AUTOMATIC CLASSIFICATION OF PATIENT-GENERATED MESSAGES FROM A PATIENT PORTAL

By

Robert Michael Cronin, II

Thesis

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

MASTER OF SCIENCE

in

Biomedical Informatics

December, 2015

Nashville, Tennessee

Approved:
Gretchen Purcell Jackson, M.D., Ph.D.
Joshua Denny, M.D., M.S.
S. Trent Rosenbloom, M.D., M.P.H

ACKNOWLDEGEMENTS

This work would not be possible without the financial support of the National Library of Medicine Fellowship in Biomedical Informatics through the training grant 5T15LM007450-12. I am grateful to the mentorship and direction of my Dissertation Committee, Gretchen Jackson, Trent Rosenbloom, and Josh Denny, who have guided and contributed invaluable insights into this work. Each of the members of my Dissertation Committee has provided me extensive personal and professional guidance and taught me a great deal about both scientific research and life in general. I am indebted to Michael Matheny and Daniel Fabbri, whose guidance and expertise allowed me to complete this work. I am also grateful to the annotators of the gold standard corpus: Jared Shenson, Ebone Ingram, and Mary Masterman. They eagerly dedicated hours of annotations to this work. I am also grateful to everyone in the Department of Biomedical Informatics who has helped me through my fellowship.

TABLE OF CONTENTS

Page

ACKNOWLDEGEMENTS ii

LIST OF TABLES iv

LIST OF FIGURES v

INTRODUCTION 1

BACKGROUND 3

Health Information Needs 3

Technologies that Address Health Information Needs 4

Text Classification 7

Categorization Schemes for Consumer Health Information Needs 10

METHODS 13

Overview 13

Setting/Gold Standard (Outputs) 14

Automated Classifiers 15

Evaluation and Statistical Analysis 19

RESULTS 22

Gold Standard 22

Predictive Performance for Individual Communication Types 23

Predictive Performance for All Communication Types Within A Message 32

Features 35

DISCUSSION 37

Gold Standard 37

Performance for single communication types 38

Features 41

Limitations 43

Future Directions 44

REFERENCES 46

LIST OF TABLES

Table Page

1. Definitions of communication types 16

2. Words used for major communication types the basic classifier 17

3. Words used for subtypes for the basic classifier 17

4. Distribution of messages across time 22

5. Distribution of communication types among messages 23

6. Communication type co-occurrence matrix 23

7. Most common communication subtypes 23

8. Area under the curves for clinical information types 24

9. Area under the curves for logistical types 24

10. Area under the curves for social types 25

11. Area under the curves for medical types 25

12. Area under the curves for social/acknowledgement subtype 29

13. Area under the curves for logistical/contact information subtype 29

14. Area under the curves for medical/appointments/scheduling subtype 30

15. Area under the curves for medical/prescription subtype 30

16. Area under the curves for medical/problems subtype 31

17. Area under the curves for medical/follow-up subtype 31

18. Area under the curves for medical/tests subtype 32

19. Jaccard indices for major types and subtypes 33

20. Top 10 features for random forests 36

LIST OF FIGURES

Figure Page

1. The taxonomy of consumer health information needs or communications 12

2. Example of input, classifiers, and output system 13

3. Example message labelled by communication types 17

4. Example message with selected features 19

5. Classification process and evaluation metrics 20

6. Area under the curves of different major communication types 26

7. Bar charts of Jaccard indices.. 34

INTRODUCTION

People with questions about their health have been seeking answers from online resources with increasing frequency [1-4]. These questions result from deficiencies in personal knowledge and can include questions such as “what is this lump on my leg” or “how do I get to my flu shot”. These deficiencies in knowledge have been termed: consumer health information needs [5]. People with consumer health information needs seek information from a variety of online sources, such as general internet search engines, curated health websites, social networking sites, and patient portals [1,2,4,6-9]. Patient portals, electronic applications that allow patients to interact with their healthcare providers, have had increasing adoption because of consumer demand and governmental regulations [10]. Secure messaging is one of the most popular functions of patient portals, and it allows for individuals to communicate with their healthcare providers [6,11-17]. This thesis proposes methods to identify the types of information needs expressed in secure messaging through a patient portal.

One method to identify health information needs in secure messages is text classification. Text classification involves the assignment of the words in a document to one or more categories [18]. Text classification can be carried out through automated methods, including a basic rule-based approach that assigns categories based on rules, or machine learning techniques that learn categories [19-25]. Previous literature has examined classification of secure messages through patient portals predominately in primary care settings only using manual methods [26,27]. By automatically classifying secure messages, the types of communications occurring in patient portals can be better understood, and health information needs expressed in messages could potentially be triaged to resources that resolve those needs.

This thesis uses text processing methods to explore the diverse communications types within the contents of secure messages. Some messages contain requests for clinical information, such as questions that might be answered by a textbook. Others express needs for medical care, such as the communication of a new symptom or the request for a particular medical service. Secure messages through patient portals may also communicate requests for logistical information, such as directions to the hospital, or social interactions, such as an expression of gratitude or a complaint.

This thesis describes the development and evaluation of automated classifiers to categorize patient-generated secure portal messages into communication types, such as clinical information, medical, logistical, or social types. We evaluated the ability of our classifiers to identify a single communication type within a message and to predict all communication types within a single message. We compared our classifiers against a gold standard manually labeled message corpus.

BACKGROUND

Health Information Needs

This section reviews health information needs and motivates this research to develop methods for automated identification and classification of consumer health information needs. This section then describes the ways consumers can meet their needs using health information technologies and motivates the use of patient portals as a means for expressing and potentially meeting consumer health information needs. The following sections define health information needs, describe variations by stakeholder, and explain the importance of information needs in health care.

Belkin et al. defined health information needs as an “anomaly” in a person’s state of health-related knowledge [5]. For example, a person who developed a new lump in the leg might have a question about that lump. A health information need arises as they do not have knowledge about the cause of the lump. Health information needs vary according to the person's role [28-39]. The person’s role can include the patient themselves; their health care team members, such as clinicians, nurses, and case managers; or their caregivers, such as family members. Information needs of health care team members have been well studied [28-35], and they are often distinct from needs of patients and their caregivers [36]. Health information needs of caregivers also differ from those of the patients themselves [37-39].

This topic is significant; research has demonstrated that meeting patients’ health information needs can improve health outcomes, such as hospital readmission rates, quality of life, and mortality rates [40-44]. Consumers of health care have also increased their demand for resources to meet health information needs [1-4]. Prior to the 2000’s, consumers resolved their health information needs primarily through their health care providers or the public library [45]. In 2000, Jones envisioned technologies that consumers might use to answer their questions during the next decade, such as consumer sites on the world wide web, electronic mail messages between consumers and physicians, and technology-based communications among consumers with similar conditions [46]. These predictions have largely come to pass [47-52]. The proportion of American online users who search for health related topics has increased to approximately 80% [4]. Since about 15% of adults do not use the internet [53], about two-thirds of all adults search for health information online. In a Harris Poll, 57% of consumers reported discussing internet searches with their physicians and 57% searched the internet after seeing their physician [8]. Consumers have confidence in the internet's ability to answer their health questions; 90% felt their search were very or somewhat successful [8]. Analysis of postings content from social networking sites demonstrate that consumers most frequently exchange information, describe their experiences, support each other, interact with peers, and promote behavior change [48]. These studies demonstrate a shift in the way consumers meet their health information needs toward increased use of online tools.

Technologies that Address Health Information Needs

Consumers' use a variety of technologies to address their health information needs [1,2,4,6-9], including general search engines such as Google; dedicated health web sites, which contain vetted health information; social networking sites, where consumers can talk with others about their health information needs; and patient portals, technologies that support interaction between consumers and health institutions through a variety of resources and tools. Each type of technology supplying consumer health information has advantages and shortcomings. General search engines can perform comprehensive searches and generate diverse results, but they may yield a lot of irrelevant information or low quality answers. An older study of internet searches for the treatment of childhood diarrhea demonstrated that 20% of results failed to match then-current American Academy of Pediatrics guidelines [54]. Many consumers seek information from websites created by reputable health organizations, such as WebMD [55], MedlinePlus [56], UpToDate [57], and Mayo Clinic [58]. These sites have extensive health libraries with materials about treatments, drug information, symptom searches, and medical news for men, women, and children. They are particularly useful for common problems and established knowledge, but may fall short in addressing needs related to rare diseases and emerging areas of medical science.

As a consumer health information resource, social networks are communities formed through connections among people based on similar features or interests in their living or working environments [59]. Websites designed to support social networking, such as Facebook or internet forums, allow people to interact with large social networks of individuals spread over wide geographic areas. On these websites, interactions commonly take place when people create, share, and exchange information and ideas. Through these interactions, users can make connections that may not occur otherwise and disseminate consumer health information [60]. There are also social networking websites that are built around an interest in healthcare issues or diseases, including PatientsLikeMe [61], MedHelp [62], DailyStrength [63], and Tudiabetes [64]. These websites can help healthcare providers and consumers communicate about medical knowledge, literature, management, self-care, patient engagement, and personal health data sharing [65,66]. However, concerns exist about the accuracy of information provided by such sites [67-70].

Patient portals, a consumer health information resource, are web-based applications that enable patients and their caregivers to interact with healthcare systems and health information [71-79]. Patient portals typically allow users to schedule appointments, access parts of the electronic health record (EHR), manage medical bills, receive personalized health information, and communicate with healthcare providers through secure messaging [80]. Hundreds of institutions have implemented patient portals [81-89], with increasing adoption being driven by consumer demand and government mandates such as Meaningful Use criteria [10]. These portals have been implemented in diverse settings including large academic medical centers [86,87,90-94], community practices [95], adult and pediatric primary care [96,97], and specialty care [84,98].   Patient portals have been shown to increase satisfaction with care, enhance communication between patients and providers, expand access to health information, and improve outcomes for patients with selected diseases [86,99-105].

Using technologies such as secure messaging in patient portals, consumers can resolve their information needs [6,11-17]. Educational resources, such as a flu tool [106], may address common consumer health questions, and secure messaging allows consumers to express their specific needs and receive personalized answers. Secure patient-provider messaging is one of the most popular functions of patient portals [92,97,107,108] and has been widely adopted across clinical specialties [109]. Prior research has shown that clinical care is delivered through secure message exchanges [98,110]. For example, patients may report new problems, and secure messages may facilitate further evaluation and treatment [111]. Thus, messaging within a patient portal is more than a new communication modality to support administrative tasks. Instead, portal messaging can be considered an evolving form of outpatient interaction through which healthcare is delivered.

Only a few research studies have focused specifically on the nature of secure messaging through patient portals. These studies examined portals in primary care settings, with manual annotation of secure messages. North et al. manually classified 323 messages, demonstrating 37% of messages were medication related, 23% were symptom related, 20% were test related, 7% had to do with medical questions, 6% were acknowledgements, and 9% had greater than one issue [27]. Haun et al. had senders classify their messages in the following categories: 59% in the general category (condition management/report, specialty/procedure request, correspondence request, medication refill request, test results, appointment requests, treatment/appointment follow-up), 24% in appointments (confirmations, cancellations, specialty appointment requests), 16% had to do with refill requests and medication inquiries, and 2% had to do with test requests [26]. As patient portal and secure messaging adoption increases, understanding the nature of these interactions and their implications for provider workload becomes more important. With millions of messages exchanged each year, automated techniques for understanding their content are needed.

Text Classification

The task of categorizing the content of secure messages is an example of a text classification task. This section reviews the methods applicable to the task of classifying unstructured text in patient portals to categories of health information needs, or more generally communication types. Text classification is defined, and the machine learning methods that can classify text into categories are described. Text classification is a broad topic which has been widely applied in healthcare applications and reviewed extensively elsewhere [112]. This section focuses on literature specifically relevant to the task of secure message classification.

Classification consists of assigning items to categories [113]. Text classification involves the assignment of the words in a document to one or more categories [18]. Classification tasks can be done manually or using automated techniques. In manual classification, a human will read document text and assign categories [114]. This task can be time consuming and expensive. A sufficiently skilled person who knows the domain very well can create a set of rules that can classify text well; however, such a person may be difficult to find, or may not exist. Machine learning classification uses sets of text that have already been classified to determine rules automatically that can be used on unclassified text for categorization. The inputs or features that are used for machine learning methods can vary. In the bag of words model, the features are the individual words in document and their frequency of occurrence [115]. More sophisticated methods can be used to determine the features, such as topic modeling, n-grams, and natural language processing (NLP) [115]. Topic modeling employs probabilistic models that uncover the underlying semantic structure of a collection of documents [116]. N-grams are a contiguous sequence of n items from a given sequence of text [117]. Natural language processing (NLP) is a field of study that bridges the gap between textual and structured data, allowing humans to interact using familiar natural language while enabling computer applications to process data effectively [118]. NLP analyzes unstructured text written or typed by people and translates these text into a structured representation, such as concepts of a controlled vocabulary [119]. For example, a NLP program might analyze a clinical note and extract disease concepts that can be mapped to the International Classification of Diseases version 10 (ICD-10).

The Unified Medical Language System (UMLS) is a set of files and software that brings together many health and biomedical vocabularies and standards to enable interoperability between terminologies and computer systems [120]. The UMLS contains unique concept identifiers (CUIs), such as myocardial infarction, which represents heart attacks, and semantic type identifiers (STYs), such as *Pharmacologic Substance*, which represents medications, that allow computer programs to map shared entities across vocabularies. Concepts and semantic types, such as those represented in the UMLS, can be used in NLP and machine learning approaches to text classification [121].

A number of different machine learning approaches have been applied to text classification activities with encouraging results. Naïve Bayes [19-22], logistic regression [24,25], and random forests [23] have demonstrated good performance, with accuracies over 0.95, in text classification tasks. Naïve Bayes is a probabilistic learning method, which uses Bayes theorem to determine the probability that a document will belong to a certain category [20]. Major shortcomings of the Naïve Bayes method for text classification include considering the presence and absence of terms instead of the frequency of terms, failing to recognize that fewer words in shorter documents are more important than in longer documents, and difficulty detecting interaction between terms [122]. Logistic regression uses a logistic function that measures the relationship between a category and the features within a document to determine the probability that a document belongs to a category [123]. Limitations of logistic regression are only using linear features or needing to change the equation for nonlinear features, overfitting from a distinctly large number of features f as compared to the number of observations n ($f >> n$), and lacking detection of complex nonlinear relationships [124]. Random forests are ensemble learning methods, which use different features to traverse decision trees to determine the probability that a document belongs to a category [125]. Random forests have advantages over logistic regression and Naïve Bayes by not requiring linear features or features that interact linearly and performing better even if the number of features is distinctly larger than the number of observations [126-129]. However, random forests do not always outperform other machine learning methods [130].

Text classification applications have been evaluated in the health care domain. Several studies have demonstrated the ability to classify unstructured text written by medical personnel. Haas et al. described a system that used unstructured text from chief complaints and triage notes for syndromic surveillance by determining if those documents belonged to three different syndromes (gastrointestinal, respiratory, and fever-rash) [131]. Another system classified mortality risk based on nursing notes in the intensive care unit (ICU) using elastic nets [132]. These studies demonstrate excellent accuracies with sensitivities and specificities over 0.90 and areas under the operator-receiver curve over 0.88.

Researchers have attempted to identify adverse drug reactions from consumer-generated text [133,134]. Yang et al. analyzed online discussion threads of drugs from Medhelp to discover adverse drug events (ADE) for clarithromycin, lansoprazole and fluvoxamine. Sarker et al. used Twitter, DailyStrength, and a publicly available ADE corpus to discover ADEs. They employed multiple techniques to determine the inputs to the different classifiers including NLP techniques (n-grams, as well as concepts and semantic types from UMLS), topic modeling, and semantic analysis. They demonstrated that using NLP and machine learning techniques improved automatic classification of social media text to determine ADEs by tuning certain parameters in their techniques. Huh et al. used text classification to determine whether discussion in WebMD's online diabetes community needed a moderator's attention using a machine learning classifier [135]. Although text classification has been successfully done for consumer-generated text from online forums, social media, and consumer generated text, this approach has not been applied to and evaluated in secure messages from patient portals.

Categorization Schemes for Consumer Health Information Needs

This section briefly reviews the relevant literature on categorization schemes for health information needs, with an emphasis on consumer needs. Although consumer health information needs are known to be different than those of healthcare providers in particular situations, there is overlap in need types. Thus, this section will briefly review taxonomies of providers' information needs before summarizing the literature about types of consumer health information needs.

