Learning Analyses

Coauthors
Kimberly Shoenbill; Department of Family Medicine; Program on Health and Clinical Informatics; University of North Carolina-Chapel Hill; Chapel Hill, NC, USA

Yiqiang Song; Department of Biostatistics and Medical Informatics; University of Wisconsin-Madison; Madison, WI, USA

Mark Craven; Department of Biostatistics and Medical Informatics; Department of Computer Sciences; University of Wisconsin-Madison; Madison, WI, USA

Heather Johnson; Department of Medicine, Division of Cardiovascular Medicine; University of Wisconsin-Madison; Madison, WI, USA

Maureen Smith; Department of Population Health Sciences; Department of Family Medicine; University of Wisconsin-Madison; Madison, WI, USA

Eneida A. Mendonca; Department of Biostatistics and Medical Informatics; Department of Pediatrics; University of Wisconsin-Madison; Madison, WI, USA

Keywords - MeSH Terms
Electronic Health Records, Health Behavior, Hypertension, Machine Learning

To be submitted for publication

Abstract

The Gap: Over 45% of the 85.7 million US adults with hypertension have uncontrolled blood pressure resulting in increased risks of cardiovascular disease including stroke, heart failure, and myocardial infarction. Guidelines on hypertension management include lifestyle modification (e.g., diet, exercise) and medication initiation as first line treatment. To understand current hypertension treatment efforts and improve hypertension control, it is important to determine the frequency and inter-relatedness of lifestyle modification and hypertension medication initiation.

Methods: Electronic health record data from 14,360 adult hypertension patients at an academic medical center were analyzed using statistical and machine learning methods to determine the relationships between documentation of lifestyle modification and hypertension medication initiation and predictors of medication initiation.

Results: Within one year of hypertension onset, over 78% of patients had documented lifestyle modification. Overall, only 38% of patients with ongoing elevated blood pressures had documentation of an initial antihypertensive prescription within one year. However, 69% of patients without an antihypertensive prescription had documented lifestyle modification. The best machine learning classifier for medication initiation within one year = false was random forest with an AUROC of 0.709 on the test set with recall (sensitivity) = 82%.

Conclusion: Analyzing textual data and coded data from EHRs can provide a more complete assessment of hypertension care. Knowledge gained from this approach to EHR data evaluation can inform and improve hypertension treatment, care process, and

metric development efforts to decrease morbidity and mortality related to uncontrolled hypertension.

INTRODUCTION

Hypertension is the number one modifiable risk for cardiovascular disease.[1] Unfortunately only 54.4% of hypertension patients have controlled blood pressure.[2] Causes of inadequate control of blood pressure can manifest throughout the care continuum including lack of timely hypertension diagnosis, lack of timely hypertension treatment with lifestyle modification and/or medication, lack of patient adherence to treatment, and lack of treatment titration.[3-8] Barriers to each of these steps have been explored previously in many studies, usually based on surveys of patients or providers. Although those studies continue to inform understanding and optimization of hypertension care, this study undertook a novel approach of using computational methods to evaluate the poorly understood problem of timely medication initiation. Electronic health record data from 14,360 patients that met criteria for hypertension were analyzed using statistical and machine learning methods. Machine learning methods are appropriate when the goal is to identify informative patient phenotypes with high predictive accuracy, rather than explain causal effects. It can be used to determine if combinations of variables are more predictive of a class variable than individual variables alone. Machine learning methods have been applied to a variety of medical problems to find previously unrecognized relationships within existing data.[9,10] To our knowledge, this study is the first to apply machine learning methods to the problem of timely hypertension medication initiation and its potential predictors.[11,12]

METHODS AND MATERIALS

Institutional and clinical settings: UW Health is the academic health system for the University of Wisconsin-Madison. UW Health adopted Epic Systems Corporation's electronic health record in 2004 and has created a Health Information Management Center, which is devoted to the integrity of system-wide electronic health record data and facilitation of its use for improved patient care and health.

Data Source: All relevant full-text documents from patients meeting inclusion criteria were retrieved from the UW Health electronic health record. Data were obtained from notes throughout the outpatient clinical encounter including nursing notes, provider notes, patient instructions, nutrition consultation, and exercise consultation. Data were linked and deidentified for analysis with 16 of the 18 HIPAA identifiers removed to create a Limited Data Set. These data in this set included all person-level information used to construct the sample: sociodemographics, insurance information, and encounter details (e.g., diagnosis and procedure codes) coupled with administrative data including provider specialty, provider age, provider gender, and visit type (primary, specialty, or urgent care). The University of Wisconsin IRB approved this study.

Data Retrieval: Study inclusion and exclusion criteria are detailed in Table 1.[13]

Inclusion criteria: (from this retrospective patient population)
• Adult patients ≥18 years old "managed" at a UW Health practice between January 01, 2008 and December 31, 2011. "Managed" is defined as having at least two billable office encounters in an outpatient, non-urgent care primary care setting, or one primary care encounter and one office encounter in an urgent care setting (regardless of diagnosis code).

• Adult patients (regardless of gender, race, or ethnicity) meeting criteria for the diagnosis of hypertension defined as follows by guidelines from the Seventh Report of the Joint National Committee on the Prevention, Detection, Evaluation and Treatment of High Blood Pressure which were the active guidelines during this study.[14-17] Hypertension is defined as:
  • ≥ **three** separate elevated blood pressures within a two year period, measured at least 30 days apart with: systolic ≥ 140 mmHg or diastolic ≥ 90 mmHg
  • Or ≥ **two** elevated blood pressures within a two year period, measured at least 30 days apart with: systolic ≥ 160 mmHg or diastolic ≥ 100 mmHg

Exclusion criteria:
• Patients pregnant during the retrospective study period were excluded 1 year before, during, and 1 year after pregnancy.
• Patients with a pre-existing diagnosis of hypertension including essential hypertension, hypertensive heart disease, hypertensive renal disease, hypertensive heart and renal disease, secondary hypertension, or any anti-hypertensive medication prescription.
• Children under the age of 18 (because we employed guidelines on lifestyle modification for adults).
• Prisoners (as a vulnerable population) were excluded as IRB approval was not obtained for inclusion of this population due to their limited freedom to choose lifestyle activities.

Table 1: Inclusion and exclusion criteria

The hypertension diagnosis criteria were based on JNC 7 criteria, reflecting the guidelines available during the time of the hypertension dataset creation.[17] Lifestyle modification was retrieved from the electronic health records using natural language processing methods with details reported in another manuscript.[15] Blood pressure measurements were retrieved from structured fields within the EHR. Preexisting conditions were identified using ICD9 codes.

Data Division:

14,860 patients met inclusion criteria with 500 patients' records used for NLP tool development for lifestyle modification retrieval referenced above. Of the remaining 14,360 patients, this study identified 9701 patients with persistent hypertension by one year of meeting and sustaining criteria for hypertension. These patients either had

medication initiated or were still eligible for medication treatment of hypertension by one year after meeting criteria.

Individual variable analysis was completed on the dataset of 9,701 patients. Machine learning methods were applied with 8,655 patients' records used for training the models and 1,046 used for testing of the models.

A schematic of the dataset and data division is provided in Figure 1.

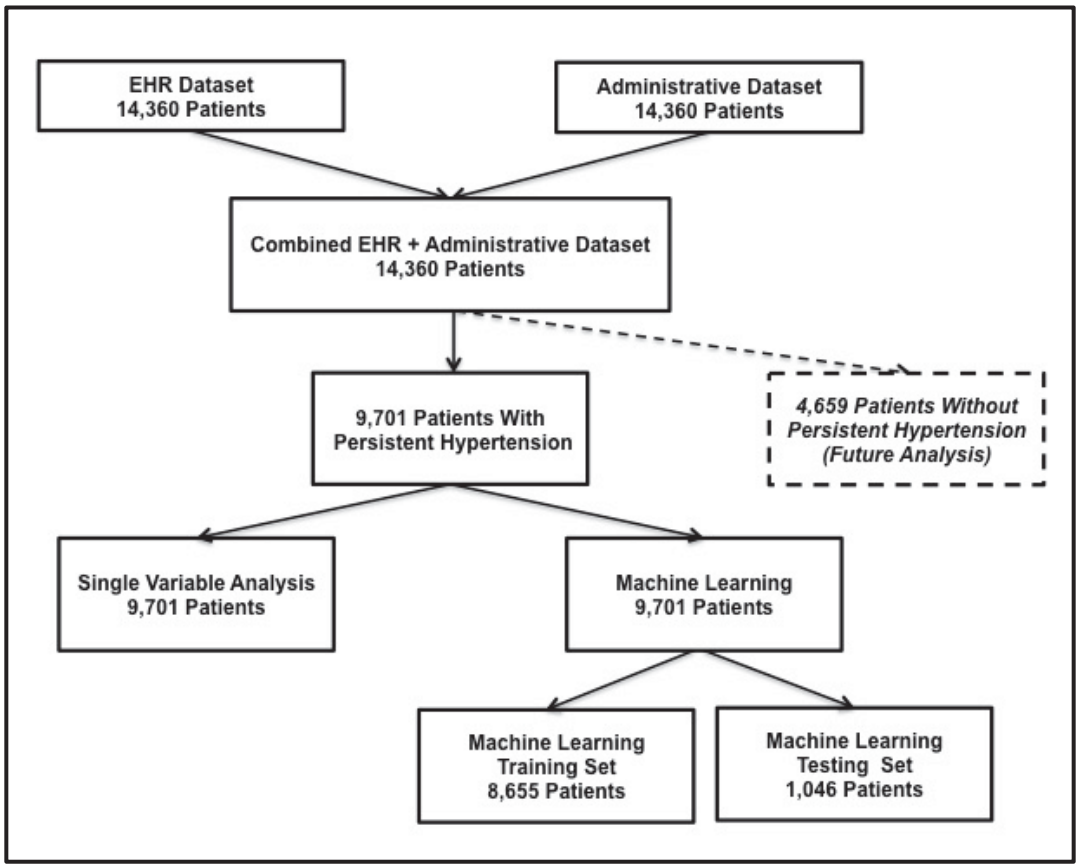


Figure 1. Data division

Data Analysis: This study identified and evaluated medication initiation within one year of hypertension onset and lifestyle modification documentation by looking at frequencies of each. This study also identified and evaluated independent variables consisting of

patient, provider, clinic, and lifestyle modification characteristics as potential predictors of delays in medication initiation for hypertension. Although comorbidities and cardiovascular risk guide how long to continue lifestyle modification as sole treatment of persistent hypertension, this study used the longest proposed period: one year.[16,18-20] Variables were selected for retrieval based on an iterative approach using literature, clinical domain knowledge, and data from the electronic health record (the top 10 most frequent diagnosis codes within this dataset).