Several researchers have examined the types of healthcare provider information needs [31,33,34,136]. Ely and Osheroff previously described two taxonomies of clinical questions asked by primary care doctors, one for question topics and one for generic questions [137,138]. The topics taxonomy was based on specialties, and modified from a system used to file journal articles in family practice [139]. The most common topics in this taxonomy were drug prescribing (19%), obstetrics and gynecology (9%), and adult infectious diseases (8%). The taxonomy of generic questions was based on methods used to classify Medline searches [140]. The most common themes in the generic questions were "What is the cause of symptom X?" "What is the dose of drug X" and "How should I manage disease or finding X?" The generic question taxonomy was modified to include four hierarchical levels. Their highest level consisted of five broad areas: diagnosis, treatment, management, epidemiology, and non-clinical questions, and the taxonomy included 64 quaternary categories. Inter-rater reliability in the application of this taxonomy to clinical questions by primary care doctors showed substantial agreement.

Classifying consumer health information needs is an evolving research area with existing taxonomies being incomplete or difficult to use [38,45,49,141-151]. Boot and Meijman investigated the ability of two taxonomies, the International Classification of Primary Care (ICPC-2), and the Taxonomy for Generic Clinical Questions (TGCQ) to classify health questions asked by the public to health care providers. They discovered missing items from the taxonomies and ill-defined information needs. Alzougool and colleagues proposed a disease specific taxonomy of information needs from a single caregiver of a diabetic child. They used a qualitative case study approach, looking at a 52-year-old woman and her three children, with her younger son being a type I diabetic. The mother was asked to keep activity diaries and had semi-structured interviews to discover deeper insight into her information needs. Neither of these studies created or evaluated a comprehensive taxonomy for consumer health information needs.

Our research team has developed a taxonomy of consumer health information needs and communication types shown in Figure 1. This taxonomy provides a comprehensive model of the semantic types of consumer health communications and has been employed on a diverse set of communications including patient journals, patient and caregiver interviews, and secure messages [152]. This taxonomy divides information needs and communications into five main communication types: clinical information, medical, logistical, social, and other. The taxonomy is described as containing communication types because it can be employed to categorize both consumer questions and the answers to these questions. *Clinical information* communication types include questions that require medical knowledge, such as those that could be answered by a medical textbook or consumer health information resource. This component of the model has been employed to structure online medical textbooks [136]. *Medical* communication types are about the delivery of medical care, such as the expression of a new symptom requiring management or the communication of a test result. *Logistical* communication types address pragmatic information, such as the location of a clinic or the copy of a medical record. The *social* communication types include personal exchanges such as an expression of gratitude or a complaint. The *other* communication types cover communications that are incomplete, unintelligible, or not captured in other parts of the taxonomy. Secure messages can contain more than one type of communication.

I. Clinical Information Needs or Communications

- A. Problems (Diseases or Observations)
  - 1. Definition
  - 2. Epidemiology
  - 3. Risk factors
  - 4. Etiology
  - 5. Pathogenesis/natural history
  - 6. Clinical presentation
  - 7. Differential diagnosis
  - 8. Related diagnoses
  - 9. Prognosis
- B. Management
  - 1. Goals/strategy
  - 2. Tests
  - 3. Interventions
  - 4. Sequence/timing
  - 5. Personnel/setting
- C. Interventions
  - 1. Definition
  - 2. Goals
  - 3. Mechanism of action
  - 4. Efficacy
  - 5. Indications
  - 6. Contraindications
  - 7. Preparation
  - 8.Technique/administration
  - 9. Monitoring
  - 10. Post-intervention care
  - 11. Advantages/benefits
  - 12. Costs/disadvantages
  - 13. Adverse effects
- D. Tests
  - 1. Definition
  - 2. Goals
  - 3. Physiologic basis
  - 4. Efficacy
  - 5. Indications
  - 6. Contraindications
  - 7. Preparation
  - 8.Technique/administration
  - 9. Interpretation
  - 10. Post-test care
  - 11. Advantages/benefits
  - 12. Costs /disadvantages
  - 13. Adverse effects

II. Medical Needs or Communications

- A. Appointments/scheduling
- B. Medical equipment
- C. Personnel/referrals
- D. Prescriptions
- E. Problems
- F. Follow-up
- G. Management
- H. Tests
- I. Interventions

III. Logistical Needs or Communications

- A. Contact information
- B. Facility/policies
- C. Insurance/billing
- D. Interventions
- E. Medical records
- F. Personal documentation
- G. Portal/health information technologies
- H. Tests

IV. Social Needs or Communications

- A. Acknowledgment
- B. Complaints
- C. Relationship communications
- D. Miscellaneous

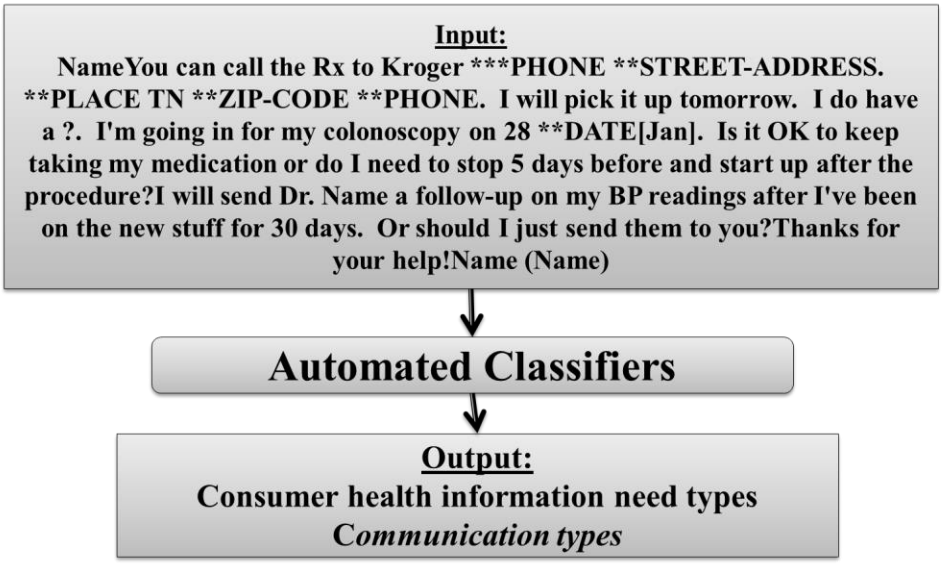
V. Other

Figure 1. The taxonomy of consumer health information needs or communications.

METHODS

Overview

This dissertation describes a project to develop and evaluate methods for automatically categorizing the content of patient-generated secure messages into types of consumer health information needs, or communication types, as some messages do not have information needs in them (Figure 2). We created a gold standard of patient-generated secure messages manually classified using the taxonomy of consumer health communication types presented in Figure 1. We then developed automatic classifiers to identify these communication types in secure messages. Classification used four methods: one rule based classifier and three machine learning techniques including Naïve Bayes, logistic regression, and random forests. We evaluated the classifiers on their ability to determine the presence of a single type of communication in a secure message, as well as their accuracy in capturing all types of communications in the message.



NameYou can call the Rx to Kroger ***PHONE **STREET-ADDRESS. **PLACE TN **ZIP-CODE **PHONE. I will pick it up tomorrow. I do have a ?. I'm going in for my colonoscopy on 28 **DATE[Jan]. Is it OK to keep taking my medication or do I need to stop 5 days before and start up after the procedure?I will send Dr. Name a follow-up on my BP readings after I've been on the new stuff for 30 days. Or should I just send them to you?Thanks for your help!Name (Name)

Figure 2: Example with the input (patient-generated secure message), automated classifiers, and output (communication types from our taxonomy described in the background)

Setting/Gold Standard (Outputs)

This study was conducted at Vanderbilt University Medical Center (VUMC), a private, non-profit institution that provides primary and regional referral care to over 500,000 patients annually. VUMC is located in middle Tennessee and serves both adults and children with over 900 inpatient beds, 50,000 inpatient admissions, and over 1 million outpatient visits per year. The VUMC Institutional Review Board approved this study as non-human subjects research.

VUMC launched a patient portal called My Health at Vanderbilt (MHAV) in 2005, with pediatric accounts added in 2007. The MHAV patient portal is available to any patient who receives medical care at VUMC. MHAV functions include access to parts of the EHR, secure messaging with healthcare providers, and delivery of personalized health information [153,154]. Adult patients can assign "delegate" MHAV users to access their health information and interact with providers on their behalf. For pediatric patients younger than 13 years, access to the pediatric patient's health information is done through a "surrogate" MHAV account by a parent or legal guardian. Patients of ages 13 through 17 years may also access and use their own MHAV account. MHAV currently has over 327,000 cumulatively registered users, including more than 21,000 pediatric accounts or relationships, with almost 300,000 logins per month.

The secure messaging function in MHAV extended VUMC provider-to-provider messaging capabilities within the EHR. MHAV users can only send messages to providers with whom they have a prior or scheduled appointment. Clinical teams manage MHAV messages, and an individual provider's messages may be answered by themselves or a staff member (e.g., nurse, administrative assistant, or allied health professional) [52,155]. Message threads are collections of messages exchanged between MHAV users and VUMC healthcare providers; a thread includes an initial message and all replies to that message by patients and providers. All MHAV message content is written to the EHR.

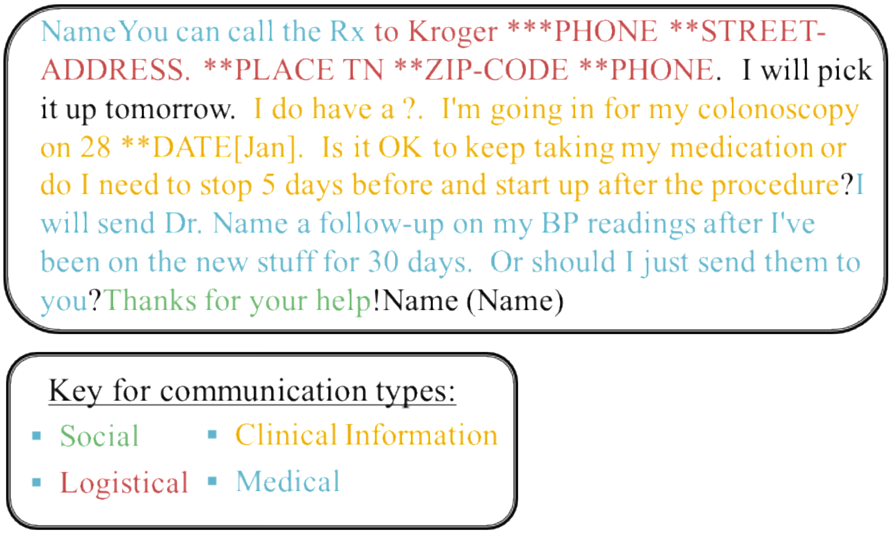
De-identified messages were extracted from the VUMC Synthetic Derivative (SD), a database containing a de-identified copy of all hospital medical records created for research purposes. Since all message content is saved to the EHR, we queried the SD for all message threads from 2005 to 2014. A message within a message thread that starts with a MHAV identification number is a patient-generated message. We used regular expressions to find messages that started with MHAV identification numbers to retrieve all patient-generated messages. Over 2.5 million patient-generated messages were present in the SD. We randomly selected 3,253 individual messages equally over the 10-year period for analysis. A gold standard was developed by manual annotation by two to three individuals who reviewed the content of all 3,253 messages and assigned all relevant communication types to each message. Annotators discussed discrepancies and achieved consensus to produce this gold standard.

Automated Classifiers

Automated classifiers predict which consumer health communication types are present within messages. Separate classifiers were built for individual communication types (Table 1). Figure 3 shows a sample secure message in which all four major communication types are present (clinical information, logistical, social, and medical). We built two sets of classifiers to identify communication types in secure messages: a rule-based classifier and three machine learning classifiers. The simple rule-based classifier, which is referred to as the basic classifier, determined if communication types were present in messages through regular expressions (Table 2 and Table 3). Words were chosen for the basic classifier based on expert knowledge of the research team. These words represent the most common expressions that would appear in a message that should be categorized to that communication type. If a word is present in a message then the basic classifier will output a 1 indicating the message belongs to that communication type, otherwise the classifier will output a 0. The machine learning classifiers included Naïve Bayes, logistic regression, and random forests classifiers. Each machine learning classifier outputs a probability between 0 and 1 based on whether the communication type is present in the message. To create the classifiers, we used python's scikit learn package [156]. We used Bernoulli Naïve Bayes with an alpha of 0.1 and random forests with 500 trees.

Table 1: Definitions of communication types

Major Type	Subtype	Definition
clinical information		Focuses on a desire to obtain knowledge about diseases, symptoms, management strategies, or other clinical topics
logistical		Concerns the pragmatic rather than medical issues
logistical	**contact information**	Includes the provision of or request for phone numbers, fax numbers, postal addresses, email addresses, or other methods of contact for any entity, including the patient
social		Related to social interactions or an interpersonal relationship that is not directly related to clinical information needs, medical needs, or logistical needs
social	**acknowledge-ment**	Expressions of gratitude or satisfaction or acknowledgement or agreement that is not directly related to medical care
medical		Expresses a desire for medical care or addresses delivery of care. This category is distinguished from *clinical information needs* with the former involving a request for or delivery of actual care and the latter expressing or fulfilling a desire for information or knowledge about a particular clinical topic
medical	**appointments**	Requests to schedule, change, or cancel appointments; confirmations of appointments; questions or concerns about a specific appointment; or requests for contact
medical	**prescriptions**	Requests for medication samples, refills of medications, or changes to existing prescriptions (e.g., change in dose amount or frequency)
medical	**problems**	Communications about a new, worsening, or changing symptom or condition
medical	**follow up**	Discussion about, confirmation of, or agreement upon a patient's care plan including updates on conditions that are being monitored when there is not a new, worsening, or changing problem that is being reported
medical	**tests**	Related to the need to undergo one or more tests, including the scheduling of that test, requests for test results, or reports of test results



NameYou can call the Rx to Kroger ***PHONE **STREET-ADDRESS. **PLACE TN **ZIP-CODE **PHONE. I will pick it up tomorrow. I do have a ?. I'm going in for my colonoscopy on 28 **DATE[Jan]. Is it OK to keep taking my medication or do I need to stop 5 days before and start up after the procedure?I will send Dr. Name a follow-up on my BP readings after I've been on the new stuff for 30 days. Or should I just send them to you?Thanks for your help!Name (Name)

Key for communication types:

- Social
- Clinical Information
- Logistical
- Medical

Figure 3: Example message labeled by communication types

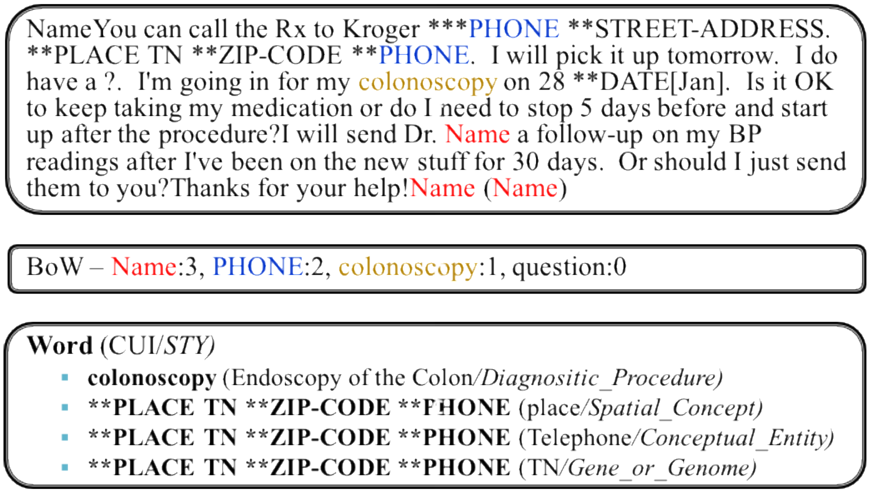
Table 2. Words used to determine if a message belongs to one of the major communication types for the basic classifier.

Communication Type	Words
clinical information	question, normal, medication, procedure
logistical	insurance, record, bill, cover
social	thank you very much, thank you so much, thanks very much, thanks so much, appreciate, your time
medical	refill, prescription, appointment, pain, hurt, lab, follow up, test, xray, ct, mri

Table 3: Words used to determine if a message belongs to one of the sub communication types for the basic classifier

Major Type	Subtype	Words
logistical	contact information	fax, phone, telephone, cell, address, street, email
social	acknowledge-ment	appreciate, time, very much
medical	appointments	call me, appointment, be seen, (Mon,Tues,Wed,Thurs,Fri)-day
medical	prescriptions	refill, prescription
medical	problems	pain, worse
medical	follow-up	better, follow up
medical	tests	lab, labs, ultrasound, CT, MRI, test

The features used for the machine learning classifiers consisted of bag of words (BoW), concept unique identifiers (CUIs), and semantic types (STYs). BoW is a representation in vector form of the number of times a word appears in a message. The KnowledgeMap Concept Indexer (KMCI) extracted concepts and semantic types from messages using NLP and the Unified Medical Language System (UMLS). KMCI is a tool designed at VUMC, and it has been validated for specific NLP tasks, such as discovering clinical concepts in clinical text in multiple studies, with high sensitivity and specificity [157-162]. The corpus of messages was represented as a matrix with each message corresponding to a row and the different features designated by the columns. For the BoW, the number of occurrences of each word in a message made up the cells in a row. CUIs and STYs were binary features, which were 0 or 1 depending on whether the CUI or STY was present in the message. Common stop words were removed from messages for the BoW representation. We used each of these features alone and in combination with each other, yielding seven feature sets used for the different classifiers (Figure 4). A different machine learning classifier was built for each major communication type and subtypes present in greater than 10% of messages, yielding a total of 231 machine learning classifiers (3 types of classifiers * 7 feature sets * 11 major and subtypes). The machine learning classifiers were trained and tested with a gold standard corpus of 3,253 documents using 10-fold cross validation. To determine what features were important for prediction, we recorded the variable importance for the random forest, measured by the decrease in impurity at the nodes using those variables. The variable importance from the complete feature set including all the BoW, CUIs, and STYs was used to determine the most important features.



NameYou can call the Rx to Kroger ***PHONE **STREET-ADDRESS. **PLACE TN **ZIP-CODE **PHONE. I will pick it up tomorrow. I do have a ?. I'm going in for my colonoscopy on 28 **DATE[Jan]. Is it OK to keep taking my medication or do I need to stop 5 days before and start up after the procedure?I will send Dr. Name a follow-up on my BP readings after I've been on the new stuff for 30 days. Or should I just send them to you?Thanks for your help!Name (Name)

BoW – Name:3, PHONE:2, colonoscopy:1, question:0

Word (CUI/*STY)*

- **colonoscopy** (Endoscopy of the Colon/*Diagnositic_Procedure)*
- ****PLACE TN **ZIP-CODE **PHONE** (place/*Spatial_Concept)*
- ****PLACE TN **ZIP-CODE **PHONE** (Telephone/*Conceptual_Entity)*
- ****PLACE TN **ZIP-CODE **PHONE** (TN/*Gene_or_Genome)*

Figure 4: Example message with selected features including bag of words (BoW), concepts (CUIs), and semantic types (STYs). In this message, the word "Name" appears three times, "PHONE" twice, "colonoscopy" once, and "question" zero times. The word "colonoscopy" gets mapped to the "Endoscopy of the Colon" CUI, which belongs to the "*Diagnostic_Procedure*" STY.

Evaluation and Statistical Analysis

Figure 2 depicts the classification process and evaluation metrics. The different classifiers process message content for each major communication type (clinical information, logistical, social, and medical), as well as subtypes that appeared in at least 10% of all messages. We determined the ability for classifiers to predict a single category of communication type in a message and to predict all of the major types or subtypes within a message using area under the receiver operator curves (AUCs) and the Jaccard Index. We evaluated the ability to predict a single major category in the machine learning classifiers with AUCs. We considered an AUC of 0.90-1 as excellent; 0.80-0.90 as good; 0.70-0.80 as fair; 0.60-0.70 as poor (an AUC of 0.50 is the same as random chance). For our basic classifier, we determined the ability to identify a communication type through the presence of several representative words found in those categories within the message.

We used the Jaccard index [163] to measure how well the classifiers are able to identify the set of communication types in a single message. The Jaccard index is a measure of the similarity between two sets:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

We chose the Jaccard Index for its ability to determine similarities between two sets of binary outcomes. It has similar performance in text classification tasks as other similarity metrics such as Pearson's correlation coefficient [164]. A Jaccard Index of 1 indicates that the sets A and B have the same elements, and a Jaccard Index of 0 means the sets A and B have no common elements. In our study, the gold standard annotated set represents A and the predicted set from the different classifiers represents B. We averaged the Jaccard indices for each message to give an overall estimation of the ability to predict the set of communication types across the entire corpus of messages.

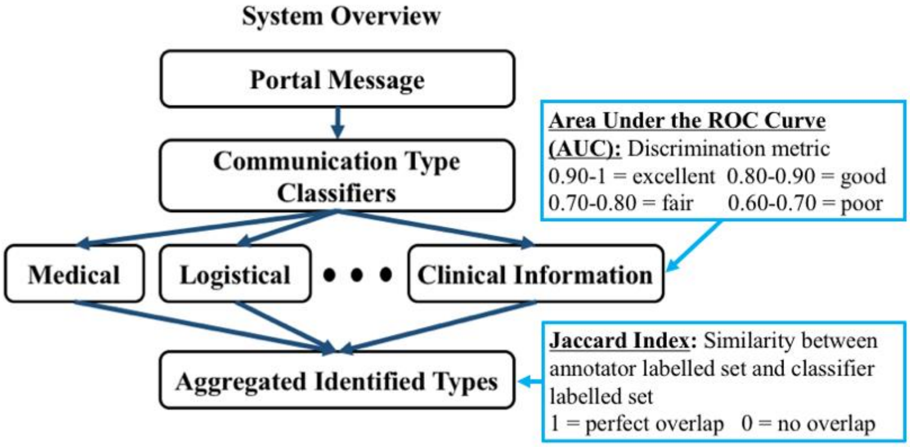

System Overview
Portal Message
Communication Type Classifiers
Medical
Logistical
• • •
Clinical Information
Aggregated Identified Types
Area Under the ROC Curve (AUC): Discrimination metric
0.90-1 = excellent 0.80-0.90 = good
0.70-0.80 = fair 0.60-0.70 = poor
Jaccard Index: Similarity between annotator labelled set and classifier labelled set
1 = perfect overlap 0 = no overlap

Figure 5. Classification process and evaluation metrics.

To determine the best Jaccard index for a classifier, a threshold probability that determines whether a message contains or does not contain a communication type must be set. Each classifier for a specific communication type generates a probability that the message contains the communication type. We set the threshold probability for whether a communication type was present in a message by calculating the average Jaccard for all threshold probabilities between zero and one at increments of 0.05. The threshold probability that yielded the maximum average Jaccard index was used across all major or sub communication types for the classifiers with the same set of features. For example, if a probability of 0.60 yielded the maximum Jaccard index for the BoW classifier, that Jaccard index was reported in the results.

RESULTS

Gold Standard

The gold standard contained 3,253 patient-generated messages, which were sent about 3,116 unique patients. The majority of the patients were female (1,937; 62.2%) and Caucasian (2,772; 89.0%). The median age of the patients was 50 years of age, with a range of 1 month to 112 years. The proportion of messages from each year was approximately 0.15% (Table 4).

Table 4: Distribution of messages across time

Year	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Message Corpus	2	31	73	155	264	475	744	920	578	11
Total Messages	820	21,219	52,760	97,831	184,469	288,518	467,169	592,189	363,220	7,142
% Message Corpus of Total Messages	0.24%	0.15%	0.14%	0.16%	0.14%	0.16%	0.16%	0.16%	0.16%	0.15%

The gold standard contained 2,351 medical communications, 922 social communications, 404 clinical information communications, and 806 logistical communications (Table 5). Of the messages, 114 (3.5%) had content categorized as other, which consisted of incomplete or incomprehensible text, or content not belonging to another category in the taxonomy. The number of major communication types in each of the remaining messages was one for 2,019 messages (62.1%), two for 912 messages (28.0%), three for 192 messages (5.9%), and four for 16 messages (0.5%). A co-occurrence matrix (Table 6) shows how often communication types coexisted in a single message (i.e., the value in each cell represents the percentage of messages that have communication type indicated in the row heading, which also have the communication type indicated in the column heading). In addition to determining the distribution of major communication types, we looked at the subtypes distributions. The following subtypes appeared in at least 10% of messages: logistical/contact information, social/acknowledgement, medical/appointments/scheduling, medical/prescriptions, medical/problems, medical/follow-up, and medical/tests (Table 7).

Table 5: Distribution of communication types among messages

Distribution of communication types among messages N (% of total messages)

Communication Type	# of Messages (%)
clinical information	404 (12.4%)
logistical	806 (24.8%)
social	922 (28.3%)
medical	2351 (72.3%)

Table 6: Communication type co-occurrence matrix.

Co-occurrence of communication types

Communication Type	clinical information	logistical	social	medical
clinical information	100%	17%	17%	78%
logistical	8%	100%	20%	69%
social	7%	18%	100%	45%
medical	13%	24%	18%	100%

Table 7: Most common communication subtypes.

Commonly Occurring Subtypes (% of total messages)

Major Type	Subtype	# of Messages (%)
logistical	contact information	802 (24.7%)
social	acknowledgement	482 (14.8%)
medical	appointments/scheduling	513 (15.8%)
medical	prescriptions	491 (15.1%)
medical	problems	492 (15.1%)
medical	follow-up	835 (25.7%)
medical	tests	429 (13.2%)

Predictive Performance for Individual Communication Types

The predictive performance for the classifiers is shown in Tables 7-17. AUCs for major communication types ranged from 0.604 for the Naive Bayes classifier in social types to 0.925 for the random forest machine learning classifier in logistical types (Tables 8-11, Figure 6).

Table 8: The area under the curves (AUCs) of the machine learning classifiers for clinical information types with each type of input: Bag of Words (BoW), unique concept identifiers (CUIs), and semantic types (STYs). The highest AUC is bolded for each classifier.

Feature Sets	**# of Features**	**Basic Classifier**		
Words	4	0.842 (0.836,0.848)		
		Naive Bayes	**Logistic Regression**	**Random Forest**
BoW	9,643	0.759 (0.735,0.783)	0.783 (0.759,0.808)	0.825 (0.807,0.844)
CUI	6,040	0.706 (0.665,0.747)	0.770 (0.753,0.787)	0.783 (0.763,0.804)
STY	200	0.652 (0.615,0.689)	0.772 (0.754,0.789)	0.765 (0.742,0.788)
BoW, CUI	15,683	**0.763 (0.734,0.793)**	0.788 (0.768,0.808)	0.827 (0.806,0.848)
BoW, STY	9,843	0.761 (0.737,0.784)	0.811 (0.789,0.832)	**0.830 (0.812,0.847)**
CUI, STY	6,240	0.714 (0.672,0.756)	0.795 (0.769,0.821)	0.803 (0.782,0.824)
BoW, CUI, STY	15,883	0.762 (0.732,0.793)	**0.806 (0.785,0.827)**	0.829 (0.810,0.848)

Table 9: The area under the curves (AUCs) of the machine learning classifiers for logistical types with each type of input: Bag of Words (BoW), unique concept identifiers (CUIs), and semantic types (STYs). The highest AUC is bolded for each classifier.

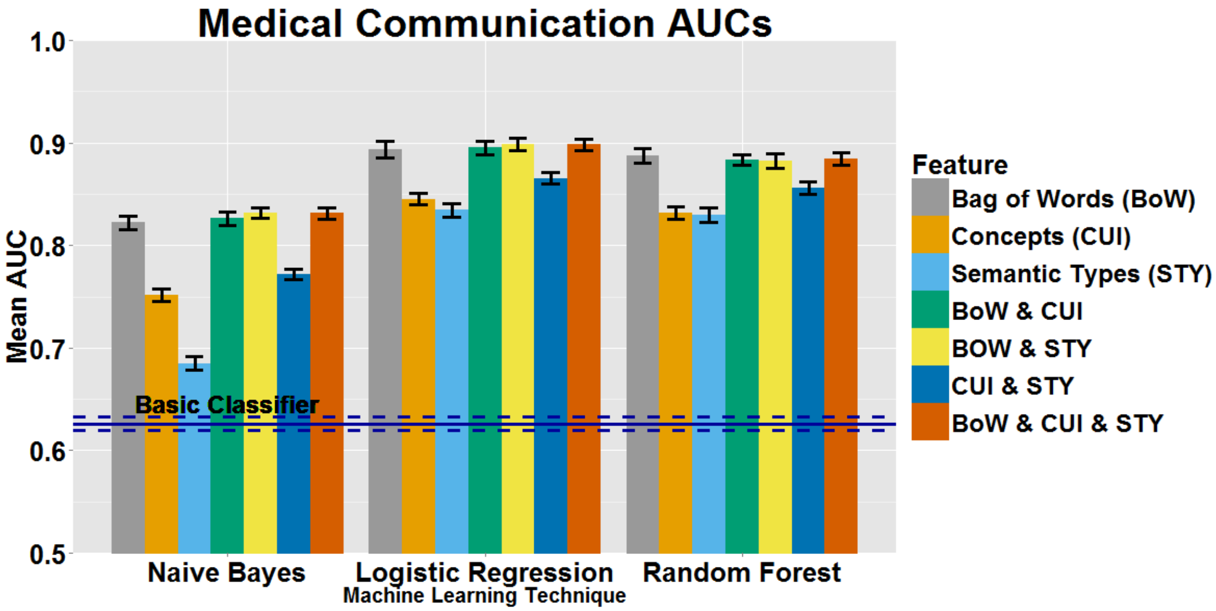
Feature Sets	**# of Features**	**Basic Classifier**		
Words	4	0.794 (0.788,0.801)		
		Naive Bayes	**Logistic Regression**	**Random Forest**
BoW	9,643	0.813 (0.796,0.831)	0.903 (0.892,0.915)	**0.925 (0.917,0.933)**
CUI	6,040	0.784 (0.770,0.799)	0.864 (0.840,0.888)	0.871 (0.853,0.890)
STY	200	0.718 (0.705,0.730)	0.796 (0.781,0.812)	0.799 (0.785,0.812)
BoW, CUI	15,683	0.826 (0.812,0.840)	0.903 (0.890,0.916)	**0.925 (0.916,0.934)**
BoW, STY	9,843	0.817 (0.800,0.834)	0.904 (0.894,0.915)	0.921 (0.912,0.929)
CUI, STY	6,240	0.795 (0.780,0.810)	0.867 (0.846,0.887)	0.869 (0.853,0.884)
BoW, CUI, STY	15,883	**0.828 (0.814,0.841)**	**0.905 (0.892,0.917)**	0.922 (0.913,0.930)

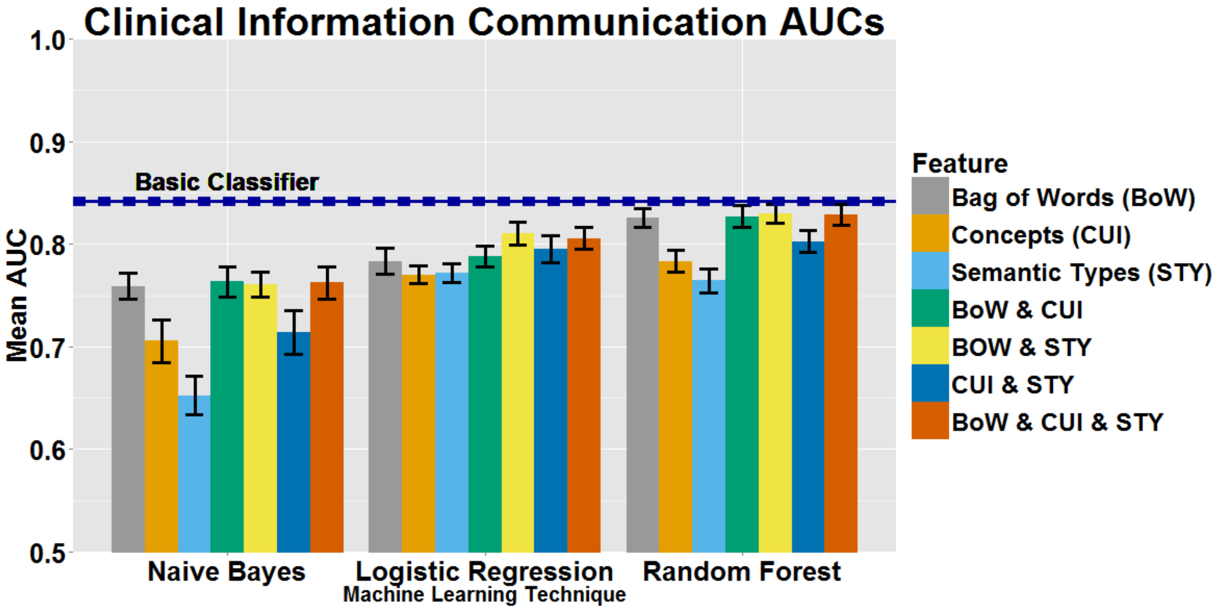
Table 10: The area under the curves (AUCs) of the machine learning classifiers for social types with each type of input: Bag of Words (BoW), unique concept identifiers (CUIs), and semantic types (STYs). The highest AUC is bolded for each classifier.