All variables are listed in Table 4 and include eight types:

1. Discrete demographic variables
2. Provider and clinic variables from the administrative dataset
3. Existing coded variables of diagnosis codes or findings (e.g., ICD9 for any anemia: 280-286) or findings (e.g., BMI = 30 kg/m^2)
4. Combined variables of multiple ICD9 codes (Table 2)
5. Family history variables – NLP-extracted family history details from EHR text
    a) Two variables from NLP extraction were mapped directly to the Unified Medical Language System (UMLS) Concept Unique Identifier Code (CUI) codes: Family History of Alcohol Abuse or Dependence (C29111218), and Family History of Obesity (C0455373)
    b) Two variables were combinations of NLP-extracted variables and mapped to UMLS CUI codes: Family History of Cardiovascular Disease (C0455404C), and Family History of Metabolic Disorder (C0455367C) (Table 2)
6. Lifestyle modification assessment – NLP-extracted data transformed to UMLS CUI codes

7. Lifestyle modification advice – NLP-extracted data transformed to UMLS CUI codes

8. Time periods of first lifestyle modification documentation

Variables were combined using Boolean logic and if/then expressions. Combinations were performed to decrease dimensionality of the data in order to improve strength of association with a higher level diagnosis (e.g., all diabetes) compared to potentially no identified correlations with the class with multiple lower level diagnoses (e.g. diabetes mellitus type 1 and diabetes mellitus type 2 analyzed separately).

Combined Variable Name and ICD9 or UMLS Concept Unique Identifier Codes	**Component Variable Names and ICD9 or UMLS Concept Unique Identifier Codes**
Depression Combined, 311C	311 Depression 296.2x Major depressive disorder single episode 296.3x Major depressive disorder recurrent
Alcohol Dependence Combined, 303C	305-305.04 Alcohol abuse 303X Alcohol dependence syndrome
Tobacco Abuse Combined, 305.1C	305.1 Tobacco use disorder V15.82 History of tobacco use
Menopause Combined, 256.31C	256.31-256.391 Menopause 627.8-627.91 Menopause disorder NEC 256.2 Post-ablative ovarian failure
Inflammatory Disease Combined, 714C	714.0-714.91 Rheumatoid arthritis and other inflammatory disorders 710 Systemic lupus erythematosus 695.4 Lupus
Cerebrovascular Disease Combined, 430C	430-439 Cerebrovascular disease V12.54 Transient ischemic attack
Diabetes Mellitus Combined, 250C	250x Diabetes mellitus 249x Secondary diabetes mellitus
Long Term Use of Medications Combined, V58C	V58.66 Long term use of aspirin V58.69 Long term use of meds NEC
Family History of Metabolic Disorder Combined, C0455367C	C0455367 Family history of metabolic disorder C0455369 Family history of raised lipids C1445950 Family history of polycystic ovaries C1313937 Family history of diabetes mellitus C2317125 Family history of impaired glucose tolerance
Family History of Cardiovascular Disease Combined, C0455404	C0455405 Family history of hypertension C0455407 Family history of aneurysm of artery or atherosclerosis C1261367 Family history of stroke C0559128 Family history of cardiac disorder C0475701 Family history of transient ischemic

	attack

Table 2. Combined variables defined with component concept unique identifier codes and descriptors

The dependent variable (also called the class variable in machine learning terminology) was medication initiation by one year (yes, no). Delay in medication initiation was defined as no hypertension medication documentation in patients with persistent hypertension within one year of meeting criteria for hypertension (Table 1). Persistent hypertension was defined as systolic blood ≥140mmHg or diastolic blood pressure ≥90mmHg based on guidelines at the time of this data creation. The data were evaluated using individual variable analysis and machine learning methods.

Individual Variable Analysis: Chi-square tests were used to evaluate each patient, provider, clinic, and lifestyle modification characteristic in relation to the dependent variable (medication initiation by one year yes, no) for the 9,701 patients with persistent hypertension at one year. A p-value <0.05 was considered statistically significant. Multiple comparisons were corrected for using the Bonferroni method.

Machine Learning Analysis: WEKA, a suite of open-source machine learning tools, was used to perform the machine learning analyses.[14,21] Data from 8,655 patients were used for model training and data from 1,046 patients were used for model testing. Baseline characteristics of subjects in the training and test sets are provided (Table 3).

Baseline Characteristics	**Training Subset: 8,655 Patients Number of patients (% of subset)**	**Test Subset: 1,046 Patients Number of patients (% of subset)**
Age (years, mean, median, standard deviation)	18-96, Mean 49.05, Median 49.00, SD 14.41	18-93, Mean 49.32, Median 49.00, SD 14.08
Gender		

Female	3939 (45.6%)	465 (44.5%)
Male	4716 (54.4%)	581 (55.5%)
Race/ethnicity		
African American/Black	401 (4.6%)	50 (4.8%)
Asian	130 (1.5%)	13 (1.2%)
American Indian/Alaska Native	26 (0.3%)	5 (0.5%)
White	7644 (88.3%)	931 (89.0%)
Hispanic/Latino	162 (1.9%)	14 (1.3%)
Other/Native Hawaiian/Pacific Islander/Multiple	42 (0.5%)	1 (0.1%)
Unknown	250 (2.9%)	32 (3.1%)

Table 3: Study subjects' baseline characteristics

In this study, the input data consisted of patient, provider, clinic, and lifestyle modification variables for each patient with persistent hypertension at one year, coupled with the known class variable for that patient: hypertension medication initiation was provided within one year of meeting hypertension criteria. The machine learning methods identified models of the given variables that were predictive of receipt of medication initiation within one year for patients with persistent hypertension. Such models can provide insight into the factors that explain the class variable, and can be applied to previously unseen patients to predict their class variables. Machine learning algorithms used were logistic regression, decision trees, and random forest. They are briefly described in Figure 2 along with zero R as a baseline classifier used for comparison. Models learned using each of these algorithms were internally validated on the training set using a 10-fold cross validation method. This means that training data were separated into 10 partitions (using the entire data set each time) with 90% of each version of the data set used for training and 10% used for testing. Thus, at the end of

this validation, all of the 8,655 patients' data in the training set were sequentially used as a training and a validation set.

Zero R

- A learned model is a rule that assigns the majority class value to each new instance based on the training data
- The learning algorithm infers a model by predicting the majority class (mode class value) for each instance.
- The model classifies new instances as having the majority class value.
- Provides a baseline understanding of almost random prediction based solely on the most frequent class value in the dataset

Logistic regression.

- In a learned model a patient is represented as a vector of variables corresponding to a point in the hypothesis space of all clinic, patient and provider variables.
- The learning algorithm infers a model by transforming the input variables using the logistic function to the output variable. The output variable is a probability between 0 and 1. The coefficient for each variable is determined from the data to most accurately reproduce each patient's class in the training set.
- Class assignment is determined by applying a threshold value to the weighted sum of the variables for each patient. This value separates new instances into membership into one of the identified classes: 1) medication initiation yes; 2) medication initiation no.

Decision tree learner

- A learned model is represented as an upside down tree with the “root” (at the top) containing all the data with a mixture of class variable values down to the “leaves” which contain data with only one (ideal) or a few (acceptable) class variable values
    - Looking at the tree, an observer can determine which characteristics (or combination of characteristics) is/are predictive (found in the branch path) leading out to the particular class variable value at the end of the branching (a leaf)
- A learning algorithm Iteratively splits the training set on input variable values, into subsets with purer collections of data based on the targeted class with the goal being to decrease heterogeneity in each subsequent subset.
- The model classifies new instances using the variables to split the data as per the training set to achieve a tree with pure leaves which contain instances of a single (or mostly single) class value.

Random Forests

- A learned model is represented as a collection of decision trees from repeated resampling of the data to achieve the most homogenous leaves (classes) at the end of the branches
- The learning algorithm infers a model as in decision trees, but assigns the majority class of the many trees that were created for a given instance.
- Classification is done with assignment of the majority class value from all the tree models to a new nominal instance.

Figure 2. Classifier descriptions in machine learning analysis.

Data was evaluated in multiple ways including all variables (excluding redundant variables), the ten most frequent diagnosis variables from the EHR data retrieval, only ICD9 variables, only NLP-retrieved variables, only demographic and administrative variables, and combinations of each subset. Looking at all variables in a single dataset and using WEKA’s variable selector provided models with the highest predictive capacity.

Evaluation of the model prediction accuracy included:

- %Correct (with mean absolute error) – the number of instances identified with the correct class
- AUROC – Area under the receiver operating characteristic curve – the area under the curve formed by plotting the false positive rate (x axis) against the true positive rate (y axis)
- Recall (sensitivity) – the number of true positives identified by the model divided by the total number of true positives in the sample
- Precision (positive predictive value) – the number of true positives identified by the model divided by the number of all positives identified by the model
- F-measure – the harmonic mean of precision and recall:
    - 2 × (precision × recall) / (precision + recall)

Accuracy was measured by area under the receiver operating characteristic curve (AUROC) with 95% confidence intervals. Comparisons of model accuracy were performed to identify whether machine learning could produce a classifier that is significantly better than chance, and whether one machine learning classifier significantly outperformed others. After determining the best model in predicting the correct class given a new patient record, its performance was tested on the held-aside test set of 1,046 patients' records.

RESULTS

Individual Variable Analysis Results

From the 14,360 patients in the dataset, 9,701 met criteria for persistent hypertension by one year (with blood pressure ≥140/90mmHg or hypertension medication initiation);

analysis of the patients not requiring medication by one year will be described in a separate manuscript.

This study's individual variable analysis results identified most variables as statistically significant with Bonferroni corrected P-Value <0.05 (Table 4).