Feature Sets	# of Features	Basic Classifier		
Words	6	0.747 (0.742,0.753)		
		Naive Bayes	**Logistic Regression**	**Random Forest**
BoW	9,643	0.705 (0.674,0.737)	0.811 (0.793,0.828)	0.857 (0.841,0.874)
CUI	6,040	0.604 (0.588,0.620)	0.703 (0.680,0.726)	0.759 (0.744,0.775)
STY	200	0.620 (0.610,0.630)	0.661 (0.646,0.677)	0.715 (0.705,0.725)
BoW, CUI	15,683	0.700 (0.671,0.730)	0.813 (0.794,0.833)	0.870 (0.853,0.886)
BoW, STY	9,843	**0.719 (0.688,0.750)**	0.825 (0.809,0.840)	0.871 (0.859,0.883)
CUI, STY	6,240	0.625 (0.608,0.643)	0.735 (0.716,0.753)	0.780 (0.767,0.794)
BoW, CUI, STY	15,883	0.709 (0.679,0.738)	**0.826 (0.809,0.844)**	**0.875 (0.862,0.887)**

Table 11: The area under the curves (AUCs) of the machine learning classifiers for medical types with each type of input: Bag of Words (BoW), unique concept identifiers (CUIs), and semantic types (STYs). The highest AUC is bolded for each classifier.

Feature Sets	# of Features	Basic Classifier		
Words	11	0.626 (0.614,0.638)		
		Naive Bayes	**Logistic Regression**	**Random Forest**
BoW	9,643	0.822 (0.810,0.835)	0.894 (0.878,0.909)	**0.887 (0.873,0.902)**
CUI	6,040	0.752 (0.740,0.763)	0.845 (0.834,0.856)	0.832 (0.820,0.843)
STY	200	0.685 (0.672,0.698)	0.834 (0.822,0.847)	0.830 (0.815,0.844)
BoW, CUI	15,683	0.826 (0.814,0.839)	0.895 (0.883,0.908)	0.884 (0.873,0.894)
BoW, STY	9,843	**0.832 (0.821,0.842)**	**0.899 (0.887,0.911)**	0.883 (0.869,0.896)
CUI, STY	6,240	0.772 (0.762,0.782)	0.866 (0.854,0.877)	0.856 (0.844,0.869)
BoW, CUI, STY	15,883	0.831 (0.821,0.841)	0.898 (0.888,0.909)	0.884 (0.872,0.896)


Medical Communication AUCs
Mean AUC
1.0
0.9
0.8
0.7
0.6
0.5
Basic Classifier
Naive Bayes
Logistic Regression
Random Forest
Machine Learning Technique
Feature
Bag of Words (BoW)
Concepts (CUI)
Semantic Types (STY)
BoW & CUI
BOW & STY
CUI & STY
BoW & CUI & STY


Clinical Information Communication AUCs
Mean AUC
1.0
0.9
0.8
0.7
0.6
0.5
Basic Classifier
Naive Bayes
Logistic Regression
Random Forest
Machine Learning Technique
Feature
Bag of Words (BoW)
Concepts (CUI)
Semantic Types (STY)
BoW & CUI
BOW & STY
CUI & STY
BoW & CUI & STY

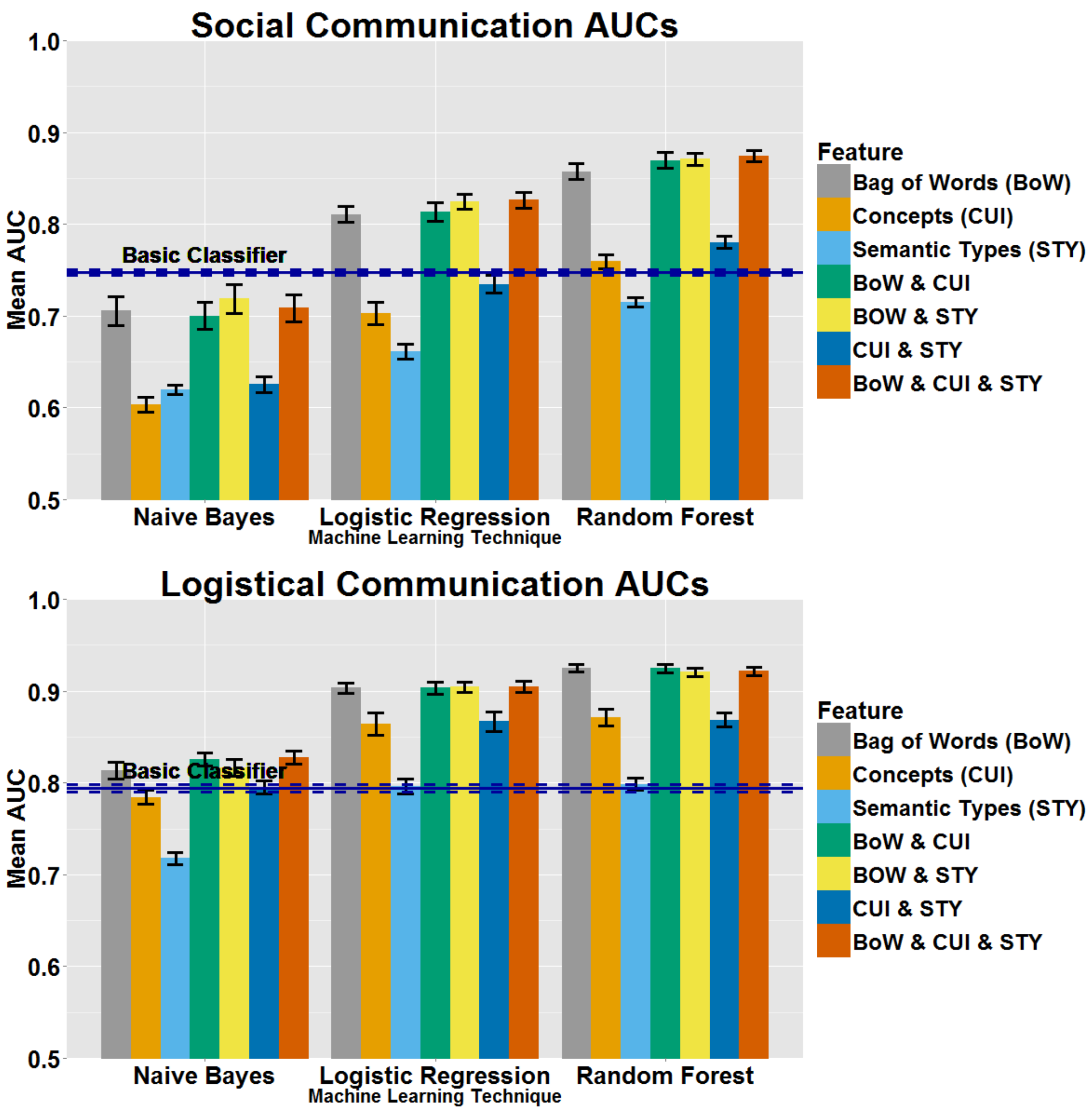


Figure 6: Area under the curve (AUC) of the different major communication types

For clinical information communication types, the classifier with the best performance characteristics was the basic classifier (Median: 0.842; 95% CI: 0.836,0.848), followed by the random forest classifier using BoW, STY (Median: 0.830; 95% CI: 0.812,0.847), logistic regression classifier using BoW, CUI, STY (Median: 0.806; 95% CI: 0.785,0.827), and Naïve Bayes classifier using BoW, CUI (Median: 0.763, 95% CI: 0.734,0.793).

In social communication types, the best performing classifier was the random forest classifier with BoW, CUI, STY (Median: 0.875; 95% CI: 0.862,0.887), followed by the logistic regression classifier using BoW, CUI, STY (Median: 0.826; 95% CI: 0.809,0.844), basic classifier (Median: 0.747; 95% CI: 0.742,0.753), and Naïve Bayes classifier using BoW, STY (Median: 0.719, 95% CI: 0.688,0.750).

For logistical communication types, the classifier with the highest AUC was the random forest classifiers using BoW, with or without CUI (Median: 0.925; 95% CI: 0.917,0.933), followed by the logistic regression classifier using BoW, CUI, STY (Median: 0.905; 95% CI: 0.892,0.917), Naïve Bayes classifier using BoW, CUI, STY (Median: 0.828; 95% CI: 0.814,0.841), and the basic classifier (Median: 0.794, 95% CI: 0.788,0.801).

For medical communication types, the classifier with the best performance characteristics was the logistic regression classifier with BoW, STY (Median: 0.899; 95% CI: 0.887,0.911), followed by the random forest classifier using BoW (Median: 0.887; 95% CI: 0.873,0.902), Naïve Bayes using BoW, STY (Median: 0.832; 95% CI: 0.821,0.842), and basic classifier (Median: 0.747, 95%CI: 0.742,0.753).

The AUCs for the subtypes ranged from 0.609 for the Naïve Bayes classifier in both medical/follow-up and social/acknowledgement to 0.963 for the random forest classifier in logistical/contact information (Tables 12-18).

Table 12: The area under the curves (AUCs) of the machine learning classifiers for social/acknowledgements subtypes with each type of input: Bag of Words (BoW), unique concept identifiers (CUIs), and semantic types (STYs). The highest AUC is bolded for each classifier.

Feature Sets	**# of Features**	**Basic Classifier**		
Words	3	0.770 (0.765,0.774)		
ML		**Naive Bayes**	**Logistic Regression**	**Random Forest**
BoW	9,643	0.734 (0.716,0.752)	0.841 (0.826,0.855)	0.877 (0.866,0.887)
CUI	6,040	0.609 (0.590,0.628)	0.726 (0.705,0.747)	0.768 (0.747,0.789)
STY	200	0.616 (0.599,0.633)	0.691 (0.663,0.720)	0.720 (0.697,0.743)
BoW, CUI	15,683	0.720 (0.702,0.737)	0.848 (0.831,0.864)	0.888 (0.875,0.901)
BoW, STY	9,843	**0.746 (0.728,0.764)**	0.856 (0.841,0.871)	0.888 (0.877,0.900)
CUI, STY	6,240	0.626 (0.609,0.642)	0.751 (0.734,0.769)	0.787 (0.777,0.796)
BoW, CUI, STY	15,883	0.727 (0.711,0.744)	**0.860 (0.845,0.875)**	**0.892 (0.883,0.900)**

Table 13: The area under the curves (AUCs) of the machine learning classifiers for logistical/contact information subtypes with each type of input: Bag of Words (BoW), unique concept identifiers (CUIs), and semantic types (STYs). The highest AUC is bolded for each classifier.

Feature Sets	**# of Features**	**Basic Classifier**		
Words	7	0.844 (0.839,0.850)		
ML		**Naive Bayes**	**Logistic Regression**	**Random Forest**
BoW	9,643	0.810 (0.789,0.831)	**0.948 (0.936,0.960)**	0.961 (0.949,0.972)
CUI	6,040	0.774 (0.750,0.799)	0.889 (0.880,0.898)	0.911 (0.899,0.922)
STY	200	0.738 (0.721,0.756)	0.829 (0.807,0.850)	0.821 (0.793,0.849)
BoW, CUI	15,683	**0.824 (0.802,0.846)**	0.944 (0.930,0.957)	**0.963 (0.954,0.972)**
BoW, STY	9,843	0.814 (0.795,0.834)	0.945 (0.931,0.959)	0.961 (0.949,0.973)
CUI, STY	6,240	0.791 (0.771,0.810)	0.886 (0.873,0.899)	0.910 (0.894,0.925)
BoW, CUI, STY	15,883	**0.824 (0.803,0.846)**	0.943 (0.930,0.956)	0.963 (0.953,0.972)

Table 14: The area under the curves (AUCs) of the machine learning classifiers for medical/appointments/scheduling subtypes with each type of input: Bag of Words (BoW), unique concept identifiers (CUIs), and semantic types (STYs). The highest AUC is bolded for each classifier.

Feature Sets	**# of Features**	**Basic Classifier**		
Words	8	0.855 (0.844,0.867)		
ML		**Naive Bayes**	**Logistic Regression**	**Random Forest**
BoW	9,643	0.796 (0.776,0.816)	**0.863 (0.846,0.880)**	0.875 (0.859,0.891)
CUI	6,040	0.714 (0.690,0.737)	0.800 (0.781,0.820)	0.797 (0.775,0.819)
STY	200	0.681 (0.664,0.699)	0.705 (0.683,0.727)	0.704 (0.687,0.722)
BoW, CUI	15,683	0.797 (0.777,0.817)	0.862 (0.847,0.877)	0.878 (0.861,0.895)
BoW, STY	9,843	**0.803 (0.783,0.822)**	0.855 (0.838,0.872)	**0.880 (0.864,0.895)**
CUI, STY	6,240	0.735 (0.709,0.762)	0.794 (0.772,0.816)	0.807 (0.787,0.828)
BoW, CUI, STY	15,883	0.800 (0.780,0.819)	0.855 (0.840,0.869)	0.878 (0.862,0.895)

Table 15: The area under the curves (AUCs) of the machine learning classifiers for medical/prescriptions subtypes with each type of input: Bag of Words (BoW), unique concept identifiers (CUIs), and semantic types (STYs). The highest AUC is bolded for each classifier.

Feature Sets	**# of Features**	**Basic Classifier**		
Words	2	0.936 (0.932,0.941)		
ML		**Naive Bayes**	**Logistic Regression**	**Random Forest**
BoW	9,643	0.896 (0.881,0.912)	0.943 (0.932,0.955)	0.961 (0.953,0.969)
CUI	6,040	0.866 (0.850,0.883)	0.907 (0.892,0.923)	0.915 (0.903,0.927)
STY	200	0.795 (0.767,0.824)	0.836 (0.804,0.868)	0.837 (0.815,0.859)
BoW, CUI	15,683	**0.902 (0.887,0.918)**	0.941 (0.932,0.951)	**0.962 (0.954,0.970)**
BoW, STY	9,843	0.895 (0.881,0.910)	**0.948 (0.938,0.958)**	0.961 (0.953,0.969)
CUI, STY	6,240	0.869 (0.855,0.883)	0.916 (0.900,0.931)	0.915 (0.899,0.930)
BoW, CUI, STY	15,883	0.901 (0.886,0.915)	0.944 (0.934,0.954)	**0.962 (0.954,0.970)**

Table 16: The area under the curves (AUCs) of the machine learning classifiers for medical/problems subtypes with each type of input: Bag of Words (BoW), unique concept identifiers (CUIs), and semantic types (STYs). The highest AUC is bolded for each classifier.

Feature Sets	**# of Features**	**Basic Classifier**		
Words	2	0.871 (0.864,0.878)		
ML		**Naive Bayes**	**Logistic Regression**	**Random Forest**
BoW	9,643	0.869 (0.850,0.889)	**0.904 (0.893,0.915)**	**0.934 (0.924,0.944)**
CUI	6,040	0.814 (0.787,0.840)	0.852 (0.837,0.866)	0.891 (0.874,0.908)
STY	200	0.772 (0.742,0.802)	0.859 (0.839,0.880)	0.872 (0.850,0.894)
BoW, CUI	15,683	0.863 (0.843,0.884)	0.902 (0.888,0.916)	0.932 (0.922,0.942)
BoW, STY	9,843	**0.870 (0.852,0.889)**	0.901 (0.889,0.914)	0.932 (0.922,0.942)
CUI, STY	6,240	0.824 (0.800,0.847)	0.876 (0.858,0.895)	0.903 (0.887,0.919)
BoW, CUI, STY	15,883	0.864 (0.844,0.884)	0.902 (0.886,0.917)	0.931 (0.920,0.942)

Table 17: The area under the curves (AUCs) of the machine learning classifiers for medical/follow-up subtypes with each type of input: Bag of Words (BoW), unique concept identifiers (CUIs), and semantic types (STYs). The highest AUC is bolded for each classifier.