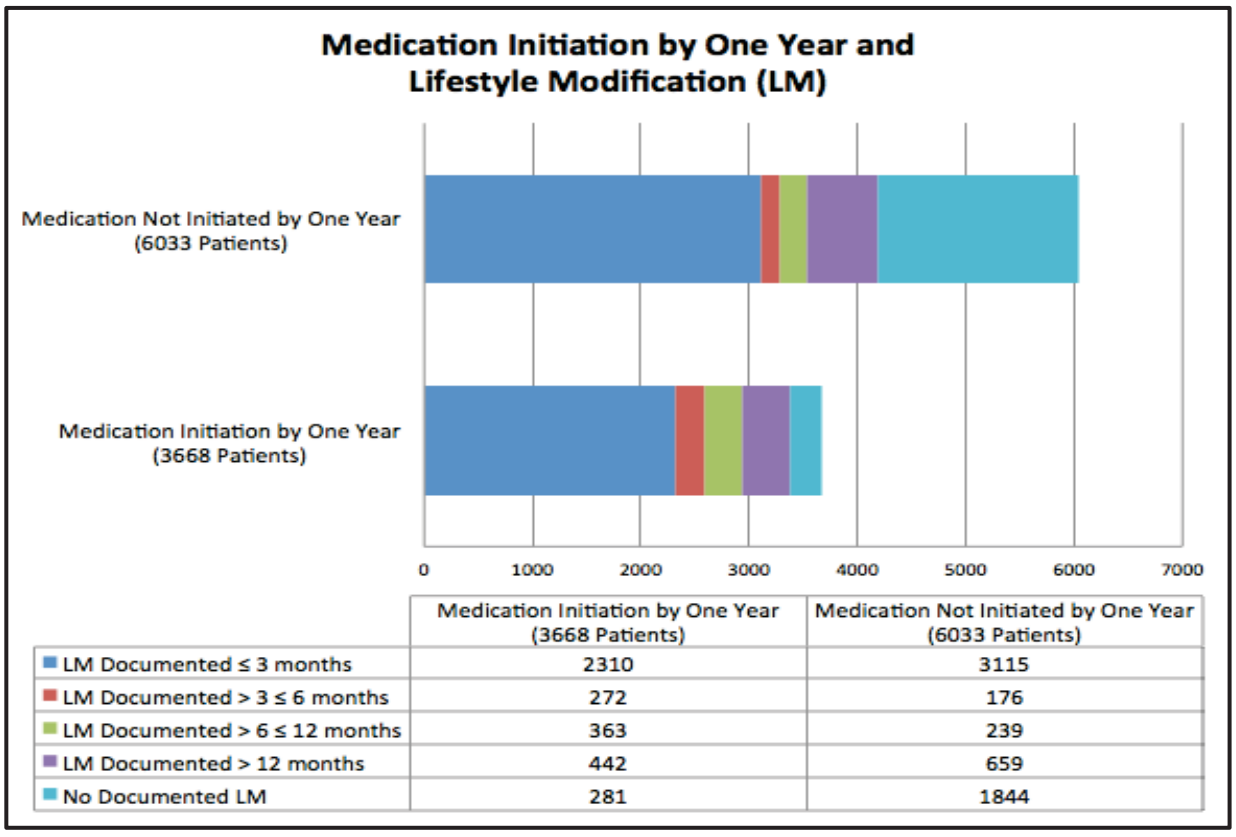
Chi Square Analysis Results Hypertension Medication By One Year True/False	
Variable Type (in bold) and Variable Names	**Adjusted P-Value**
Demographic Variables	
Age (< 40, ≥ 40 years old)	<0.01
Gender (Female, Male)	<0.01
Language (English, Other)	<0.01
Marital Status (Married/Partner, Other)	
Medicaid (Yes, No)	<0.01
Race (Black, White)	<0.01
Race (Black, Other)	0.21
Race (Other, White)	0.04
Administrative Data: Clinic and Provider Variables	
Provider Age (< 40, ≥ 40 years old)	0.03
Provider Gender (Female, Male)	<0.01
Provider Specialty (Primary Care, Other)	0.06
Visits, Primary Care (< 3, ≥ 3 visits)	0.66
Visits, Specialty Care (< 3, ≥ 3 visits)	0.03
Visits, Urgent Care (< 3, ≥ 3 visits)	0.49
Existing Coded Data Variables	
Anemia 280-286	<0.01
Anxiety 300.0 – 300.91	0.02
BMI (< 25 kg/m^2, ≥ 25kg/m^2)	0.05
Chronic Kidney Disease – 585x	<0.01
Diabetes Mellitus – 250*	Combined
Disease of arteries, arterioles, and capillaries – 440 - 450	<0.01
Dysmetabolic Syndrome – 277.7	1.00
Family History of Cardiovascular Disease – V17.49	<0.01
Family History of Ischemic Heart Disease – V17.3	<0.01
Family History of Stroke – V17.1	0.02
Family History of Sudden Cardiac Death – V17.41	0.71
Hyperlipidemia – 272.4*	<0.01
Hypertension NOS – 401.9*	<0.01
Hypertensive Disease– 401 - 406	<0.01
Lipoid Metabolism Disorders – 272x	<0.01
Long-term Use Aspirin NEC – V58.66*	Combined
Long-term Use Meds NEC – V58.69*	Combined
Lumbago – 724.2*	<0.01

Migraines 346.0 – 346.931	<0.01
Sleep Apnea – 327.2x	<0.01
Stress – 308.0 – 308.91	0.13
Obstructive Sleep Apnea – 327.3*	<0.01
Osteoporosis – 733x	<0.01
Overweight, obesity, other hyperalimentation – 278 – 278.03	<0.01
Physical Therapy NEC – V57.1*	<0.01
Polycystic Ovaries – 256.4	0.52
Post-procedural State – V54.89*	<0.01
Routine Medical Exam – V70.0*	<0.01
Combined Variables	
Abnormal glucose without Diabetes Mellitus – 790.2R (revised – patients with only abnormal glucose diagnosis (not diabetes mellitus) counted in 790.2R)	<0.01
Alcohol Abuse-Dependence Combined – 303C (combined variable)	0.01
Cardiovascular Disease in Patient – 430C (combined variable)	<0.01
Depression Combined – 311C (combined variable)	<0.01
Diabetes Mellitus Combined – 250C (combined variable)	<0.01
Heart Disease Combined– 420C (combined variable)	<0.01
Long term use of any medication – V58C* (combined variable)	<0.01
Menopause Combined – 256.3C (combined variable)	0.40
Rheumatology Disorder Combined – 714C (combined variable)	0.13
Tobacco Abuse Combined – 305.1C (combined variable)	<0.01
Family History **NLP-Retrieved Data: Transformed to Coded Variables**	
Family History of Alcohol Abuse/Dependence – C2911218	0.11
Family History of Cardiovascular Disease – C0455404C (combined variable)	0.01
Family History of Metabolic Disorder – C0455367C (combined variable)	0.73
Family History of Obesity – C0455373	0.09
Lifestyle Modification Advice **NLP-Retrieved Data: Transformed to Coded Variables**	
Alcohol consumption counseling – C1531491	<0.01
Dietary mgmt. education, guidance, counseling – C1828150	<0.01
Drug addiction counseling – C0199403	0.02
Exercise education – C0582396	<0.01
Patient advised about weight management – C3697318	<0.01
Smoking cessation assistance – C1692317	<0.01
Lifestyle Modification Assessment **NLP-Retrieved Data: Transformed to Coded Variables**	
Assessment of alcohol use – C4076406	<0.01
Assessment of drug use – C4075408	<0.01
Dietary history – C042501	<0.01
Exercise history – C1287528	<0.01
Smoking assessment – C3853073	<0.01
Weight finding – C1265588	<0.01
Time Periods of First Lifestyle Modification Documentation **NLP-Retrieved Data**	
Lifestyle modification ≤ 3 mos	<0.01
Lifestyle modification > 3 mos - ≤ 6mos	<0.01

Lifestyle modification > 6 mos - ≤ 12mos	<0.01
Lifestyle modification > 12 mos	0.10
Lifestyle modification documented at any time within the study period	<0.01

Table 4. Individual variable analysis results. Bonferroni adjusted p-values. *Top 10 most frequent diagnoses in dataset.

Of the 9,701 patients with ongoing elevated blood pressures over one year, 3,668 (38%) received a prescription for hypertension medication. Of the patients needing hypertension medication treatment, 6033 (62%) did not have a documented prescription for hypertension medication by one year of hypertension diagnosis criteria being met and sustained. Lifestyle modification was documented for 92% of patients prescribed hypertension medication, compared to 69% lifestyle documentation among patients not prescribed hypertension medication (Figure 3).



	Medication Initiation by One Year (3668 Patients)	Medication Not Initiated by One Year (6033 Patients)
LM Documented ≤ 3 months	2310	3115
LM Documented > 3 ≤ 6 months	272	176
LM Documented > 6 ≤ 12 months	363	239
LM Documented > 12 months	442	659
No Documented LM	281	1844

Figure 3. Medication Initiation by 1 year relative to timing of first lifestyle modification documentation

In both groups, the most common time period for initial lifestyle modification documentation was in the first three months of hypertension onset with 63% of the medication group and 52% of the non-medication group having lifestyle modification documentation by three months. 31% of patients without medication prescribed did not have documented lifestyle modification. 8% of patients with medication initiated by one year did not have documented lifestyle modification. In this study population with 9,701 patients having persistent hypertension, 1,844 (19%) had no documented intervention of lifestyle modification or hypertension medication initiation by one year.

Machine Learning Analysis Results

Results from evaluation of different machine learning classifiers on the training set are summarized in Table 5. The variables producing these results were selected as the most informative of the entire variable list using WEKA's variable selector (CFS Subset with a Best First Search method) using the training data. The variables selected were then used on the testing set.

Hypertension Medication Initiation Within One Year of Hypertension Onset					
Classifier	**Analysis Results on Internal Cross Validation on the Training Set (8655 Patients' Data) Results for Class = Medication Initiation 1 year = False**				
	% Correct (MAE)	AUROC	Recall (Sensitivity)	Precision (Positive Predictive Value)	F-Measure
Baseline Accuracy (Weka's Zero R)	62.32 (0.47)	0.500	1.000	0.623	0.768
Logistic Regression	67.97 (0.40)	0.725	0.825	0.709	0.763
Decision Tree	68.13 (0.40)	0.693	0.837	0.706	0.766
Random Forest	68.34 (0.40)	0.725	0.862	0.700	0.772
Machine Learning Classifier	**Analysis Results on Test Set (1046 Patients' Data) Results for Class = Medication Initiation 1 year = False**				
	% Correct (MAE)	AUROC	Recall	Precision	F-Measure
Zero R	61.09 (0.48)	0.497	1.000	0.611	0.758
Random Forest	68.26 (0.41)	0.709	0.817	0.708	0.759

Table 5. Machine learning results on the training and test sets

Variables incorporated into these models consisted of diagnoses, NLP-retrieved types of lifestyle modification, NLP-retrieved timing of lifestyle modification, and NLP-retrieved family history (Figure 4).