Feature Sets	**# of Features**	**Basic Classifier**		
Words	2	0.750 (0.739,0.761)		
ML		**Naive Bayes**	**Logistic Regression**	**Random Forest**
BoW	9,643	0.701 (0.675,0.728)	0.752 (0.733,0.771)	0.785 (0.771,0.798)
CUI	6,040	0.659 (0.631,0.687)	0.711 (0.688,0.734)	0.744 (0.726,0.762)
STY	200	0.609 (0.578,0.639)	0.708 (0.689,0.728)	0.712 (0.697,0.727)
BoW, CUI	15,683	0.708 (0.681,0.734)	0.748 (0.732,0.764)	0.786 (0.773,0.800)
BoW, STY	9,843	0.703 (0.676,0.731)	**0.762 (0.741,0.783)**	**0.789 (0.777,0.800)**
CUI, STY	6,240	0.660 (0.631,0.689)	0.722 (0.700,0.744)	0.750 (0.738,0.762)
BoW, CUI, STY	15,883	**0.709 (0.682,0.735)**	0.759 (0.743,0.776)	0.785 (0.772,0.797)

Table 18: The area under the curves (AUCs) of the machine learning classifiers for medical/tests subtypes with each type of input: Bag of Words (BoW), unique concept identifiers (CUIs), and semantic types (STYs). The highest AUC is bolded for each classifier.

Feature Sets	**# of Features**	**Basic Classifier**		
Words	6	0.873 (0.869,0.877)		
ML		**Naive Bayes**	**Logistic Regression**	**Random Forest**
BoW	9,643	0.770 (0.746,0.793)	0.837 (0.819,0.855)	0.881 (0.865,0.897)
CUI	6,040	0.706 (0.686,0.726)	0.758 (0.735,0.780)	0.793 (0.771,0.815)
STY	200	0.721 (0.696,0.746)	0.761 (0.742,0.780)	0.750 (0.725,0.775)
BoW, CUI	15,683	0.777 (0.752,0.802)	0.837 (0.826,0.848)	0.881 (0.867,0.896)
BoW, STY	9,843	0.783 (0.761,0.806)	0.843 (0.827,0.859)	**0.885 (0.870,0.901)**
CUI, STY	6,240	0.735 (0.718,0.751)	0.793 (0.780,0.807)	0.816 (0.792,0.840)
BoW, CUI, STY	15,883	**0.786 (0.762,0.810)**	**0.845 (0.834,0.856)**	**0.885 (0.869,0.901)**

Predictive Performance for All Communication Types Within A Message

The Jaccard Indices for the major communication types ranged from 0.663 (95% CI: 0.657,0.669) for the basic classifier, to 0.861 (95% CI: 0.855,0.867) for logistic regression classifier using BoW and STY features (Table 19, Figure 7). The best random forest classifier's Jaccard index 95% CI overlapped with logistic regression, but both did not overlap with the basic classifier and best performing Naïve Bayes classifiers.

Table 19: Jaccard indices for the basic and machine learning classifiers with a single feature for the (A) major types and (B) subtypes

(A)

Classifier	**# Features**	**Basic Classifier**		
Words	24	0.674 (0.663,0.684)		
		Naive Bayes	**Logistic Regression**	**Random Forest**
BoW	9,643	0.794 (0.787,0.800)	0.857 (0.850,0.863)	0.858 (0.852,0.864)
CUI	6,040	0.767 (0.760,0.774)	0.816 (0.810,0.823)	0.819 (0.813,0.825)
STY	200	0.766 (0.759,0.773)	0.799 (0.792,0.805)	0.803 (0.796,0.811)
BoW, CUI	15,683	0.793 (0.786,0.800)	0.858 (0.851,0.864)	**0.859 (0.853,0.865)**
BoW, STY	9,843	**0.799 (0.792,0.805)**	**0.861 (0.855,0.867)**	0.857 (0.850,0.863)
CUI, STY	6,240	0.777 (0.770,0.783)	0.824 (0.817,0.831)	0.829 (0.823,0.836)
BoW, CUI, STY	15,883	0.795 (0.789,0.802)	**0.861 (0.855,0.867)**	0.858 (0.851,0.864)

(B)

Classifier	**# Features**	**Basic Classifier**		
Words	30	0.674 (0.663,0.684)		
		Naive Bayes	**Logistic Regression**	**Random Forest**
BoW	9,643	0.842 (0.837,0.846)	0.879 (0.875,0.883)	0.876 (0.872,0.880)
CUI	6,040	0.834 (0.830,0.839)	0.855 (0.851,0.859)	0.855 (0.851,0.859)
STY	200	0.833 (0.828,0.837)	0.836 (0.832,0.840)	0.837 (0.832,0.842)
BoW, CUI	15,683	0.836 (0.831,0.841)	0.881 (0.877,0.885)	0.879 (0.874,0.883)
BoW, STY	9,843	**0.844 (0.840,0.849)**	0.881 (0.877,0.886)	**0.880 (0.876,0.884)**
CUI, STY	6,240	0.837 (0.833,0.842)	0.856 (0.852,0.860)	0.861 (0.857,0.865)
BoW, CUI, STY	15,883	0.838 (0.833,0.842)	**0.882 (0.877,0.886)**	0.879 (0.875,0.883)

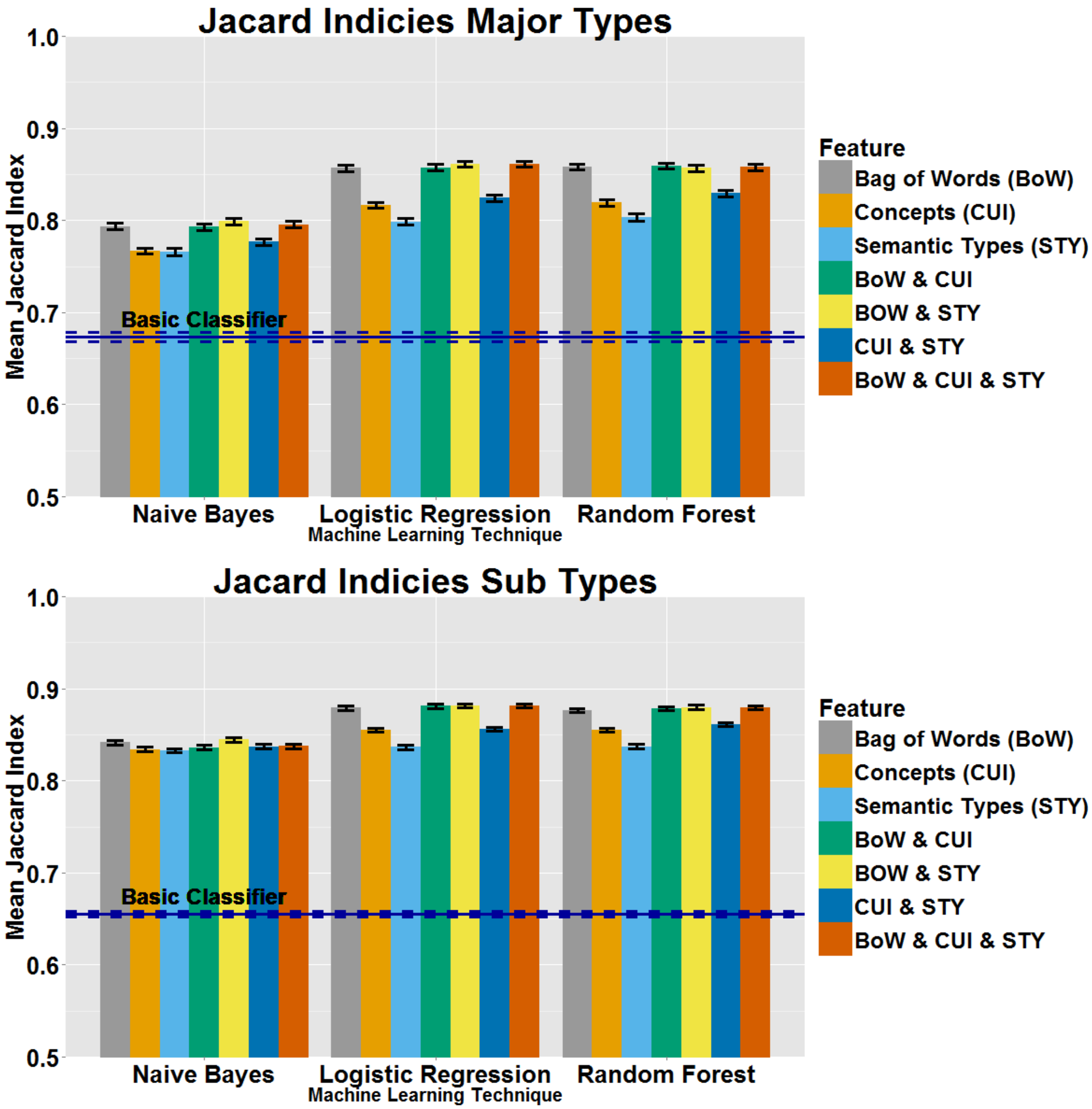


Figure 7: Bar charts of the Jaccard Indices of the different communication types

Features

There were a total of 9,643 words, 6,040 CUIs, and 200 STYs that were used by the machine learning classifiers. The basic classifier used between 2 and 11 words (Tables 2-3). Random forest variables of importance for each major communication type are shown in Table 20.

For clinical information communication types, the majority of the top five most important variables were semantic types. The semantic type "Finding" is identified by the words "lab work" and "latest test results", "Qualitative_Concept" STY is identified by the words "critical" and "high", and "Pharmacologic_Substance" is identified by the word "medication". The words ranked in the top ten included "question", "name", "red", "normal", and "dr".

Logistical communication types contained similar numbers of CUIs and words, with the concept "Telephone" being identified by the word "phone" and the concept "Insurances" identified by the word "insurance". "Conceptual_Entity" semantic types are identified by the words "phone", "doctor", and "fax". The words ranked in the top 5 included "phone", "insurance", and "please".

The four of the top five variables of importance for identifying Social communication were words rather than concepts or semantic types. The semantic type "Intellectual_Product", identified by the word "name", was the third most important variable. Most of the words involve common expressions of gratitude such as "thank", "thanks", and "much".

In medical communication types, the majority of important variables were STYs. "Temporal_Concept" is identified by the words "morning", "day", "evenings", and "Friday", "Idea_or_Concept" is identified by the words "appointment" and "refill", "Quantitative_Concept" has the abbreviation "mg" within it, and "Intellectual_Product" is identified by the word "prescription". The only word ranked in the top five was "name".

Table 20: Top 10 features for random forests. Green boxes contain concepts (CUIs), and yellow boxes represent semantic types (STYs). The white boxes represent un-stemmed words as they appear in the message.

Key	Word	Concept	Semantic Type

Importance Rank	Clinical Information	Logistical	Social	Medical
# 1	question	phone	thank	Temporal_Concept
# 2	Finding	Telephone	thanks	Idea_or_Concept
# 3	Qualitative_Concept	Conceptual_Entity	name	name
# 4	name	insurance	Intellectual_ Product	Quantitative_Concept
# 5	Pharmacologic_Substance	Insurances	much	Intellectual_Product
# 6	Statistical_mean	please	Idea_or_Concept	Qualitative_Concept
# 7	red	call	dr	Organic_Chemical__Pharmacologic_Substance
# 8	Functional_Concept	fax	Temporal_Concept	appointment
# 9	normal	address	you	Finding
# 10	dr	Manufactured_Object	help	dr

DISCUSSION

In this project, we developed and evaluated a set of methods to identify automatically the communication types in contents of secure messages sent through a widely deployed patient portal at an academic medical center. This research provides evidence that patient-generated secure messages can be automatically classified into communication type categories with good accuracy. As adoption of patient portals increases, automated techniques may be needed to assist in understanding and managing growing volumes of secure messages. Automated classification of patient-generated secure messages has several important potential applications. First, the ability to specify communication types in secure messages may aid in connecting patients to needed resources and in triaging secure messages. In addition, automated classifiers could support consumer health informatics research to understand the nature of communications and types of care delivered within patient portals. Such work could lead to better resources for commonly expressed information needs and might support compensation for care delivered online.

Gold Standard

In this study, the majority of messages contained medical communications. Medical communications typically involve requests for or the delivery of medical care such as a patient's communication of new or worsening symptoms or medications or a physician's plans for managing them. This finding supports prior research [26,97] demonstrating that patient portals are used as a conduit for delivering healthcare. The classification systems used in the Haun et al. and North et al. studies generally included subtypes found in our taxonomy of consumer health communications. North et al manually classified 323 messages, demonstrating 37% of messages were medication related, 23% were symptom related, 20% were test related, 7% had to do with medical questions, 6% were acknowledgements, and 9% had greater than one issue. In our population, similarly medical communication types were the majority of messages; however, our population had more messages that contained acknowledgements and multiple issues. Our study also described logistical communication types, such as directions and insurance information, which to our knowledge have not been described previously in the literature. Therefore, this work extends the previously reported types of communications identified in secure messages [26,97,98,110].

Performance for single communication types

Our classifiers' performance varied by communication type. The best performing classifiers had good predicting power with AUCs over 0.82 for all major communication types and every subtype except for the medical/follow-up subtype. The best machine learning classifiers also outperformed the basic classifiers in all subtypes and all major communication types except for clinical information types. The basic classifier may have outperformed the machine learning classifiers for two reasons. First, clinical information communication types were the least likely to be present and most often occurred with another communication types. Therefore, it would be more difficult for machine learning algorithms to detect these because there is too much noise in the messages to detect the signal of the few most predictive words. Second, there may be a few common terms that occur in those messages that never occur in other ones. These findings demonstrate that identification of most major communication types and subtypes requires a sophisticated machine learning technique as the expression of information categories among messages are more complex. Logistic regression and random forests performed similarly across most communication types, and both outperformed Naïve Bayes and basic classifiers. Random forest had the highest performance for every subtype and most major types.

The good performance of the classifiers for major communication types is important as messages can potentially be triaged to the most appropriate person or resource based on the type of communication. For example, messages with medical communications are most likely need response from a healthcare provider, such as a nurse or physician, as these messages typically have to do with delivery of medical care. Messages with logistical communications may be better addressed by an administrative assistant or office manager as they often involve pragmatic questions rather than medical knowledge. The machine learning classifiers performed best in identifying the following subtypes of communications:

logistical/contact information, medical/prescriptions, and medical/problems. Medical/problems, present in 15% of messages, are descriptions of new or worsening symptoms. Medical/prescription, also present in 15% of messages, have to do with medication management. These communication types have significant implications in the delivery of care for a patient. Triaging and determining symptoms from a patient is an activity traditionally done in outpatient clinic visit. When done through secure messaging, financial incentives are lacking [165]. Several proposals for compensating online care have been developed, including billing codes for transition of care and telehealth services [166], but few payers reimburse for this type of care. By being able to identify the volume of messages containing medical communications, hospital administrators may be able to get a better estimate of the volume of care delivered through this modality and clinicians may be able to lobby for better models for compensation.

More complex machine learning methods with NLP performed better in identifying the communication types in the more complex parts of the taxonomy. Automated text classification systems have categorized documents based on specific tasks, such as classification of syndromes [131], mortality risk in the ICU [132], adverse drug events [133,134], and the need for moderation in WebMD's diabetes community [135]. These systems showed good predictive power of machine learning techniques, such as Naïve Bayes, support vector machines, and elastic nets (which are based on regression models). Some of these studies also used NLP techniques including concepts and semantic types from UMLS to improve classification. Contrary to prior studies [133], the performance improvement with a combination of BoW, CUIs, and STYs in this study was marginal. The 95% CI for all classifiers overlapped with the 95% CI of a single feature (BoW, CUI, or STY) classifier. The gains in AUCs when adding NLP CUIs and STYs over just BoW were modest in most cases, with the best AUC improvement being 0.023. For some communication types, the BoW alone outperformed all other classifiers. The CUI or STY classifiers never outperformed the other classifiers. These findings suggest that the words within the document predict communication types well, and that adding NLP techniques do not add much.