Lifestyle Modification Type	First Lifestyle Modification Time	Diagnoses	Family history

- Dietary mgmt. education, guidance, counseling: C1828150 - Exercise education: C0582396 - Patient advised about weight management: C3697318 - Dietary history: C042501 - Exercise history: C1287528 - Smoking assessment: C3853073 - Weight finding: C1265588	- > 3 to ≤ 6 mos. - > 6 to ≤ 12 mos.	- Hypertensive disease: 401-406 - Long-term use of any medication: V58C	- Family History of Cardiovascular Disease: C0455404C

Figure 4. Variables used in machine learning models

Coefficients for Training Set Logistic Regression Model **Medication Initiation At 1 Year = False**
Hypertensive disease = N * 0.37
Dietary mgmt. education, guidance, counseling = N * 0.36
Dietary history = N * 0.07
Exercise history = N * 0.19
Weight finding = N * 0.1
Smoking assessment = N * 0.13
>3months and <=6months = T * -0.47
>6months and <=12months = T * -0.46
Family history of cardiovascular disease = Not Mentioned * 0.28
Family history of cardiovascular disease = N * 0.1
Long term use of any medication = N * 0.11

Figure 5. Coefficients of variables in logistic regression model

Comparison of the different classifiers based on AUROC with a 95% confidence interval showed random forest to be the marginally best classifier equal to logistic regression in the AUROC but slightly better in % correctly classified, recall (sensitivity), and precisions (PPV) as shown in Figure 6.

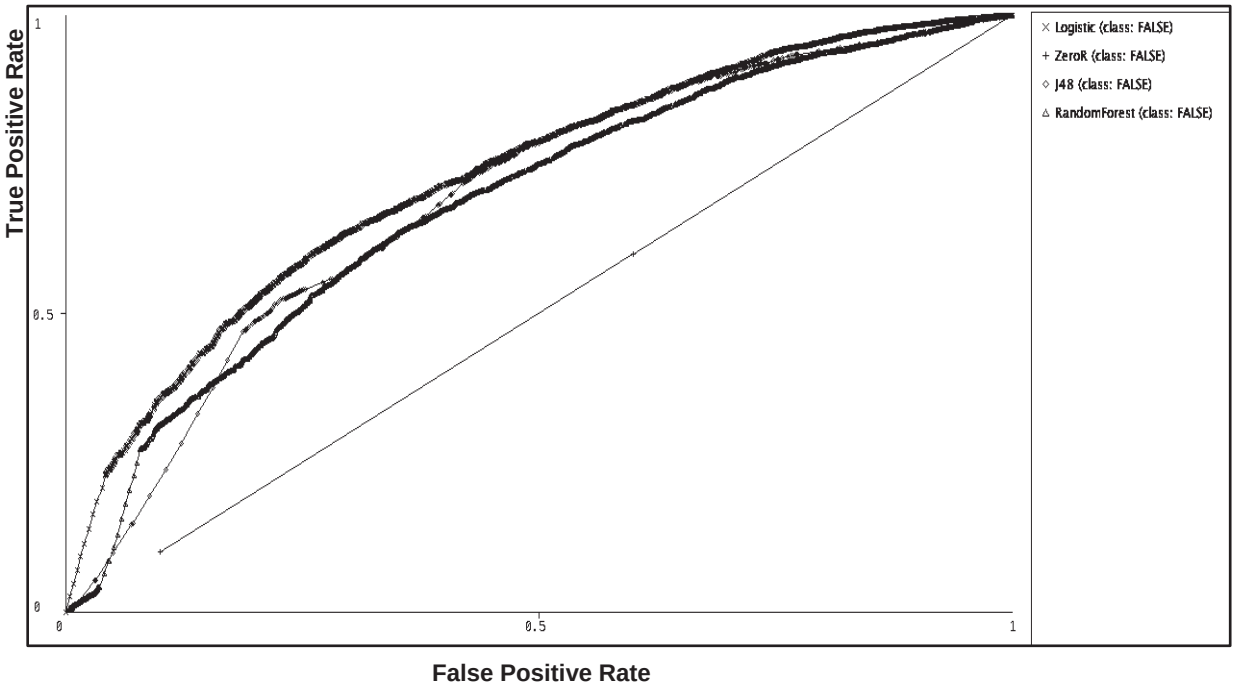


Figure 6. AUROC comparison with baseline Zero R classifier (diagonal straight line)

The baseline classifier used was Weka's Zero R. The machine learning classifiers used were logistic regression, J48 decision tree, and random forest. These were very similar in their AUROCs as shown in the graph (Figure 5). All were significantly better than chance as represented by the baseline classifier Zero R. Random forest and logistic regression had the same AUROC and slightly higher than the J48 decision tree as shown in Table 5.

DISCUSSION

Our goal in this study was to understand the relationship of patient, provider, clinic, and lifestyle modification characteristics as related to hypertension medication initiation within one year of hypertension onset. We believed that knowledge of these relationships could allow providers to identify potential barriers or facilitators of medication initiation due to identified patient, provider, and clinic characteristics. We also hoped that with this knowledge providers could better tailor plans for hypertension treatment to optimize current hypertension guidelines on use of lifestyle modification and hypertension medication initiation.

Over 78% of all patients with persistent hypertension by one year of onset had lifestyle modification documentation as appropriate first line treatment of hypertension. This included 69% of patients who were not started on hypertension medications by one year. These patients may appear to be receiving no treatment of their hypertension if evaluation is comprised only of structured/coded data. Using natural language processing to analyze the records, it becomes clear that many patients have documented lifestyle modification, which is first line treatment of hypertension. However, only 38% of patients needing medication by one year after hypertension onset were prescribed hypertension medication. This is not consistent with current guidelines for hypertension treatment.[22,23] Patients who have documented hypertension medication by one year are also more likely to have documented lifestyle modification (92%) than patients with no hypertension medication documentation (69%). This study provides clear evidence that efforts are needed to improve treatment of hypertension and/or its documentation. Focusing on the 19% of patients having no documentation of either lifestyle modification or medication may be a prime starting point.

Although our findings are informative of hypertension treatment patterns, we would like to continue research to find even stronger predictors of medication initiation for hypertension treatment within one year. One caveat to this analysis is that if the goal is to decrease the number of patients **not** receiving medication within one year as needed, the analysis must look at the ability of a classifier to identify the patients who did **not** have medication initiation (i.e., the “medication initiation by 1 year = false” group) as reported here. The random forest AUROC on the test set was 0.709 and the recall (sensitivity) was almost 82%. This is evidence that we can accurately identify determinants of medication initiation within a population with hypertension. The variables that were used in these models are listed above with their coefficients in the logistic regression model. Variables that increase the likelihood that a patient will not get hypertension medication initiated in a timely manner include absence of a diagnosis of hypertensive disease, absence of a documented family history of cardiovascular disease, no long-term medication use, and lack of lifestyle modifications. This machine learning analysis provided evidence for a more succinct number of variables being significant in determining medication initiation compared to individual variable analysis. These results specifically identify the importance of timely medication initiation and providers diagnosing hypertension and using lifestyle modification for all hypertension patients (regardless of family history of cardiovascular disease).

It is interesting that administrative data used in these analyses were not informative in these machine learning models and most ICD9 diagnosis codes were also not informative. In analyses using only ICD9 codes, the best AUROC was only 0.606 (with only 65.8% correctly identified instances). The most consistently informative ICD9 code

was the combined 401-406 code for hypertensive disease. Of the 2277 patients with one or more of these hypertension ICD9 codes, 56% had medication initiation within one year. 35% of patients having documentation of medication initiation within one year had a diagnosis of hypertension, compared to only 16% of patients without medication initiation having a hypertension diagnosis.

In predicting medication initiation, there are several potential reasons that the classifiers did not perform better overall. Many reasons have been identified in the literature and/or offered by our research group including: patient preferences or biases for or against medication; provider preferences for or against treatment (which may manifest as clinical inertia); patient financial or time constraints; time constraints during office visits; patient complexity or acuity; and lack of consistent follow-up with a single provider.[24-27] Although our study attempted to evaluate some of these factors (e.g. looking at comorbidities, family history, provider specialty), many components potentially causing variability within the clinical encounter were not captured (e.g., complexity of an encounter when blood pressure was elevated; consistency of the same provider seeing the patient; and provider prescribing patterns with his/her patient panel). Variables such as these may be the hidden keys to unlocking the pattern of who and who does not receive medication initiation within one year.

Limitations: One limitation of this study is its retrospective design with secondary use of EHR data. Secondary use of data can present data quality issues (e.g., missing data, reporting bias, recording bias).[28-30] Despite these difficulties, multiple studies have successfully used EHR data for many purposes and this dataset has been successfully used in the prior study retrieving lifestyle modification by this research team. Another

limitation of this study is the possibility for overfitting given that data from only one health system were evaluated and may reflect specific practice and documentation patterns unique to that system. Finally, each patient encounter is an amalgam of stated and unstated, congruent and incongruent: preferences, biases, goals, agendas, evidence, data, and constraints involving all stakeholders including the patient, provider, clinic, payers, and professional bodies issuing guidelines. This reality of potential variability in care delivery limits the accuracy of our classification models.

CONCLUSION

This study has presented new evidence that most patients (69%) who were not on hypertension medication, and appeared to have no treatment for their ongoing hypertension by one year of onset, actually had documented lifestyle modification. Certainly, by one year, this is not sufficient if blood pressure is still elevated. However, it is evidence that providers are at least beginning acknowledgement of and/or interventions for hypertension. With only 38% of eligible (persistently hypertensive) patients receiving medication by one year, there is much work to be done in improving timely initiation of antihypertensive medication initiation. Even more patients will be needing lifestyle modification interventions if current recommendations for diagnosing hypertension at lower thresholds are widely accepted.[31,32]

To extend this effort, future analysis of this data with comparisons of patients who do and do not require medication by one year of hypertension onset will be undertaken by our team. Additionally, more research into “hidden” variables, and extension of this work to include ICD10 diagnosis codes, may prove beneficial in uncovering even stronger predictors of medication initiation.