There are several possible reasons for these observations. NLP tools may incorrectly identify text due to misspellings and undefined abbreviations. Homonyms can be difficult for NLP tools to differentiate. *Left* can indicate laterality such as pain in the *left* leg, or an action such as the patient *left* the hospital. *Growth* can be an abnormal physiologic process, the *growth* of a tumor, or a normal one, the *growth* of a child. Clinical messages can be very compact and omit inferred information, "a mass" may implicitly mean "a mass on a breast". A lack of standardized structure for patient-generated messages can cause misidentification by NLP tools. Periods may not be used to demarcate the end of a sentence and sentences may have missing punctuation. Patient-generated text is likely to have less formal biomedical content and thus may have fewer identifiable UMLS concepts than formal medical texts. Higher order NLP methods, such as negation, also likely have little impact on content type. NLP is only going to do as well as the underlying tools, and these tools are based on controlled terminologies inside UMLS. Many of the component terminologies have been proven most useful in specific contexts. For example, SNOMED-CT performs well when looking at clinical documents [167]. RxNorm has good coverage for discovering drug names in documents [168]. The consumer health vocabulary (CHV) was designed to discover consumer language in documents [169-171]. CHV is open, allowing anyone to add concepts to the vocabulary, but representative portal users or consumer health informatics researchers may not be contributing to this vocabulary. One review described studies looking at the language used by patients [172], and concluded that the current standardized vocabularies may not be sufficient. This suggests that the non-CHV terminologies for this project may not be optimal for NLP of patient-generated messages sent through patient portals.

Performance for all communication types in a single message

The logistic regression classifier was able to identify all of the communication types in a single message significantly better than the Naïve Bayes and basic classifiers when evaluated using the Jaccard index. The random forest classifier's Jaccard index was only slightly lower than logistic regression. Random forests had the highest AUCs for determining each single major communication type except for the medical type, where logistic regression had the highest AUC. Since medical communication types occurred most often, the Jaccard index for logistic regression may have been higher than random forests because of their superior predictive power for the most common type of communication seen in secure messages. The best Jaccard Index was observed for a classifier using the BoW and STY feature set, with or without CUIs. The BoW alone classifier performed marginally worse. As each classifier performed better for different communication types, a hybrid of communication types classifiers might best determine all of the communication types in a single message.

Features

The features that were most important for the full BoW, CUI, and STY random forest classifiers varied by communication type. Classifiers to identify clinical information needs employed more STYs than words or CUIs. The STYs corresponded to words that could appear more often in these types of messages ("lab work", "test results", "critical", and "high"), as these messages frequently involved questions about what test results mean, which would be classified in clinical information/tests/interpretation in our taxonomy. Examining the most important variables to predict clinical information communication types led to several interesting findings. The word "red" ranked seventh in importance. In our system, "red" designates an abnormal laboratory result. In this case, patients were asking for test interpretations of their abnormal laboratory results. "Statistical_mean" was a CUI ranked sixth in importance in clinical information communication types. This concept was identified inappropriately by the words "mean" or "means" which might appear in the phrase "what does this test mean?" and gets mapped to the concept of "Statistical_mean". Although interpreted incorrectly, this mapping reflects a characteristic language within these messages, which is acceptable for identification, but not good for interpretation.

For classifying logistical communication types, the word "phone" was the most important variable. This word is the de-identified version of phone numbers, so in logistical communication types, phone numbers or the word phone was most important. This important word was an artifact of the de-identification process when messages are written to the Synthetic Derivative. For classifying logistical communication types, CUIs were often important variables. These CUIs were mapped to words in logistical communication types including "phone" and "insurance”. These words would be present in logistical/contact information or logistical/insurance/billing subtypes in our taxonomy. However, these CUIs may also help discover variations of words within logistical communications, such as “phone”, “telephone”, and "phone number" that are all identified by the CUI "Telephone". These words could all be important for logistical communication types, however may not have ranked in the top 10. The semantic type that ranked third in logistical communication types was identified by the words “phone” and “fax”.

The most important variables for classifying social communication types were words. Most of these words describe expressions of gratitude for care, including “thank" and "much” present in "thank you very much", and the word “thanks”, which would be present in messages classified by social/acknowledgement in our taxonomy. All names of people including doctor names and patient names were replaced by the de-identified word “name.” Names were important in our classifiers for social communication types as they likely designated the person they were thanking. Names were also important in classifying communication types, which occur in communications in which patients are asking questions of or talking about certain people, such as their doctors.

In medical communication types, STYs appeared most often among predictive features. STYs are likely important because of the diversity of words within medical communications. For example, Temporal_Concept is identified by certain times of day, such as "morning" or "evening”, as well as the day of the week, and Idea_Or_Concept contains the words “refills” and “appointment”. Both of these STYs are mapped to words that can appear in communications of the types medical/prescriptions or medical/appointments. There can also be variability of words within a single subtype, such as medical/problems. In this subtype, the words communicating many different symptoms such as pain or nausea in a message would be mapped to the STY Finding.

The difference in the variables of importance across communication types reflects the diversity of language within types of communications. For the secure message classification task, NLP outputs like CUIs and STYs did not add much to the BoW approach although they were all present in the top ten most important variables. There were several instances of inaccurate identification of CUIs and STYs, such as the "Statistical_Mean" CUI being identified by "what does this test mean?" and the word "red" present in questions about abnormal laboratory values. The de-identified data set used as input for the classifier may have normalized some of the variability in words such as "name" for names, and "phone" for telephone numbers. This normalization was important for some of our classifiers and these de-identified mappings may have helped in the performance of our classifiers, which would not be present in fully identified data.

Limitations

The performance of these classifiers may be limited by several factors. First, patient-generated messages may include misspellings, which may adversely affect communication type identification. These messages may also contain abbreviations not commonly used, as well as different abbreviations for the same word. Building a vocabulary that maps abbreviations to the same word could solve this problem. Second, automatic derivation of meaning from patient-generated texts using computers is an ongoing challenge. Our classifiers may not be able to understand the meaning of the text, and therefore cannot determine the category of communication type. Third, we used UMLS for our classifiers; however, this collection of terminologies may not capture different ways of expressing concepts.

This study was conducted at a single institution with a locally developed patient portal. Although the communication types seen in these messages are common communication types that have been seen in other papers about patient portal messaging [26,97], our results may be limited by the unique policies and procedures developed for MHAV. Second, this study trained the classifiers using a small data set. Therefore, all communication types and the full breadth of their expression may not be adequately represented. Third, this study has older data including the years 2005-2014, and some of the content of messages may have become antiquated based on secular trends. Fourth, these data were from a de-identified data set the unique features described above. The classifiers may not have similar performance characteristics on fully identified messages. Finally, our machine learning models are built on large feature sets that could lead to overfitting. However, random forests performed just as well if not better than most other methods and are robust to prevent overfitting [129], [126-128]. Our ongoing research projects will evaluate these methodologies on larger data sets, and explore the performance of automated classifiers across clinical specialties where vocabularies and distributions of communication types may differ.

Future Directions

This research project demonstrates promise in identifying the types of communication in patient-generated secure messages and suggests several directions for future work. First, while our performance in the classifiers was good, they could be improved. We could accomplish this through examination of where failures occur in secure message classification as well as continuing to validate our taxonomy in different specialties. Using other features in the classifiers could also improve performance, such as n-grams, topic modeling, or word2vec [173], which takes a text corpus and places words into a vector space where groups of words would be in similar places in the vector space.

These automated classifiers were tested on a small data set from randomly selected secure messages. Different clinical specialties may use messaging for different reasons, and the communications are likely to cover substantially different topics. Exploring messages across different clinical specialties may help determine differences in how clinical populations express themselves in messaging, as well as the amount and types of care that are delivered through messaging. Sociodemographic groups may express themselves differently in secure messages. Classifying messages in different racial or ethnic groups or individuals with different levels of health literacy may require different approaches. Finally, these classifiers may be applicable to other consumer-generated text such as that found on social networking sites or online forums. Determining the strengths and weaknesses of classifiers in these various populations and text sources is an area of interest to our research group.

Automatic classification of communication types in patient portals has several potentially important applications. First, it could allow triaging of patient-generated messages to different members of the health care team or information resources; therefore, automatic classification might enable routing of these messages appropriately without human intervention. Second, classifiers might be used to detect levels of urgency in messages. North et al. showed that occasionally patients will send potentially life-threatening symptoms through patient portals [97]. Utilizing automated classifiers to detect urgent messages could prevent adverse events by prioritizing responses or alerting a provider through an alternative means of communication. Finally, these classifiers could be used to determine communication types that result in patient care being delivered. Financial models for reimbursement for this type of care are lacking, and by exploring the nature of patient-generated secure messages, we may be able to help develop models for reimbursement of care that is being delivered.

REFERENCES

1. Cline, R. J., & Haynes, K. M. (2001). Consumer health information seeking on the Internet: the state of the art. *Health education research, 16*(6), 671-692.
2. Wald, H. S., Dube, C. E., & Anthony, D. C. (2007). Untangling the Web—The impact of Internet use on health care and the physician–patient relationship. *Patient Education and Counseling, 68*(3), 218-224.
3. McMullan, M. (2006). Patients using the Internet to obtain health information: how this affects the patient–health professional relationship. *Patient Education and Counseling, 63*(1), 24-28.
4. Fox, S. (2011). Health Topics. Retrieved 6/14/2015, 2015, from http://www.pewinternet.org/2011/02/01/health-topics-2/
5. Belkin, N. J., Oddy, R. N., & Brooks, H. M. (1982). ASK for information retrieval: Part I. Background and theory. *Journal of documentation, 38*(2), 61-71.
6. Calabretta, N. (2002). Consumer-driven, patient-centered health care in the age of electronic information. *Journal of the Medical Library Association: JMLA, 90*(1), 32-37.
7. Van de Belt, T. H., Engelen, L. J. L. P. G., Berben, S. A. A., Teerenstra, S., Samsom, M., & Schoonhoven, L. (2013). Internet and social media for health-related information and communication in health care: preferences of the Dutch general population. *Journal of Medical Internet Research, 15*(10). doi: 10.2196/jmir.2607
8. Harris Interactive, H. P. (2011). The Growing Influence and Use Of Health Care Information Obtained Online. Retrieved 6/14/2015, 2015, from http://www.harrisinteractive.com/NewsRoom/HarrisPolls/tabid/447/ctl/ReadCustom%20Default/mid/1508/ArticleId/863/Default.aspx
9. Hanberger, L., Ludvigsson, J., & Nordfeldt, S. (2013). Use of a web 2.0 portal to improve education and communication in young patients with families: randomized controlled trial. *Journal of Medical Internet Research, 15*(8). doi: 10.2196/jmir.2425
10. Shapochka, A. (2012). Providers Turn to Portals to Meet Patient Demand, Meaningful Use | Journal of AHIMA.
11. Tang, P. C., & Lansky, D. (2005). The missing link: bridging the patient-provider health information gap. *Health Affairs (Project Hope), 24*(5), 1290-1295. doi: 10.1377/hlthaff.24.5.1290
12. Koonce, T. Y., Giuse, D. A., Beauregard, J. M., & Giuse, N. B. (2007). Toward a more informed patient: bridging health care information through an interactive communication portal. *Journal of the Medical Library Association: JMLA, 95*(1), 77-81.
13. Bussey-Smith, K. L., & Rossen, R. D. (2007). A systematic review of randomized control trials evaluating the effectiveness of interactive computerized asthma patient education programs. *Ann Allergy Asthma Immunol, 98*(6), 507-516; quiz 516, 566. doi: 10.1016/S1081-1206(10)60727-2
14. Jackson, C. L., Bolen, S., Brancati, F. L., Batts-Turner, M. L., & Gary, T. L. (2006). A systematic review of interactive computer-assisted technology in diabetes care. Interactive information technology in diabetes care. *Journal of General Internal Medicine, 21*(2), 105-110. doi: 10.1111/j.1525-1497.2005.00310.x
15. Neve, M., Morgan, P. J., Jones, P. R., & Collins, C. E. (2010). Effectiveness of web-based interventions in achieving weight loss and weight loss maintenance in overweight and obese adults: a systematic review with meta-analysis. *Obesity reviews: an official journal of the International Association for the Study of Obesity, 11*(4), 306-321. doi: 10.1111/j.1467-789X.2009.00646.x

16. Vandelanotte, C., Spathonis, K. M., Eakin, E. G., & Owen, N. (2007). Website-delivered physical activity interventions a review of the literature. *American Journal of Preventive Medicine, 33*(1), 54-64. doi: 10.1016/j.amepre.2007.02.041

17. Walters, S. T., Wright, J. A., & Shegog, R. (2006). A review of computer and Internet-based interventions for smoking behavior. *Addict Behav, 31*(2), 264-277. doi: 10.1016/j.addbeh.2005.05.002

18. Aphinyanaphongs, Y., Tsamardinos, I., Statnikov, A., Hardin, D., & Aliferis, C. F. (2005). Text categorization models for high-quality article retrieval in internal medicine. *Journal of the American Medical Informatics Association: JAMIA, 12*(2), 207-216. doi: 10.1197/jamia.M1641

19. Friedman, J. H. (1997). On bias, variance, 0/1—loss, and the curse-of-dimensionality. *Data mining and knowledge discovery, 1*(1), 55-77.

20. Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian network classifiers. *Machine learning, 29*(2-3), 131-163.

21. McCallum, A., & Nigam, K. (1998). *A comparison of event models for naive bayes text classification.* Paper presented at the AAAI-98 workshop on learning for text categorization.

22. Sahami, M. (1996). *Learning Limited Dependence Bayesian Classifiers.* Paper presented at the KDD.

23. Fette, I., Sadeh, N., & Tomasic, A. (2007). *Learning to detect phishing emails.* Paper presented at the Proceedings of the 16th international conference on World Wide Web.

24. Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., & Lin, C.-J. (2008). LIBLINEAR: A library for large linear classification. *the Journal of machine Learning research, 9*, 1871-1874.

25. Genkin, A., Lewis, D. D., & Madigan, D. (2007). Large-scale Bayesian logistic regression for text categorization. *Technometrics, 49*(3), 291-304.

26. Haun, J. N., Lind, J. D., Shimada, S. L., Martin, T. L., Gosline, R. M., Antinori, N., Stewart, M., & Simon, S. R. (2014). Evaluating user experiences of the secure messaging tool on the Veterans Affairs' patient portal system. *Journal of Medical Internet Research, 16*(3). doi: 10.2196/jmir.2976

27. North, F., Crane, S. J., Stroebel, R. J., Cha, S. S., Edell, E. S., & Tulledge-Scheitel, S. M. (2013). Patient-generated secure messages and eVisits on a patient portal: are patients at risk? *Journal of the American Medical Informatics Association: JAMIA, 20*(6), 1143-1149. doi: 10.1136/amiajnl-2012-001208

28. Graham, M. J., Currie, L. M., Allen, M., Bakken, S., Patel, V., & Cimino, J. J. (2003). Characterizing information needs and cognitive processes during CIS use. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 852.

29. Currie, L. M., Graham, M., Allen, M., Bakken, S., Patel, V., & Cimino, J. J. (2003). Clinical information needs in context: an observational study of clinicians while using a clinical information system. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 190-194.

30. Collins, S., Bakken, S., Cimino, J. J., & Currie, L. M. (2007). Information needs related to antibiotic prescribing while using CPOE. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 916.

31. Allen, M., Currie, L. M., Graham, M., Bakken, S., Patel, V. L., & Cimino, J. J. (2003). The classification of clinicians' information needs while using a clinical information system. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 26-30.

32. Hyun, S., Currie, L. M., Hodorowski, J. K., Joyce, M. P., Lee, N. J., Velez, O., & Bakken, S. (2008). Nurses' use and perceptions of usefulness of National Cancer Institute's tobacco-related Cancer Information Service (CIS) resources. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 985.