Future work may also be undertaken to apply these methods to the analysis of treatment of other disorders that use lifestyle modification as first-line treatment such as diabetes, obesity, hyperlipidemia, and peripheral vascular disease.

ACKNOWLEDGEMENTS

We are grateful to the Health Innovation Program at the University of Wisconsin – Madison for assistance with data acquisition.

COMPETING INTERESTS

Each author claims no competing interests.

FUNDING

Funding for this project **will be supplied after review to preserve anonymity per journal peer review policy.**

REFERENCES

1. Valderrama AL, Tong X, Ayala C, Keenan NL. Prevalence of Self-Reported Hypertension, Advice Received From Health Care Professionals, and Actions Taken to Reduce Blood Pressure Among US Adults-HealthStyles, 2008. *The Journal of Clinical Hypertension*. 2010;12(10):784–792. doi:10.1111/j.1751-7176.2010.00323.x.

2. Benjamin EJ, Virani SS, Callaway CW, et al. Heart Disease and Stroke Statistics—2018 Update: A Report From the American Heart Association. *Circulation*. January 2018:CIR.0000000000000558. doi:10.1161/CIR.0000000000000558.

3. Gil-Guillén V, Orozco-Beltrán D, Pérez RP, et al. Clinical inertia in diagnosis and treatment of hypertension in primary care: quantification and associated factors. *Blood Press*. 2010;19(1):3–10. doi:10.3109/08037050903350762.

4. Huebschmann AG, Mizrahi T, Soenksen A, Beaty BL, Denberg TD. Reducing clinical inertia in hypertension treatment: a pragmatic randomized controlled trial. *Journal of Clinical Hypertension (Greenwich, Conn)*. 2012;14(5):322–329. doi:10.1111/j.1751-7176.2012.00607.x.

5. Phillips LS, Branch WT, Cook CB, et al. Clinical inertia. *Annals of Internal Medicine*. 2001;135(9):825–834.

6. Donahue KE, Vu MB, Halladay JR, et al. Patient and Practice Perspectives on Strategies for Controlling Blood Pressure, North Carolina, 2010–2012. *Prev Chronic Dis*. 2014;11:130157. doi:10.5888/pcd11.130157.

7. Faria C, Wenzel M, Lee KW, Coderre K, Nichols J, Belletti DA. A narrative review of clinical inertia: focus on hypertension. *J Am Soc Hypertens*. 2009;3(4):267–276. doi:10.1016/j.jash.2009.03.001.

8. Johnson HM, Warner RC, Bartels CM, LaMantia JN. “They’re younger… it’s harder.” Primary providers’ perspectives on hypertension management in young adults: a multicenter qualitative study. *BMC Research Notes*. 2017;10(1):94. doi:10.1111/j.1742-1241.2009.02290.x.

9. Mendonca EA, Haas J, Shagina L, Larson E, Friedman C. Extracting information on pneumonia in infants using natural language processing of radiology reports. *Journal of Biomedical Informatics*. 2005;38(4):314–321. doi:10.1016/j.jbi.2005.02.003.

10. Kawaler E, Cobian A, Peissig P, Cross D, Yale S, Craven M. Learning to predict post-hospitalization VTE risk from EHR data. *AMIA Annu Symp Proc*. 2012;2012:436–445.

11. Krittanawong C, Bomback AS, Baber U, Bangalore S, Messerli FH, Wilson Tang

WH. Future Direction for Using Artificial Intelligence to Predict and Manage Hypertension. *Current Hypertension Reports*. 2018;20(9):591. doi:10.1161/CIRCULATIONAHA.118.034390.

12. Ye C, Fu T, Hao S, et al. Prediction of Incident Hypertension Within the Next Year: Prospective Study Using Statewide Electronic Health Records and Machine Learning. *J Med Internet Res*. 2018;20(1):e22. doi:10.1016/j.ijmedinf.2015.06.007.

13. Johnson HM, Thorpe CT, Bartels CM, et al. Undiagnosed hypertension among young adults with regular primary care use. *Journal of Hypertension*. 2014;32(1):65–74. doi:10.1097/HJH.0000000000000008.

14. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten I. *The WEKA Data Mining Sofware: an Update*. Vol 11. SIGKDD Explorations; 2009:10–18.

15. Shoenbill K, Song Y, Gress L, Johnson H, Smith M, Mendonca E. Natural Language Processing of Lifestyle Modification Documentation. *Health Informatics Journal*. (In press.).

16. James PA, Oparil S, Carter BL, et al. 2014 Evidence-Based Guideline for the Management of High Blood Pressure in Adults. *JAMA*. December 2013. doi:10.1001/jama.2013.284427.

17. Chobanian AV. Seventh Report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure. *Hypertension*. 2003;42(6):1206–1252. doi:10.1161/01.HYP.0000107251.49515.c2.

18. Go AS, Bauman MA, Coleman King SM, et al. An effective approach to high blood pressure control: a science advisory from the American Heart Association, the American College of Cardiology, and the Centers for Disease Control and Prevention. *Hypertension*. 2014;63(4):878–885. doi:10.1161/HYP.0000000000000003.

19. Weber MA, Schiffrin EL, White WB, et al. Clinical Practice Guidelines for the Management of Hypertension in the Community. *The Journal of Clinical Hypertension*. December 2013:n/a–n/a. doi:10.1111/jch.12237.

20. Moser M. Are lifestyle interventions in the management of hypertension effective? How long should you wait before starting specific medical therapy? An ongoing debate. *Journal of Clinical Hypertension (Greenwich, Conn)*. 2005;7(6):324–326.

21. Witten IH, Frank ES. *Data Mining: Practical Machine Learning Tools and Techniques*. 3rd ed. San Francisco: Elsevier; 2011.

22. Shrout T, Rudy DW, Piascik MT. Hypertension update, JNC8 and beyond. *Current Opinion in Pharmacology*. 2017;33:41–46. doi:10.1016/j.coph.2017.03.004.

23. Whelton PK, Carey RM, Aronow WS, et al. 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA Guideline for the Prevention, Detection, Evaluation, and Management of High Blood Pressure in Adults: Executive Summary. *Hypertension*. November 2017:HYP.0000000000000066. doi:10.1161/HYP.0000000000000066.

24. Borzecki AM, Oliveria SA, Berlowitz DR. Barriers to hypertension control. *Am Heart J*. 2005;149(5):785–794. doi:10.1016/j.ahj.2005.01.047.

25. Khatib R, Schwalm J-D, Yusuf S, et al. Patient and Healthcare Provider Barriers to Hypertension Awareness, Treatment and Follow Up: A Systematic Review and Meta-Analysis of Qualitative and Quantitative Studies. Barengo NC, ed. *PLoS ONE*. 2014;9(1):e84238. doi:10.1371/journal.pone.0084238.s002.

26. Gee ME, Bienek A, Campbell NRC, et al. Prevalence of, and Barriers to, Preventive Lifestyle Behaviors in Hypertension (from a National Survey of Canadians With Hypertension). *The American Journal of Cardiology*. 2012;109(4):570–575. doi:10.1016/j.amjcard.2011.09.051.

27. Sessoms J, Reid K, Williams I, Hinton I. Adherence to National Guidelines for Managing Hypertension in African Americans. *International Journal of Hypertension*. October 2015:1–7. doi:10.1155/2015/498074.

28. Hersh WR, Weiner MG, Embi PJ, et al. Caveats for the Use of Operational Electronic Health Record Data in Comparative Effectiveness Research. *Medical Care*. 2013;51:S30–S37. doi:10.1097/MLR.0b013e31829b1dbd.

29. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *Journal of the American Medical Association*. 2013;20(1):144–151. doi:10.1136/amiajnl-2011-000681.

30. Weiskopf NG, Hripcsak G, Swaminathan S, Weng C. Defining and measuring completeness of electronic health records for secondary use. *J Biomed Inform*. 2013;46(5):830–836. doi:10.1016/j.jbi.2013.06.010.

31. Muntner P, Carey RM, Gidding S, et al. Potential U.S. Population Impact of the 2017 American College of Cardiology/American Heart Association High Blood Pressure Guideline. *Circulation*. November 2017:CIRCULATIONAHA.117.032582. doi:10.1161/CIRCULATIONAHA.117.032582.

32. The SPRINT Research Group. A Randomized Trial of Intensive versus Standard Blood-Pressure Control. *N Engl J Med*. 2015;373(22):2103–2116. doi:10.1056/NEJMoa1511939.

5. Conclusion

5.1. Discussion

This work was undertaken to accomplish the three specific aims detailed above. Each aim was accomplished, although with varying degrees of success. I have shown that lifestyle modification as treatment of hypertension can be extracted from electronic health records to provide a more complete picture of hypertension care compared to analysis of only discrete field or coded data from electronic health records. This information can inform future studies and care improvement efforts for hypertension. This study provided new, objective evidence that:

1. Lifestyle modification documentation can be extracted from electronic health record narrative data with a high degree of recall and precision using augmented natural language processing methods.

2. Over 78% of hypertension patients had documented lifestyle modification in this study population with most receiving it within the first 3 months of hypertension onset.

3. Lifestyle modification documentation was difficult to predict in this population except for the classes “≤ 3 months” and “any time”. Both of these time periods also had more statistically significant variables than the other three time periods. The most informative variables were BMI, provider specialty, family history of alcohol abuse/dependence, family history of metabolic disorder, and family history of cardiovascular disease. Specific results and challenges regarding lifestyle modification predictor analysis are addressed in the manuscript (chapter 4, section 4.2).

3. Only 38% of patients needing hypertension medication (within one year of reaching criteria for hypertension) had documented prescriptions for it.

4. Most patients (69%) who did not have a documented prescription for hypertension medication, did have documented lifestyle modification. This is new, encouraging evidence that hypertension in these patients is being addressed, although not adequately with the lack of medication initiation present by one year after reaching and persisting in hypertension blood pressure levels.

5. Predictors of medication initiation within one year of hypertension onset were discovered that provided recall of 79.3% and precision of 71%.