33. Schnall, R., Cimino, J. J., Currie, L. M., & Bakken, S. (2011). Information needs of case managers caring for persons living with HIV. *Journal of the American Medical Informatics Association: JAMIA, 18*(3), 305-308. doi: 10.1136/jamia.2010.006668
34. Covell, D. G., Uman, G. C., & Manning, P. R. (1985). Information needs in office practice: are they being met? *Ann Intern Med, 103*(4), 596-599.
35. Forsythe, D. E., Buchanan, B. G., Osheroff, J. A., & Miller, R. A. (1992). Expanding the concept of medical information: an observational study of physicians' information needs. *Comput Biomed Res, 25*(2), 181-200.
36. Liu, F., Antieau, L. D., & Yu, H. (2011). Toward automated consumer question answering: automatically separating consumer questions from professional questions in the healthcare domain. *Journal of Biomedical Informatics, 44*(6), 1032-1038. doi: 10.1016/j.jbi.2011.08.008
37. Bar-Tal, Y., Barnoy, S., & Zisser, B. (2005). Whose informational needs are considered? A comparison between cancer patients and their spouses' perceptions of their own and their partners' knowledge and informational needs. *Soc Sci Med, 60*(7), 1459-1465. doi: 10.1016/j.socscimed.2004.08.003
38. Palisano, R. J., Almarsi, N., Chiarello, L. A., Orlin, M. N., Bagley, A., & Maggs, J. (2010). Family needs of parents of children and youth with cerebral palsy. *Child: Care, Health and Development, 36*(1), 85-92. doi: 10.1111/j.1365-2214.2009.01030.x
39. Washington, K. T., Meadows, S. E., Elliott, S. G., & Koopman, R. J. (2011). Information needs of informal caregivers of older adults with chronic health conditions. *Patient Education and Counseling, 83*(1), 37-44. doi: 10.1016/j.pec.2010.04.017
40. Gustafson, D. H., Hawkins, R., Boberg, E., Pingree, S., Serlin, R. E., Graziano, F., & Chan, C. L. (1999). Impact of a patient-centered, computer-based health information/support system. *American Journal of Preventive Medicine, 16*(1), 1-9.
41. Hailey, D., Roine, R., & Ohinmaa, A. (2002). Systematic review of evidence for the benefits of telemedicine. *Journal of Telemedicine and Telecare, 8 Suppl 1*, 1-30.
42. Martinez, A., Everss, E., Rojo-Alvarez, J. L., Figal, D. P., & Garcia-Alberola, A. (2006). A systematic review of the literature on home monitoring for patients with heart failure. *Journal of Telemedicine and Telecare, 12*(5), 234-241. doi: 10.1258/135763306777889109
43. Louis, A. A., Turner, T., Gretton, M., Baksh, A., & Cleland, J. G. (2003). A systematic review of telemonitoring for the management of heart failure. *Eur J Heart Fail, 5*(5), 583-590.
44. Jennett, P. A., Affleck Hall, L., Hailey, D., Ohinmaa, A., Anderson, C., Thomas, R., Young, B., Lorenzetti, D., & Scott, R. E. (2003). The socio-economic impact of telehealth: a systematic review. *Journal of Telemedicine and Telecare, 9*(6), 311-320. doi: 10.1258/135763303771005207
45. Phillips, S. A., & Zorn, M. J. (1994). Assessing consumer health information needs in a community hospital. *Bulletin of the Medical Library Association, 82*(3), 288-293.
46. Jones, R. (2000). Developments in consumer health informatics in the next decade. *Health Libraries Review, 17*(1), 26-31.
47. Cobb, N. K. (2010). Online consumer search strategies for smoking-cessation information. *American Journal of Preventive Medicine, 38*(3 Suppl), S429-432. doi: 10.1016/j.amepre.2009.12.001
48. Mercado-Martínez, F. J., & Urias-Vázquez, J. E. (2014). [Hispanic American kidney patients in the age of online social networks: content analysis of postings, 2010 - 2012]. *Revista Panamericana De Salud Pública = Pan American Journal of Public Health, 35*(5-6), 392-398.
49. Abdulla, S., Vielhaber, S., Machts, J., Heinze, H.-J., Dengler, R., & Petri, S. (2014). Information needs and information-seeking preferences of ALS patients and their carers. *Amyotrophic Lateral Sclerosis & Frontotemporal Degeneration*, 1-8. doi: 10.3109/21678421.2014.932385

50. Katz, S. J., Moyer, C. A., Cox, D. T., & Stern, D. T. (2003). Effect of a triage-based E-mail system on clinic resource use and patient and physician satisfaction in primary care: a randomized controlled trial. *Journal of General Internal Medicine, 18*(9), 736-744.

51. Roter, D. L., Larson, S., Sands, D. Z., Ford, D. E., & Houston, T. (2008). Can e-mail messages between patients and physicians be patient-centered? *Health Communication, 23*(1), 80-86. doi: 10.1080/10410230701807295

52. White, C. B., Moyer, C. A., Stern, D. T., & Katz, S. J. (2004). A content analysis of e-mail communication between patients and their providers: patients get the message. *Journal of the American Medical Informatics Association: JAMIA, 11*(4), 260-267. doi: 10.1197/jamia.M1445

53. Zickuhr, K. (2013). Who's Not Online and Why. Retrieved 6/5/2015, 2015, from http://www.pewinternet.org/2013/09/25/whos-not-online-and-why/

54. McClung, H. J., Murray, R. D., & Heitlinger, L. A. (1998). The Internet as a source for current patient information. *Pediatrics, 101*(6), E2.

55. WebMd. (2015). WebMD - Better information. Better health. Retrieved 6/14/2015, 2015, from http://www.webmd.com/default.htm

56. MedlinePlus. (2015). MedlinePlus - Health Information from the National Library of Medicine. Retrieved 6/14/2015, 2015, from http://www.ncbi.nlm.nih.gov/pubmed/

57. UpToDate. (2015). UpToDate. Retrieved 6/14/2015, 2015, from http://www.uptodate.com/home

58. Mayo Clinic. (2015). Mayo Clinic. Retrieved 6/14/2015, 2015, from http://www.mayoclinic.org/

59. Burns, L., Bradley, E., & Weiner, B. (2011). *Shortell and Kaluzny's Healthcare Management: Organization Design and Behavior*: Cengage Learning.

60. Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication, 13*(1), 210-230.

61. Bensley, R. J., & Brookins-Fisher, J. (2003). *Community health education methods: A practical guide*: Jones & Bartlett Learning.

62. International, M. (1994). MedHelp - Health community, health information, medical questions, and medical apps.

63. DailyStrength, I. (2006). Online Support Groups and Forums at DailyStrength.

64. Weitzman, E. R., Kelemen, S., & Mandl, K. D. (2011). Surveillance of an Online Social Network to Assess Population-level Diabetes Health Status and Healthcare Quality. *Online J Public Health Inform, 3*(3). doi: 10.5210/ojphi.v3i3.3797

65. Amir, M., Sampson, B. P., Endly, D., Tamai, J. M., Henley, J., Brewer, A. C., Dunn, J. H., Dunnick, C. A., & Dellavalle, R. P. (2014). Social networking sites: emerging and essential tools for communication in dermatology. *JAMA dermatology, 150*(1), 56-60. doi: 10.1001/jamadermatol.2013.6340

66. Ho, Y.-X., O'Connor, B. H., & Mulvaney, S. A. (2014). Features of Online Health Communities for Adolescents With Type 1 Diabetes. *Western journal of nursing research*. doi: 10.1177/0193945913520414

67. Dunn, A. G., Leask, J., Zhou, X., Mandl, K. D., & Coiera, E. (2015). Associations Between Exposure to and Expression of Negative Opinions About Human Papillomavirus Vaccines on Social Media: An Observational Study. *Journal of Medical Internet Research, 17*(6).

68. Househ, M., Borycki, E., & Kushniruk, A. (2014). Empowering patients through social media: the benefits and challenges. *Health informatics journal, 20*(1), 50-58.

69. Seltzer, E., Jean, N., Kramer-Golinkoff, E., Asch, D., & Merchant, R. (2015). The content of social media's shared images about Ebola: a retrospective study. *Public health*.

70. Seymour, B., Getman, R., Saraf, A., Zhang, L. H., & Kalenderian, E. (2015). When Advocacy Obscures Accuracy Online: Digital Pandemics of Public Health Misinformation Through an Antifluoride Case Study. *Journal Information, 105*(3).

71. Patient portal - Wikipedia, the free encyclopedia.

72. Archer, N., Fevrier-Thomas, U., Lokker, C., McKibbon, K. A., & Straus, S. E. (2011). Personal health records: a scoping review. *Journal of the American Medical Informatics Association: JAMIA, 18*(4), 515-522. doi: 10.1136/amiajnl-2011-000105

73. HealthIT.gov. (2014). What is a patient portal? | FAQs | Providers & Professionals | HealthIT.gov.

74. Tang, P. C., Ash, J. S., Bates, D. W., Overhage, J. M., & Sands, D. Z. (2006). Personal health records: definitions, benefits, and strategies for overcoming barriers to adoption. *Journal of the American Medical Informatics Association: JAMIA, 13*(2), 121-126. doi: 10.1197/jamia.M2025

75. Weingart, S. N., Rind, D., Tofias, Z., & Sands, D. Z. (2006). Who uses the patient internet portal? The PatientSite experience. *Journal of the American Medical Informatics Association: JAMIA, 13*(1), 91-95. doi: 10.1197/jamia.M1833

76. Nazi, K. M., & Woods, S. S. (2008). MyHealtheVet PHR: a description of users and patient portal use. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 1182.

77. Kaelber, D., & Pan, E. C. (2008). The value of personal health record (PHR) systems. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 343-347.

78. Detmer, D., Bloomrosen, M., Raymond, B., & Tang, P. (2008). Integrated personal health records: transformative tools for consumer-centric care. *BMC Med Inform Decis Mak, 8*, 45. doi: 10.1186/1472-6947-8-45

79. Ralston, J. D., Hereford, J., & Carrell, D. (2006). Use and satisfaction of a patient Web portal with a shared medical record between patients and providers. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 1070.

80. Goldzweig, C. L., Orshansky, G., Paige, N. M., Towfigh, A. A., Haggstrom, D. A., Miake-Lye, I., Beroes, J. M., & Shekelle, P. G. (2013). Electronic patient portals: evidence on health outcomes, satisfaction, efficiency, and attitudes: a systematic review. *Annals of Internal Medicine, 159*(10), 677-687. doi: 10.7326/0003-4819-159-10-201311190-00006

81. Baer, D. (2011). Patient-physician e-mail communication: the kaiser permanente experience. *Journal of Oncology Practice / American Society of Clinical Oncology, 7*(4), 230-233. doi: 10.1200/JOP.2011.000323

82. Chen, C., Garrido, T., Chock, D., Okawa, G., & Liang, L. (2009). The Kaiser Permanente Electronic Health Record: transforming and streamlining modalities of care. *Health Affairs (Project Hope), 28*(2), 323-333. doi: 10.1377/hlthaff.28.2.323

83. Cimino, J. J., Patel, V. L., & Kushniruk, A. W. (2002). The patient clinical information system (PatCIS): technical solutions for and experience with giving patients access to their electronic medical records. *International Journal of Medical Informatics, 68*(1-3), 113-127.

84. Earnest, M. A., Ross, S. E., Wittevrongel, L., Moore, L. A., & Lin, C.-T. (2004). Use of a patient-accessible electronic medical record in a practice for congestive heart failure: patient and physician experiences. *Journal of the American Medical Informatics Association: JAMIA, 11*(5), 410-417. doi: 10.1197/jamia.M1479

85. Goldzweig, C. L., Towfigh, A. A., Paige, N. M., Orshansky, G., Haggstrom, D. A., Beroes, J. M., Miake-Lye, I., & Shekelle, P. G. (2012). *Systematic Review: Secure Messaging Between Providers and Patients, and Patients' Access to Their Own Medical Record: Evidence on Health Outcomes, Satisfaction, Efficiency and Attitudes*. Washington (DC): Department of Veterans Affairs (US).

86. Grant, R. W., Wald, J. S., Poon, E. G., Schnipper, J. L., Gandhi, T. K., Volk, L. A., & Middleton, B. (2006). Design and implementation of a web-based patient portal linked to an ambulatory care

electronic health record: patient gateway for diabetes collaborative care. *Diabetes technology & therapeutics, 8*(5), 576-586.

87. Grant, R. W., Wald, J. S., Schnipper, J. L., Gandhi, T. K., Poon, E. G., Orav, E. J., Williams, D. H., Volk, L. A., & Middleton, B. (2008). Practice-linked online personal health records for type 2 diabetes mellitus: a randomized controlled trial. *Archives of Internal Medicine, 168*(16), 1776-1782. doi: 10.1001/archinte.168.16.1776

88. Green, B. B., Cook, A. J., Ralston, J. D., Fishman, P. A., Catz, S. L., Carlson, J., Carrell, D., Tyll, L., Larson, E. B., & Thompson, R. S. (2008). Effectiveness of home blood pressure monitoring, Web communication, and pharmacist care on hypertension control: a randomized controlled trial. *JAMA, 299*(24), 2857-2867. doi: 10.1001/jama.299.24.2857

89. Liederman, E. M., Lee, J. C., Baquero, V. H., & Seites, P. G. (2005). Patient-physician web messaging. The impact on message volume and satisfaction. *Journal of General Internal Medicine, 20*(1), 52-57. doi: 10.1111/j.1525-1497.2005.40009.x

90. Carrell, D., & Ralston, J. D. (2006, 2006). *Variation in adoption rates of a patient web portal with a shared medical record by age, gender, and morbidity level*.

91. Collmann, J., & Cooper, T. (2007). Breaching the security of the Kaiser Permanente Internet patient portal: the organizational foundations of information security. *Journal of the American Medical Informatics Association, 14*(2), 239-243.

92. Osborn, C. Y., Mayberry, L. S., Wallston, K. A., Johnson, K. B., & Elasy, T. A. (2013). Understanding patient portal use: implications for medication management. *Journal of Medical Internet Research, 15*(7). doi: 10.2196/jmir.2589

93. Schnipper, J. L., Gandhi, T. K., Wald, J. S., Grant, R. W., Poon, E. G., Volk, L. A., Businger, A., Siteman, E., Buckel, L., & Middleton, B. (2008). Design and implementation of a web-based patient portal linked to an electronic health record designed to improve medication safety: the Patient Gateway medications module. *Informatics in primary care, 16*(2), 147-155.

94. Nazi, K. M., & Woods, S. S. (2007, 2007). *MyHealtheVet PHR: a description of users and patient portal use*.

95. Epstein, J. N., Langberg, J. M., Lichtenstein, P. K., Kolb, R., Altaye, M., & Simon, J. O. (2011). Use of an Internet portal to improve community-based pediatric ADHD care: a cluster randomized trial. *Pediatrics, 128*(5), e1201-1208. doi: 10.1542/peds.2011-0872

96. Ketterer, T., West, D. W., Sanders, V. P., Hossain, J., Kondo, M. C., & Sharif, I. (2013). Correlates of patient portal enrollment and activation in primary care pediatrics. *Academic pediatrics, 13*(3), 264-271. doi: 10.1016/j.acap.2013.02.002

97. North, F., Crane, S. J., Chaudhry, R., Ebbert, J. O., Ytterberg, K., Tulledge-Scheitel, S. M., & Stroebel, R. J. (2014). Impact of patient portal secure messages and electronic visits on adult primary care office visits. *Telemedicine journal and e-health: the official journal of the American Telemedicine Association, 20*(3), 192-198. doi: 10.1089/tmj.2013.0097

98. Ross, S. E., Moore, L. A., Earnest, M. A., Wittevrongel, L., & Lin, C.-T. (2004). Providing a web-based online medical record with electronic communication capabilities to patients with congestive heart failure: randomized trial. *Journal of Medical Internet Research, 6*(2). doi: 10.2196/jmir.6.2.e12

99. Agrawal, A., & Mayo-Smith, M. F. (2004). Adherence to computerized clinical reminders in a large healthcare delivery network. *Studies in Health Technology and Informatics, 107*(Pt 1), 111-114.

100. Hess, R., Bryce, C. L., Paone, S., Fischer, G., McTigue, K. M., Olshansky, E., Zickmund, S., Fitzgerald, K., & Siminerio, L. (2007). Exploring challenges and potentials of personal health records in diabetes self-management: implementation and initial assessment. *Telemedicine*

journal and e-health: the official journal of the American Telemedicine Association, 13(5), 509-517. doi: 10.1089/tmj.2006.0089

101. Ma, C., Warren, J., Phillips, P., & Stanek, J. (2006). Empowering patients with essential information and communication support in the context of diabetes. *International Journal of Medical Informatics, 75*(8), 577-596. doi: 10.1016/j.ijmedinf.2005.09.001

102. Nordqvist, C., Hanberger, L., Timpka, T., & Nordfeldt, S. (2009). Health professionals' attitudes towards using a Web 2.0 portal for child and adolescent diabetes care: qualitative study. *Journal of Medical Internet Research, 11*(2). doi: 10.2196/jmir.1152

103. Quinn, C. C., Clough, S. S., Minor, J. M., Lender, D., Okafor, M. C., & Gruber-Baldini, A. (2008). WellDoc mobile diabetes management randomized controlled trial: change in clinical and behavioral outcomes and patient and physician satisfaction. *Diabetes technology & therapeutics, 10*(3), 160-168. doi: 10.1089/dia.2008.0283

104. Bryce, C. L., Zickmund, S., Hess, R., McTigue, K. M., Olshansky, E., Fitzgerald, K., & Fischer, G. (2008). Value versus user fees: perspectives of patients before and after using a web-based portal for management of diabetes. *Telemedicine journal and e-health: the official journal of the American Telemedicine Association, 14*(10), 1035-1043. doi: 10.1089/tmj.2008.0005

105. Tang, C. H., Li, C. C., Chang, G. H., & Chang, P. (2003). Implementing a personalized portal combined with workflow management tools used in diabetes care. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*.