5.2. Limitations

The main limitation of this study is its retrospective design with secondary use of EHR data. Although a prospective study would provide results from a more controlled environment, a similar-sized prospective trial investigating lifestyle modification would not be feasible due to such a large subject pool. Secondary use of data can present data quality issues (e.g., missing data, reporting bias, recording bias).[82-84] Despite these difficulties, multiple studies have used EHR data for many purposes, including phenotype classification, cohort identification, disease correlation, adverse drug event identification, and outcome prediction.[64-66,76-81,85]

The issue of completeness and EHR data quality for use in research has been addressed in prior papers.[82] One model of data quality defined completeness as, “the extent to which data are of sufficient breadth, depth, and scope for the task at hand.”[85] This view has been further discussed and completeness described as “contextual and is determined through an understanding of specific data needs.”[86] This study’s data had relatively few missing values and allowed the intended aims to be accomplished.

Specific handling of missing data and shortcomings for each aim are described in the manuscripts (sections 4.1-4.3).

With secondary use of EHR data there are concerns whether documentation of lifestyle modification reflects what actually occurs during a patient encounter. Evaluating the data using methods employed in this study allows for understanding of at least documentation, if not actual care, and can spur efforts to improve both documentation and care. Several studies have noted that improved documentation can improve patient care.[87,88]

5.3. Future Work

Future work, extending this thesis work, will evaluate predictors of needing hypertension medication at one year after hypertension onset compared to predictors of not needing medication. To increase generalizability, I would like to apply these methods to evaluation of all patients with hypertension (not just incident hypertension patients) and to hypertension patients in another health care system. To improve scalability, I would like to work with developers of cTAKES or other open-source NLP tools to incorporate these methods and this lexicon into an NLP tool to allow automatic identification of verbs and terms of interest in lifestyle modification evaluation.

5.4. Conclusion

Methods used and results from this study have added new findings to the scientific literature as discussed above. These methods could be used for further hypertension evaluation, adapted to future hypertension guidelines, and adapted to evaluation of other disorders that also use lifestyle modification as a core treatment modality.

Information from studies using these methods could be used to improve care delivery and metric development regarding use of lifestyle modification.

6. References

1. Benjamin EJ, Virani SS, Callaway CW, et al. Heart Disease and Stroke Statistics—2018 Update: A Report From the American Heart Association. *Circulation*. January 2018:CIR.0000000000000558. doi:10.1161/CIR.0000000000000558.

2. Zhou B, Bentham J, Di Cesare M, et al. Worldwide trends in blood pressure from 1975 to 2015: a pooled analysis of 1479 population-based measurement studies with 19·1 million participants. *The Lancet*. 2017;389(10064):37–55. doi:10.1016/S0140-6736(16)31919-5.

3. The US Burden of Disease Collaborators, Mokdad AH, Ballestros K, et al. The State of US Health, 1990-2016. *JAMA*. 2018;319(14):1444. doi:10.1001/jama.2018.0158.

4. Centers for Disease Control and Prevention, ed. *CDC High Blood Pressure*. US Department of Health and Human Services https://www.cdc.gov/bloodpressure/index.htm. Accessed November 15, 2018.

5. Chobanian AV. Seventh Report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure. *Hypertension*. 2003;42(6):1206–1252. doi:10.1161/01.HYP.0000107251.49515.c2.

6. James PA, Oparil S, Carter BL, et al. 2014 Evidence-Based Guideline for the Management of High Blood Pressure in Adults. *JAMA*. December 2013. doi:10.1001/jama.2013.284427.

7. Williams B, Mancia G, Spiering W, et al. 2018 ESC/ESH Guidelines for the management of arterial hypertension: The Task Force for the management of arterial hypertension of the European Society of Cardiology and the European Society of Hypertension: The Task Force for the management of arterial hypertension of the European Society of Cardiology and the European Society of Hypertension. *Journal of Hypertension*. 2018;36(10):1953–2041. doi:10.1097/HJH.0000000000001940.

8. Whelton PK, Carey RM, Aronow WS, et al. 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA Guideline

for the Prevention, Detection, Evaluation, and Management of High Blood Pressure in Adults: Executive Summary: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *Journal of the American College of Cardiology*. 2018;71(19):2199–2269. doi:10.1016/j.jacc.2017.11.005.

9. Rapsomaniki E, Timmis A, George J, et al. Blood pressure and incidence of twelve cardiovascular diseases: lifetime risks, healthy life-years lost, and age-specific associations in 1·25 million people. *Lancet*. 2014;383(9932):1899–1911. doi:10.1016/S0140-6736(14)60685-1.

10. Thomopoulos C, Parati G, Zanchetti A. Effects of blood pressure lowering on outcome incidence in hypertension. 1. Overview, meta-analyses, and meta-regression analyses of randomized trials. *Journal of Hypertension*. 2014;32(12):2285–2295. doi:10.1097/HJH.0000000000000378.

11. Thomopoulos C, Parati G, Zanchetti A. Effects of blood pressure lowering on outcome incidence in hypertension: 2. Effects at different baseline and achieved blood pressure levels--overview and meta-analyses of randomized trials. *Journal of Hypertension*. 2014;32(12):2296–2304. doi:10.1097/HJH.0000000000000379.

12. Thomopoulos C, Parati G, Zanchetti A. Effects of blood pressure lowering on outcome incidence in hypertension: 3. Effects in patients at different levels of cardiovascular risk--overview and meta-analyses of randomized trials. *Journal of Hypertension*. 2014;32(12):2305–2314. doi:10.1097/HJH.0000000000000380.

13. Ettehad D, Emdin CA, Kiran A, et al. Blood pressure lowering for prevention of cardiovascular disease and death: a systematic review and meta-analysis. *The Lancet*. 2016;387(10022):957–967. doi:10.1016/S0140-6736(15)01225-8.

14. Forouzanfar MH, Liu P, Roth GA, et al. Global Burden of Hypertension and Systolic Blood Pressure of at Least 110 to 115 mm Hg, 1990-2015. *JAMA*. 2017;317(2):165. doi:10.1001/jama.2016.19043.

15. Heidenreich PA, Trogdon JG, Khavjou OA, et al. Forecasting the Future of Cardiovascular Disease in the United States: A Policy Statement From the American Heart Association. *Circulation*. 2011;123(8):933–944. doi:10.1161/CIR.0b013e31820a55f5.

16. Johnson HM, Thorpe CT, Bartels CM, et al. Antihypertensive Medication Initiation Among Young Adults with Regular Primary Care Use. *J GEN INTERN MED*. 2014;29(5):723–731. doi:10.1007/s11606-014-2790-4.

17. Chiuve SE, Cook NR, Shay CM, et al. Lifestyle-Based Prediction Model for the Prevention of CVD: The Healthy Heart Score. *Journal of the American*

Heart Association. 2014;3(6):e000954–e000954. doi:10.1161/JAHA.114.000954.

18. Schoenthaler A, Luerassi L, Silver S, et al. Comparative Effectiveness of a Practice-Based Comprehensive Lifestyle Intervention vs. Single Session Counseling in Hypertensive Blacks. *American Journal of Hypertension*. 2016;29(2):280–287. doi:10.1093/ajh/hpv100.

19. Golshahi J, Ahmadzadeh H, Sadeghi M, Mohammadifard N, Pourmoghaddas A. Effect of self-care education on lifestyle modification, medication adherence and blood pressure in hypertensive adults: Randomized controlled clinical trial. *Adv Biomed Res*. 2017;4:204. doi:10.4103/2277-9175.166140.

20. Institute for Health Metrics. Institute for Health Metrics Country Profile: United States. www.healthdata.org.

21. Go AS, Bauman MA, Coleman King SM, et al. An effective approach to high blood pressure control: a science advisory from the American Heart Association, the American College of Cardiology, and the Centers for Disease Control and Prevention. *Hypertension*. 2014;63(4):878–885. doi:10.1161/HYP.0000000000000003.

22. Eckel RH, Jakicic JM, Ard JD, et al. 2013 AHA/ACC Guideline on Lifestyle Management to Reduce Cardiovascular Risk: A Report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *Circulation*. November 2013. doi:10.1161/01.cir.0000437740.48606.d1.

23. Jensen MD, Ryan DH, Apovian CM, et al. *2013 AHA/ACC/TOS Guideline for the Management of Overweight and Obesity in Adults*. 2013:n/a–n/a. doi:10.1002/oby.20660.

24. Cerezo C, Sequra J, Praga M, Ruilope LM. Guidelines updates in the treatment of obesity or metabolic syndrome and hypertension. *Current Hypertension Reports*. 2013;15(3):196–203.

25. Barry SA, Harlan DM, Johnson NL, MacGregor KL. State of Behavioral Health Integration in U.S. Diabetes Care: How Close Are We to ADA Recommendations? *Diabetes Care*. 2018;41(7):e115–e116. doi:10.2337/dc18-0642.

26. Anderson JL, Halperin JL, Albert NM, et al. Management of Patients With Peripheral Artery Disease (Compilation of 2005 and 2011 ACCF/AHA Guideline Recommendations): A Report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines. *Circulation*. 2013;127(13):1425–1443. doi:10.1161/CIR.0b013e31828b82aa.

27. Weber MA, Schiffrin EL, White WB, et al. Clinical Practice Guidelines for the Management of Hypertension in the Community. *The Journal of Clinical Hypertension*. December 2013:n/a–n/a. doi:10.1111/jch.12237.

28. Mozaffarian D, Afshin A, Benowitz NL, et al. Population approaches to improve diet, physical activity, and smoking habits: a scientific statement from the American Heart Association. *Circulation*. 2012;126(12):1514–1563. doi:10.1161/CIR.0b013e318260a20b.

29. Chobanian AV, Hill M. National Heart, Lung, and Blood Institute Workshop on Sodium and Blood Pressure : A Critical Review of Current Scientific Evidence. *Hypertension*. 2000;35(4):858–863. doi:10.1161/01.HYP.35.4.858.

30. Lin JS, O'Connor E, Evans CV, Senger CA, Rowland MG, Groom HC. Behavioral counseling to promote a healthy lifestyle in persons with cardiovascular risk factors: a systematic review for the U.S. Preventive Services Task Force. *Annals of Internal Medicine*. 2014;161(8):568–578. doi:10.7326/M14-0130.

31. Sacks FM, Lichtenstein AH, Wu JHY, et al. Dietary Fats and Cardiovascular Disease: A Presidential Advisory From the American Heart Association. *Circulation*. June 2017:CIR.0000000000000510. doi:10.1161/CIR.0000000000000510.

32. Whelton SP, Chin A, Xin X, He J. Effect of aerobic exercise on blood pressure, a meta-analysis of randomized, controlled trials. *Annals of Internal Medicine*. 2002.

33. Xin X, He J, Frontini MG, Ogden LG, Motsamai OI, Whelton PK. Effects of Alcohol Reduction on Blood Pressure: a meta-analysis of randomized controlled trials. *Hypertension*. 2001;38(5):1112–1117.

34. Vamvakis A, Gkaliagkousi E, Triantafyllou A, Gavriilaki E, Douma S. Beneficial effects of nonpharmacological interventions in the management of essential hypertension. *JRSM Cardiovascular Disease*. 2017;6:204800401668389. doi:10.1161/HYPERTENSIONAHA.111.177071.

35. Mahmood S, Shah KU, Khan TM, et al. Non-pharmacological management of hypertension: in the light of current research. *Ir J Med Sci*. 2018;317(2):165. doi:10.1038/nrneph.2009.191.

36. He J, Whelton PK, Appel LJ, Charleston J, Klag MJ. Long-term effects of weight loss and dietary sodium reduction on incidence of hypertension. *Hypertension*. 2000;35(2):544–549.

37. Sacks F, Svetkey LP, Vollmer WM, et al. Effects on blood pressure of reduced dietary sodium and the Dietary Approaches to Stop Hypertension

(DASH) diet. *New England …*. 2001;344:3–10.

38. Johnson HM, Olson AG, LaMantia JN, et al. Documented Lifestyle Education Among Young Adults with Incident Hypertension. *J GEN INTERN MED*. 2014;30(5):556–564. doi:10.1007/s11606-014-3059-7.

39. Sinclair J, Lawson B, Burge F. Which patients receive advice on diet and exercise? Do certain characteristics affect whether they receive such advice? *Can Fam Physician*. 2008;54(3):404–412.

40. Corsino L, Svetkey LP, Ayotte BJ, Bosworth HB. Patient characteristics associated with receipt of lifestyle behavior advice. *N C Med J*. 2009;70(5):391–398.

41. Shoenbill K, Song Y, Gress L, Johnson H, Smith M, Mendonca E. Natural Language Processing of Lifestyle Modification Documentation. *Health Informatics Journal*. (In press.).

42. Melvin CL, Jefferson MS, Rice LJ, et al. A systematic review of lifestyle counseling for diverse patients in primary care. *Preventive Medicine*. 2017;100:67–75. doi:10.1016/j.ypmed.2017.03.020.

43. Li Y, Pan A, Wang DD, et al. Impact of Healthy Lifestyle Factors on Life Expectancies in the US Population. *Circulation*. April 2018:CIRCULATIONAHA.117.032047. doi:10.1161/CIRCULATIONAHA.117.032047.

44. Sisti LG, Dajko M, Campanella P, Shkurti E, Ricciardi W, de Waure C. The effect of multifactorial lifestyle interventions on cardiovascular risk factors: a systematic review and meta-analysis of trials conducted in the general population and high risk groups. *Preventive Medicine*. 2018;109:82–97. doi:10.1016/j.ypmed.2017.12.027.

45. Viera AJ, Kshirsagar AV. Lifestyle modification advice for lowering or controlling high blood pressure: who's getting it? *… of Clinical Hypertension*. 2007.

46. Honda K. Factors underlying variation in receipt of physician advice on diet and exercise: applications of the behavioral model of health care utilization. *Am J Health Promot*. 2004;18(5):370–377. doi:10.4278/0890-1171-18.5.370.

47. Ahmed NU, Delgado M, Saxena A. Trends and disparities in the prevalence of physicians' counseling on diet and nutrition among the U.S. adult population, 2000–2011. *Preventive Medicine*. 2016;89:70–75. doi:10.1016/j.ypmed.2016.05.014.

48. Ahmed NU, Delgado M, Saxena A. Trends and disparities in the prevalence of physicians' counseling on exercise among the U.S. adult population,

2000–2010. *Preventive Medicine*. 2017;99:1–6. doi:10.1016/j.ypmed.2017.01.015.

49. Turchin A, Kolatkar NS, Grant RW, Makhni EC, Pendergrass ML, Einbinder JS. Using Regular Expressions to Abstract Blood Pressure and Treatment Intensification Information from the Text of Physician Notes. *Journal of the American Medical Association*. 2006;13(6):691–695. doi:10.1197/jamia.M2078.

50. Savova GK, Masanz JJ, Ogren PV, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Association*. 2010;17(5):507–513. doi:10.1136/jamia.2009.001560.

51. Mendonca EA, Cimino JJ, Johnson S. Using Narrative Reports to Support a Digital Library. In:; 2001:1–5.

52. Mendonca EA, Haas J, Shagina L, Larson E, Friedman C. Extracting information on pneumonia in infants using natural language processing of radiology reports. *Journal of Biomedical Informatics*. 2005;38(4):314–321. doi:10.1016/j.jbi.2005.02.003.

53. Murff HJ, FitzHenry F, Matheny ME, et al. Automated identification of postoperative complications within an electronic medical record using natural language processing. *JAMA*. 2011;306(8):848–855. doi:10.1001/jama.2011.1204.

54. Salmasian H, Freedberg DE, Friedman C. Deriving comorbidities from medical records using natural language processing. *Journal of the American Medical Association*. 2013;20(e2):e239–e242. doi:10.1136/amiajnl-2013-001889.

55. Lindholm C, Adsit R, Bain P, et al. A demonstration project for using the electronic health record to identify and treat tobacco users. *WMJ*. 2010;109(6):335–340.

56. Stevens V, Bailey S, Hazlehurst B, Kurtz S. PS1-1d: Use of CER Hub to Evaluate Outcomes of Smoking Cessation Services, a Behavioral Treatment. *Clin Med Res*. 2013.

57. Hazlehurst BL, Lawrence JM, Donahoo WT, et al. Automating Assessment of Lifestyle Counseling inElectronic Health Records. *AMEPRE*. 2014;46(5):457–464. doi:10.1016/j.amepre.2014.01.001.

58. Abraham C, Michie S. A taxonomy of behavior change techniques used in interventions. *Health Psychology*. 2008;27(3):379–387. doi:10.1037/0278-6133.27.3.379.

59. Elasy TA, Ellis SE, Brown A, Pichert JW. A taxonomy for diabetes educational interventions. *Patient Educ Couns*. 2001;43(2):121–127. doi:10.1016/S0738-3991(00)00150-6.

60. Michie S, Ashford S, Sniehotta FF, Dombrowski SU, Bishop A, French DP. A refined taxonomy of behaviour change techniques to help people change their physical activity and healthy eating behaviours: the CALO-RE taxonomy. *Psychol Health*. 2011;26(11):1479–1498. doi:10.1080/08870446.2010.540664.

61. Vrijens B, De Geest S, Hughes DA, et al. A new taxonomy for describing and defining adherence to medications. *British Journal of Clinical Pharmacology*. 2012;73(5):691–705. doi:10.1111/j.1365-2125.2012.04167.x.

62. Zeng QT, Tse T. Exploring and Developing Consumer Health Vocabularies. *Journal of the American Medical Association*. 2006;13(1):24–29. doi:10.1197/jamia.M1761.

63. Kawaler E, Cobian A, Peissig P, Cross D, Yale S, Craven M. Learning to predict post-hospitalization VTE risk from EHR data. *AMIA Annu Symp Proc*. 2012;2012:436–445.

64. Roque FS, Jensen PB, Schmock H, et al. Using Electronic Patient Records to Discover Disease Correlations and Stratify Patient Cohorts. Ritchie MD, ed. *PLoS Comput Biol*. 2011;7(8):e1002141. doi:10.1371/journal.pcbi.1002141.s004.

65. Kohane IS. Using electronic health records to drive discovery in disease genomics. *Nature reviews Genetics*. 2011;12(6):417–428. doi:10.1038/nrg2999.

66. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. *Nature reviews Genetics*. 2012;13(6):395–405. doi:10.1038/nrg3208.

67. Shoenbill K, Song Y, Cobb NL, Drezner MK, Mendonca EA. IRB Process Improvements: A Machine Learning Analysis. *J Clin Trans Sci*. 2017;2012:1–8. doi:10.1200/JOP.2010.000051.

68. Kingsford C, Salzberg S. What are decision trees? *Nature Biotechnology*. 2008;26(9):1011–1013.

69. Dietterich TG. Ensemble methods in machine learning. *Multiple classifier systems*. 2000.

70. Dietterich TG. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine learning*. 2000.

71. Ye C, Fu T, Hao S, et al. Prediction of Incident Hypertension Within the Next Year: Prospective Study Using Statewide Electronic Health Records and Machine Learning. *J Med Internet Res*. 2018;20(1):e22. doi:10.1016/j.ijmedinf.2015.06.007.

72. Koren G, Nordon G, Radinsky K, Shalev V. Machine learning of big data in gaining insight into successful treatment of hypertension. *Pharmacol Res Perspect*. 2018;6(3):e00396. doi:10.1016/S104366180200124X.

73. Khanji C, Lalonde L, Bareil C, Lussier M-T, Perreault S, Schnitzer ME. Lasso Regression for the Prediction of Intermediate Outcomes Related to Cardiovascular Disease Prevention Using the TRANSIT Quality Indicators. *Medical Care*. November 2018. doi:10.1097/MLR.0000000000001014.

74. Krittanawong C, Bomback AS, Baber U, Bangalore S, Messerli FH, Wilson Tang WH. Future Direction for Using Artificial Intelligence to Predict and Manage Hypertension. *Current Hypertension Reports*. 2018;20(9):591. doi:10.1161/CIRCULATIONAHA.118.034390.

75. Johnson HM, Thorpe CT, Bartels CM, et al. Undiagnosed hypertension among young adults with regular primary care use. *Journal of Hypertension*. 2014;32(1):65–74. doi:10.1097/HJH.0000000000000008.

76. Carroll RJ, Eyler AE, Denny JC. Naïve Electronic Health Record phenotype identification for Rheumatoid arthritis. *AMIA Annu Symp Proc*. 2011;2011:189–196.

77. LePendu P, Iyer SV, Bauer-Mehren A, et al. Pharmacovigilance using Clinical Text. *AMIA Jt Summits Transl Sci Proc*. 2013;2013:109.