106. Rosenbloom, S. T., Daniels, T. L., Talbot, T. R., McClain, T., Hennes, R., Stenner, S., Muse, S., Jirjis, J., & Purcell Jackson, G. (2012). Triaging patients at risk of influenza using a patient portal. *Journal of the American Medical Informatics Association: JAMIA, 19*(4), 549-554. doi: 10.1136/amiajnl-2011-000382

107. Ralston, J. D., Hirsch, I. B., Hoath, J., Mullen, M., Cheadle, A., & Goldberg, H. I. (2009). Web-based collaborative care for type 2 diabetes: a pilot randomized trial. *Diabetes care, 32*(2), 234-239. doi: 10.2337/dc08-1220

108. Weingart, S. N., Rind, D., Tofias, Z., & Sands, D. Z. (2006). Who uses the patient internet portal? The PatientSite experience. *Journal of the American Medical Informatics Association: JAMIA, 13*(1), 91-95. doi: 10.1197/jamia.M1833

109. Cronin, R. M., Davis, S. E., Shenson, J. A., Chen, Q., Rosenbloom, S. T., & Jackson, G. P. (2015). Growth of Secure Messaging Through a Patient Portal as a Form of Outpatient Interaction across Clinical Specialties. *Applied Clinical Informatics, 6*(2), 288-304. doi: 10.4338/ACI-2014-12-RA-0117

110. Barnett, T. E., Chumbler, N. R., Vogel, W. B., Beyth, R. J., Qin, H., & Kobb, R. (2006). The effectiveness of a care coordination home telehealth program for veterans with diabetes mellitus: a 2-year follow-up. *The American Journal of Managed Care, 12*(8), 467-474.

111. Stiles, R. A., Deppen, S. A., Figaro, M. K., Gregg, W. M., Jirjis, J. N., Rothman, R. L., Johnston, P. E., Miller, R. A., Dittus, R. S., & Speroff, T. (2007). Behind-the-scenes of patient-centered care: content analysis of electronic messaging among primary care clinic providers and staff. *Medical care, 45*(12), 1205-1209. doi: 10.1097/MLR.0b013e318148490c

112. Cohen, A. M., & Hersh, W. R. (2005). A survey of current work in biomedical text mining. *Briefings in bioinformatics, 6*(1), 57-71.

113. Mohri, M., Talwalkar, A., & Rostamizadeh, A. (2012). *Foundations of Machine Learning*. Cambridge, MA: MIT Press.

114. Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval* (Vol. 1): Cambridge university press Cambridge.

115. Blei, D. M., & Lafferty, J. D. (2009). Topic models. *Text mining: classification, clustering, and applications, 10*(71), 1-25.

116. Blei, D. M., & Lafferty, J. D. (2009). Topic models. *Text mining: classification, clustering, and applications, 10*(71), 71-93.
117. Broder, A. Z., Glassman, S. C., Manasse, M. S., & Zweig, G. (1997). Syntactic clustering of the web. *Computer Networks and ISDN Systems, 29*(8), 1157-1166.
118. Shortliffe, E. H., & Cimino, J. J. (2006). *Biomedical informatics : computer applications in health care and biomedicine* (3rd ed.). New York, NY: Springer.
119. Meystre, S. (2007). Electronic patient records: some answers to the data representation and reuse challenges. Findings from the section on Patient Records. *Yearb Med Inform*, 47-49.
120. Humphreys, B. L., Lindberg, D. A., Schoolman, H. M., & Barnett, G. O. (1998). The Unified Medical Language System An Informatics Research Collaboration. *Journal of the American Medical Informatics Association, 5*(1), 1-11.
121. Chen, Y., Wrenn, J., Xu, H., Spickard III, A., Habermann, R., Powers, J., & Denny, J. C. (2014). *Automated Assessment of Medical Students' Clinical Exposures according to AAMC Geriatric Competencies.* Paper presented at the AMIA Annual Symposium Proceedings.
122. Lewis, D. D. (1998). Naive (Bayes) at forty: The independence assumption in information retrieval *Machine learning: ECML-98* (pp. 4-15): Springer.
123. Bishop, C. M. (2006). *Pattern recognition and machine learning* (Vol. 1): springer New York.
124. Tu, J. V. (1996). Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *Journal of clinical epidemiology, 49*(11), 1225-1231.
125. Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5-32.
126. Casanova, R., Saldana, S., Chew, E. Y., Danis, R. P., Greven, C. M., & Ambrosius, W. T. (2014). Application of random forests methods to diabetic retinopathy classification analyses. *PloS One, 9*(6). doi: 10.1371/journal.pone.0098587
127. Liu, Y., Traskin, M., Lorch, S. A., George, E. I., & Small, D. (2014). Ensemble of trees approaches to risk adjustment for evaluating a hospital's performance. *Health Care Management Science*. doi: 10.1007/s10729-014-9272-4
128. Sowa, J.-P., Heider, D., Bechmann, L. P., Gerken, G., Hoffmann, D., & Canbay, A. (2013). Novel algorithm for non-invasive assessment of fibrosis in NAFLD. *PloS One, 8*(4). doi: 10.1371/journal.pone.0062439
129. Grömping, U. (2009). Variable importance assessment in regression: linear regression versus random forest. *The American Statistician, 63*(4).
130. Cronin, R. M., VanHouten, J. P., Siew, E. D., Eden, S. K., Fihn, T. S., Nielson, C. D., Peterson, J. F., Baker, C. R., Ikizler, T. A., Speroff, T., & Matheny, M. E. (2015). National veterans health administration inpatient risk stratification models for hospital-acquired acute kidney injury. *Journal of the American Medical Informatics Association: JAMIA*. doi: 10.1093/jamia/ocv051
131. Haas, S. W., Travers, D., Waller, A., Mahalingam, D., Crouch, J., Schwartz, T. A., & Mostafa, J. (2014). Emergency Medical Text Classifier: New system improves processing and classification of triage notes. *Online J Public Health Inform, 6*(2), e178. doi: 10.5210/ojphi.v6i2.5469
132. Marafino, B. J., John Boscardin, W., & Adams Dudley, R. (2015). Efficient and sparse feature selection for biomedical text classification via the elastic net: Application to ICU risk stratification from nursing notes. *Journal of Biomedical Informatics, 54*, 114-120. doi: 10.1016/j.jbi.2015.02.003
133. Sarker, A., & Gonzalez, G. (2015). Portable automatic text classification for adverse drug reaction detection via multi-corpus training. *Journal of Biomedical Informatics, 53*, 196-207. doi: 10.1016/j.jbi.2014.11.002
134. Yang, M., Kiang, M., & Shang, W. (2015). Filtering big data from social media - Building an early warning system for adverse drug reactions. *Journal of Biomedical Informatics, 54*, 230-240. doi: 10.1016/j.jbi.2015.01.011

135. Huh, J., Yetisgen-Yildiz, M., & Pratt, W. (2013). Text classification for assisting moderators in online health communities. *Journal of Biomedical Informatics, 46*(6), 998-1005. doi: 10.1016/j.jbi.2013.08.011

136. Purcell, G. P. (2003). Surgical textbooks: past, present, and future. *Ann Surg, 238*(6 Suppl), S34-41.

137. Ely, J. W., Osheroff, J. A., Ebell, M. H., Bergus, G. R., Levy, B. T., Chambliss, M. L., & Evans, E. R. (1999). Analysis of questions asked by family doctors regarding patient care. *BMJ (Clinical research ed.), 319*(7206), 358-361.

138. Ely, J. W., Osheroff, J. A., Gorman, P. N., Ebell, M. H., Chambliss, M. L., Pifer, E. A., & Stavri, P. Z. (2000). A taxonomy of generic clinical questions: classification study. *BMJ (Clinical research ed.), 321*(7258), 429-432.

139. Reynolds, R. D. (1995). A family practice article filing system. *J Fam Pract, 41*(6), 583-590.

140. Wilson, S. R., Starr-Schneidkraut, N., & Cooper, M. D. (1989). Use of the critical incident technique to evaluate the impact of MEDLINE: American Institutes for Research, Palo Alto, CA (USA).

141. Archibald, M. M., & Scott, S. D. (2014). The information needs of North American parents of children with asthma: a state-of-the-science review of the literature. *Journal of Pediatric Health Care: Official Publication of National Association of Pediatric Nurse Associates & Practitioners, 28*(1), 5-13.e12. doi: 10.1016/j.pedhc.2012.07.003

142. Bender, J. L., Hohenadel, J., Wong, J., Katz, J., Ferris, L. E., Shobbrook, C., Warr, D., & Jadad, A. R. (2008). What patients with cancer want to know about pain: a qualitative study. *Journal of Pain and Symptom Management, 35*(2), 177-187. doi: 10.1016/j.jpainsymman.2007.03.011

143. Galarce, E. M., Ramanadhan, S., Weeks, J., Schneider, E. C., Gray, S. W., & Viswanath, K. (2011). Class, race, ethnicity and information needs in post-treatment cancer patients. *Patient Education and Counseling, 85*(3), 432-439. doi: 10.1016/j.pec.2011.01.030

144. Harding, R., Selman, L., Beynon, T., Hodson, F., Coady, E., Read, C., Walton, M., Gibbs, L., & Higginson, I. J. (2008). Meeting the communication and information needs of chronic heart failure patients. *Journal of Pain and Symptom Management, 36*(2), 149-156. doi: 10.1016/j.jpainsymman.2007.09.012

145. Molassiotis, A., Brunton, L., Hodgetts, J., Green, A. C., Beesley, V., Mulatero, C., Newton-Bishop, J. A., & Lorigan, P. (2014). Prevalence and correlates of unmet supportive care needs in patients with resected invasive cutaneous melanoma. *Annals of oncology: official journal of the European Society for Medical Oncology / ESMO*. doi: 10.1093/annonc/mdu366

146. Roberts, K., Kilicoglu, H., Fiszman, M., & Demner-Fushman, D. (2014). Decomposing Consumer Health Questions. *ACL 2014*.

147. Shea-Budgell, M. A., Kostaras, X., Myhill, K. P., & Hagen, N. A. (2014). Information needs and sources of information for patients during cancer follow-up. *Current Oncology (Toronto, Ont.), 21*(4), 165-173. doi: 10.3747/co.21.1932

148. Umgelter, K., Anetsberger, A., Schmid, S., Kochs, E., Jungwirth, B., & Blobner, M. (2014). [Survey on the need for information during the preanesthesia visit.]. *Der Anaesthesist*. doi: 10.1007/s00101-014-2365-0

149. Zirkzee, E., Ndosi, M., Vlieland, T. V., & Meesters, J. (2014). Measuring educational needs among patients with systemic lupus erythematosus (SLE) using the Dutch version of the Educational Needs Assessment Tool (D-ENAT). *Lupus*. doi: 10.1177/0961203314544188

150. Alzougool, B., Gray, K., & Chang, S. (2009). An In-depth Look at an Informal Carer's Information Needs: A Case Study of a Carer of a Diabetic Child. *Electronic Journal of Health Informatics, 4*(1).

151. Boot, C. R. L., & Meijman, F. J. (2010). Classifying health questions asked by the public using the ICPC-2 classification and a taxonomy of generic clinical questions: an empirical exploration of the feasibility. *Health Communication, 25*(2), 175-181. doi: 10.1080/10410230903544969

152. Shenson J.A., I. E., Colon N., Jackson G.P. (In press). Application of a consumer health information needs taxonomy to questions in maternal-fetal care. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*.

153. Allphin, M. (2013). Patient Portals 2013: On Track for Meaningful Use? : KLAS research.

154. Osborn, C. Y., Rosenbloom, S. T., Stenner, S. P., Anders, S., Muse, S., Johnson, K. B., Jirjis, J., & Jackson, G. P. (2011). MyHealthAtVanderbilt: policies and procedures governing patient portal functionality. *Journal of the American Medical Informatics Association: JAMIA, 18 Suppl 1*, i18-23. doi: 10.1136/amiajnl-2011-000184

155. Hobbs, J., Wald, J., Jagannath, Y. S., Kittler, A., Pizziferri, L., Volk, L. A., Middleton, B., & Bates, D. W. (2003). Opportunities to enhance patient and physician e-mail contact. *International Journal of Medical Informatics, 70*(1), 1-9.

156. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., & Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research, 12*, 2825-2830.

157. Denny, J. C., Irani, P. R., Wehbe, F. H., Smithers, J. D., & Spickard, A. (2003). The KnowledgeMap project: development of a concept-based medical school curriculum database. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 195-199.

158. Denny, J. C., Smithers, J. D., Armstrong, B., & Spickard, A., 3rd. (2005). "Where do we teach what?" Finding broad concepts in the medical school curriculum. *Journal of General Internal Medicine, 20*(10), 943-946. doi: 10.1111/j.1525-1497.2005.0203.x

159. Denny, J. C., Miller, R. A., Waitman, L. R., Arrieta, M. A., & Peterson, J. F. (2009). Identifying QT prolongation from ECG impressions using a general-purpose Natural Language Processor. *International Journal of Medical Informatics, 78 Suppl 1*, S34-42. doi: 10.1016/j.ijmedinf.2008.09.001

160. Denny, J. C., & Peterson, J. F. (2007). Identifying QT prolongation from ECG impressions using natural language processing and negation detection. *Studies in Health Technology and Informatics, 129*(Pt 2), 1283-1288.

161. Denny, J. C., Peterson, J. F., Choma, N. N., Xu, H., Miller, R. A., Bastarache, L., & Peterson, N. B. (2009). Development of a natural language processing system to identify timing and status of colonoscopy testing in electronic medical records. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium, 2009*.

162. Denny, J. C., Spickard, A., Miller, R. A., Schildcrout, J., Darbar, D., Rosenbloom, S. T., & Peterson, J. F. (2005). Identifying UMLS concepts from ECG Impressions using KnowledgeMap. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 196-200.

163. Real, R., & Vargas, J. M. (1996). The probabilistic basis of Jaccard's index of similarity. *Systematic biology*, 380-385.

164. Huang, A. (2008). *Similarity measures for text document clustering.* Paper presented at the Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand.

165. Dixon, R. F. (2010). Enhancing primary care through online communication. *Health Affairs (Project Hope), 29*(7), 1364-1369. doi: 10.1377/hlthaff.2010.0110

166. (AAFP), A. A. o. F. P. (2014). 2014 proposed Medicare Physician Fee Schedule - American Academy of Family Physicians (AAFP). Retrieved 2014 Sep 4, from http://www.aafp.org/dam/AAFP/documents/advocacy/payment/medicare/ES-2014ProposedFeeSchedule-071913.pdf

167. Lee, D., de Keizer, N., Lau, F., & Cornet, R. (2014). Literature review of SNOMED CT use. *Journal of the American Medical Informatics Association: JAMIA, 21*(e1), e11-19. doi: 10.1136/amiajnl-2013-001636

168. Zhou, L., Plasek, J. M., Mahoney, L. M., Karipineni, N., Chang, F., Yan, X., Chang, F., Dimaggio, D., Goldman, D. S., & Rocha, R. A. (2011). Using Medical Text Extraction, Reasoning and Mapping System (MTERMS) to process medication information in outpatient clinical notes. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium, 2011*, 1639-1648.

169. Keselman, A., Tse, T., Crowell, J., Browne, A., Ngo, L., & Zeng, Q. (2007). Assessing consumer health vocabulary familiarity: an exploratory study. *Journal of Medical Internet Research, 9*(1). doi: 10.2196/jmir.9.1.e5

170. Zeng, Q. T., & Tse, T. (2006). Exploring and developing consumer health vocabularies. *Journal of the American Medical Informatics Association: JAMIA, 13*(1), 24-29. doi: 10.1197/jamia.M1761

171. Zeng, Q. T., Tse, T., Divita, G., Keselman, A., Crowell, J., Browne, A. C., Goryachev, S., & Ngo, L. (2007). Term identification methods for consumer health vocabulary development. *Journal of Medical Internet Research, 9*(1). doi: 10.2196/jmir.9.1.e4

172. Smith, C. A. (2011). Consumer language, patient language, and thesauri: a review of the literature. *Journal of the Medical Library Association: JMLA, 99*(2), 135-144. doi: 10.3163/1536-5050.99.2.005

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