78. Iyer SV, Harpaz R, LePendu P, Bauer-Mehren A, Shah NH. Mining clinical text for signals of adverse drug-drug interactions. *Journal of the American Medical Association*. 2014;21(2):353–362. doi:10.1136/amiajnl-2013-001612.

79. Hivert M-F, Grant RW, Shrader P, Meigs JB. Identifying primary care patients at risk for future diabetes and cardiovascular disease using electronic health records. *BMC Health Services Research*. 2009;9:170–170. doi:10.1186/1472-6963-9-170.

80. Dean BB, Lam J, Natoli JL, Butler Q, Aguilar D, Nordyke RJ. Review: Use of Electronic Medical Records for Health Outcomes Research: A Literature Review. *Medical Care Research and Review*. 2009;66(6):611–638. doi:10.1177/1077558709332440.

81. Ohno-Machado L. Realizing the full potential of electronic health records: the role of natural language processing. *Journal of the American Medical Association*. 2011;18(5):539. doi:10.1136/amiajnl-2011-000501.

82. Weiskopf NG, Hripcsak G, Swaminathan S, Weng C. Defining and measuring completeness of electronic health records for secondary use. *J Biomed Inform*. 2013;46(5):830–836. doi:10.1016/j.jbi.2013.06.010.

83. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. *Journal of the American Medical Association*. 2012;20(1):117–121. doi:10.1136/amiajnl-2012-001145.

84. Hersh WR, Weiner MG, Embi PJ, et al. Caveats for the Use of Operational Electronic Health Record Data in Comparative Effectiveness Research. *Medical Care*. 2013;51:S30–S37. doi:10.1097/MLR.0b013e31829b1dbd.

85. Wang RY, Strong DM. Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems*. 1996;12(4):5–33. doi:10.2307/40398176?ref=no-x-route:07e4e11659dafd6554b009c67355420c.

86. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *Journal of the American Medical Association*. 2013;20(1):144–151. doi:10.1136/amiajnl-2011-000681.

87. Boone KW. Clinical Documentation. In: *The CDA ™ Book*. Springer-Verlag London; 2011. doi:10.1007/978-0-85729-336-7_2.

88. HealthITgov, ed. *Improved Diagnostics & Patient Outcomes*. Washingtongov; 2015. https://www.healthit.gov/providers-professionals/improved-diagnostics-patient-outcomes. Published October 3, 2015. Accessed October 3, 2015.

7. Appendices

7.1. Appendix A - Example: Lifestyle Modification Terms and Objects

General LM Objects - NP	Provider VP Requiring Any LM Object = Advice
alcohol	advise/d/ing
alcohol abuse	agrees/d/ing/able
alcohol use	avoid
alcoholic beverages	begin
Alcoholics Annonymous/A	change
amphetamine/s	congratulated
bike/s/ing/d	consider/ing/d/s
binges/ing	continue
caloric	counsel/ed/ing
calorie/s	covered
carbohydrate/s	declines/d/ing
Chantix	decrease
cigarette/s	discuss/ing/ed
cigars	do not
DASH diet	don't
decrease/decreasing/decr	emphasized
diet	encourage/d/ing
dietary	give/gave/giving/given
dietician	given handouts on/for/about (LM Object)
drug abuse	given information on/for/about (LM Object)
drug use	given instruction on/for/about (LM Object)
etoh/ETOH	going to
exercise/exercising/has ex	improve
fats/fat	increase
fruit	instruction/instructed
gain/gaining/gained weigh	is/was not interested in
goal/target/healthy BMI	monitor
gym	needs to
health club	prescribe/prescribed/prescribing

7.2. Appendix B – Example: Lifestyle Modification TUI-CUI Mapping

NP-LM Object List File Columns and Terms to Be Matched with TUIs/CUIs	Concept (SNOMEDCT)	TUI	CUI	LM Object Term from Provider Notes	LM Object Category
Provider VP (Column B & C) Followed by Diet LM Object	Dietary mgmt. education, g	T058	C1828150	alcohol	Alcohol
Provider VP (Column B & C) Followed by Exercise Object	Exercise education	T058	C0582396	alcohol abuse	Alcohol
Provider VP (Column B & C) Followed by Weight Object	Patient advised about weigh	T058	C3697318	alcohol use	Alcohol
Provider VP (Column B & C) Followed by Smoking Object	Smoking cessation assistanc	T061	C1692317	alcoholic beverages	Alcohol
Provider VP (Column B & C) Followed by Alcohol Object	Alcohol consumption counse	T058	C1531491	Alcoholics Annonymous/AA	Alcohol
Provider VP (Column B & C) Followed by Drug Abuse Object	Drug addiction counseling	T061	C0199403	amphetamine/s	Drug addiction/use
				bike/s/ing/d	Exercise
Declaration Column D or Patient VP (Column E & F) Regarding Diet	Dietary history	T033	C042501	binges/ing	Diet
Declaration Column D or Patient VP (Column E & F) Regarding Exercise	Exercise history	T033	C1287528	caloric	Diet
Declaration Column D or Patient VP (Column E & F) Regarding Weight	Weight finding	T033	C1265588	calorie/s	Diet
Declaration Column D or Patient VP (Column E & F)Regarding Smoking	Smoking assessment	T058	C3853073	carbohydrate/s	Diet
Declaration Column D or Patient VP (Column E & F)Regarding Alcohol	Assessment of alcohol use	T058	C4076406	Chantix	Smoking
Declaration Column Dor Patient VP (Column E & F) Regarding Drug Abuse	Assessment of drug use	T058	C4075408	cigarette/s	Smoking
				cigars	Smoking
				DASH diet	Diet

7.3. Appendix C – Example: Diagnosis Filter List

Hypertension Diagnosis Terms (case insensitive)	Hypertension Diagnosis Acronyms (case insensitive)	Hypertension-Related Diagnosis Terms (case insensitive)	Hypertension-Related Diagnosis Acroynm
blood pressure is elevated	HBP	acute myocardial infarction	AMI
borderline blood pressure	HTN	alcohol abuse	ASCAD
elevated b/p		alcohol addiction	CAD
elevated blood pressure		alcohol dependence	CHF
elevated BP		bordeline blood sugar	CVA
high blood pressure		borderline cholesterol	CVD
hypertension		borderline lipid profile/lab	DM
hypertensive		borderline lipids	MI
increased blood pressure		cardiomyopathy	PCOS
		cardiovascular disease	PVD
		cerebral vascular accident	
		cerebral vascular attack	
		cerebral vascular disease	
		cerebrovascular accident	
		cerebrovascular attack	
		cerebrovascular disease	

7.4. Appendix D – Example: Family History Concept Mapping

FH: Cardiovascular Disease	C0455404	FH: hypertension	C0455405	n/a	n/a	n/a
FH: Cardiovascular Disease	C0455404	FH: stroke	C12611367	n/a	n/a	n/a
FH: Cardiovascular Disease	C0455404	FH: stroke	C12611367	n/a	n/a	n/a
FH: Cardiovascular Disease	C0455404	FH: stroke	C12611367	n/a	n/a	n/a
FH: Cardiovascular Disease	C0455404	FH: stroke	C12611367	n/a	n/a	n/a
FH: Cardiovascular Disease	C0455404	FH: stroke	C12611367	n/a	n/a	n/a
FH: Cardiovascular Disease	C0455404	FH: stroke	C12611367	n/a	n/a	n/a
FH: Cardiovascular Disease	C0455404	FH: stroke	C12611367	n/a	n/a	n/a
FH: Cardiovascular Disease	C0455404	FH: transient ischemic attack	C0475701	n/a	n/a	n/a
FH: Cardiovascular Disease	C0455404	FH: transient ischemic attack	C0475701	n/a	n/a	n/a
FH: Cardiovascular Disease	C0455404	FH: cardiac disorder	C0559128	FH: ischemic heart d	C1313980	FH: myocardial inf
FH: Cardiovascular Disease	C0455404	FH: cardiac disorder	C0559128	FH: ischemic heart d	C1313980	FH: myocardial inf
FH: Cardiovascular Disease	C0455404	FH: cardiac disorder	C0559128	FH: ischemic heart d	C1313980	FH: myocardial inf
FH: Cardiovascular Disease	C0455404	FH: cardiac disorder	C0559128	FH: ischemic heart d	C1313980	FH: myocardial inf
FH: Cardiovascular Disease	C0455404	FH: cardiac disorder	C0559128	FH: coronary arterio	C2317524	n/a
FH: Cardiovascular Disease	C0455404	FH: cardiac disorder	C0559128	FH: coronary arterio	C2317524	n/a

7.5. Appendix E – IRB Approval

THE UNIVERSITY of WISCONSIN MADISON

Minimal Risk IRB (Health Sciences)
1/23/2018

Submission ID number:	2014-0380-CR004
Title:	Lifestyle Modification Activities for Hypertension Management: A Machine Learning Analysis
Principal Investigator:	ENEIDA A MENDONCA
Point-of-contact:	ENEIDA A MENDONCA
IRB Staff Reviewer:	EMILY JENNER

A designated MR IRB member conducted an expedited review of the above-referenced continuing review progress report form. As part of its review, the IRB determined this study does not require continuing review either under federal regulations or institutional policy, or both. Please note, however, that although this study is not required to undergo continuing review, you must still submit the following to the IRB:

1. Changes of protocol prior to their implementation (unless the change is necessary to eliminate an apparent immediate hazard to subjects)
2. Addition of new study personnel
3. Funding updates
4. Reportable events (unanticipated problems, noncompliance, new information) in accordance with institutional policy
5. Closure report

In addition, please be aware that the type of funding that supports a study or whether the study falls under FDA regulations can affect whether continuing review may be required in future.

The study qualified for expedited review pursuant to 45 CFR 46.110 and, if applicable, 21 CFR 56.110 and 38 CFR 16.110:

Category 5: Research involving materials (data, documents, records, or specimens) that have been collected, or will be collected solely for nonresearch purposes (such as medical treatment or diagnosis)

To access the materials approved by the IRB, including any stamped consent forms and recruitment materials, please log in to your ARROW account and view the documents tab in the submission's workspace.

Please review the Investigator Responsibilities guidance

(https://kb.wisc.edu/hsirbs/page.php?id=18881), which includes a description of IRB requirements for submitting personnel changes, changes of protocol and reportable events.

If you have general questions, please contact the Health Sciences IRBs at 608-263-2362. For questions related to this submission, contact the assigned staff reviewer.