



Published in final edited form as:

Int J Med Inform. 2017 September ; 105: 110–120. doi:10.1016/j.ijmedinf.2017.06.004.

A Comparison of Rule-Based and Machine Learning Approaches for Classifying Patient Portal Messages

Robert M. Cronin, M.D., M.S.^{1,2,3}, Daniel Fabbri, Ph.D.^{1,4}, Joshua C. Denny, M.D., M.S.^{1,2}, S. Trent Rosenbloom, M.D., M.P.H.^{1,2,3}, and Gretchen Purcell Jackson, M.D., Ph.D.^{1,3,5}

¹Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, Tennessee, USA

²Department of Medicine, Vanderbilt University Medical Center, Nashville, Tennessee, USA

³Department of Pediatrics, Vanderbilt University Medical Center, Nashville, Tennessee, USA

⁴Department of Computer Science, Vanderbilt University, Nashville, Tennessee, USA

⁵Department of Pediatric Surgery, Vanderbilt University Medical Center, Nashville, Tennessee, USA

Abstract

Objective—Secure messaging through patient portals is an increasingly popular way that consumers interact with healthcare providers. The increasing burden of secure messaging can affect clinic staffing and workflows. Manual management of portal messages is costly and time consuming. Automated classification of portal messages could potentially expedite message triage and delivery of care.

Materials and Methods—We developed automated patient portal message classifiers with rule-based and machine learning techniques using bag of words and natural language processing (NLP) approaches. To evaluate classifier performance, we used a gold standard of 3253 portal messages manually categorized using a taxonomy of communication types (i.e., main categories of informational, medical, logistical, social, and other communications, and subcategories including prescriptions, appointments, problems, tests, follow-up, contact information, and acknowledgement). We evaluated our classifiers' accuracies in identifying individual communication types within portal messages with area under the receiver-operator curve (AUC). Portal messages often contain more than one type of communication. To predict all

Corresponding author: Robert M. Cronin, Department of Biomedical Informatics, Vanderbilt University Medical Center, 2525 West End Blvd., Suite 1475, Nashville, TN, 37232, USA, robert.cronin@vanderbilt.edu Phone: 615-936-5097.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Competing interests: None.

Provenance and peer review: Not commissioned; externally peer reviewed.

Author contribution statement:

RMC and GPJ were involved with the conception, design, and data collection of the study. RMC, DF, JCD, STR, and GPJ were involved with the analysis of the study. All authors were involved with the writing and editing of the manuscript. All authors approved the final submitted manuscript.

communication types within single messages, we used the Jaccard Index. We extracted the variables of importance for the random forest classifiers.

Results—The best performing approaches to classification for the major communication types were: logistic regression for medical communications (AUC: 0.899); basic (rule-based) for informational communications (AUC: 0.842); and random forests for social communications and logistical communications (AUCs: 0.875 and 0.925, respectively). The best performing classification approach of classifiers for individual communication subtypes was random forests for Logistical-Contact Information (AUC: 0.963). The Jaccard Indices by approach were: basic classifier, Jaccard Index: 0.674; Naïve Bayes, Jaccard Index: 0.799; random forests, Jaccard Index: 0.859; and logistic regression, Jaccard Index: 0.861. For medical communications, the most predictive variables were NLP concepts (e.g. Temporal Concept, which maps to ‘morning’, ‘evening’ and Idea or Concept which maps to ‘appointment’ and ‘refill’). For logistical communications, the most predictive variables contained similar numbers of NLP variables and words (e.g. Telephone mapping to ‘phone’, ‘insurance’). For social and informational communications, the most predictive variables were words (e.g. social: ‘thanks’, ‘much’, informational: ‘question’, ‘mean’).

Conclusions—This study applies automated classification methods to the content of patient portal messages and evaluates the application of NLP techniques on consumer communications in patient portal messages. We demonstrated that random forest and logistic regression approaches accurately classified the content of portal messages, although the best approach to classification varied by communication type. Words were the most predictive variables for classification of most communication types, although NLP variables were most predictive for medical communication types. As adoption of patient portals increases, automated techniques could assist in understanding and managing growing volumes of messages. Further work is needed to improve classification performance to potentially support message triage and answering.

Keywords

Patient portal; text classification; natural language processing; machine learning

1. INTRODUCTION

Patient portals, online applications that allow patients to interact with their healthcare providers and institutions, have had increasing adoption because of consumer demand and governmental regulations.[1] Secure messaging is one of the most popular functions of patient portals, and this function allows individuals to interact with their healthcare providers and health information.[2–9] The increasing burden of secure messaging has been demonstrated in multiple settings with providers having a few messages per week to multiple messages per day within a few years after patient portal implementation.[10–13] This increasing burden can affect staffing and workflows of a clinic and health care providers.[14] Being able to determine the contents of these messages in an automated fashion could potentially mitigate this burden.

Prior research has demonstrated that users express diverse health-related needs in portal messages, and substantial medical care is delivered through portal interactions.[10, 15–19]

Message content can contain informational (e.g., what is the side effect of simvastatin?), logistical (e.g., what time does the pharmacy open?), medical (e.g., I am having a new numbness in my legs), and social (e.g., Please thank your nurse for his care of my wife) communication types. Classification of messages may aid message management with triage to appropriate resources or personnel. Identifying when medical care is delivered in patient portal messages could support online compensation models beyond the limited codes for online care transitions and tele-health services.[20].

Categorization of portal message content can be viewed as a text classification problem. The most popular methods employed for text classification include a manual approach, where a human will classify each message, or automatically, through rule-based approaches based on words or phrases that appear in the text or machine learning techniques (e.g., logistic regression, random forests, support vector machines).[21–27] 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. [28, 29]. These studies demonstrated excellent areas under the operator-receiver curve over 0.88. Researchers have also attempted to identify adverse drug reactions from consumer-generated text [30–32]. Although text classification has been successfully done for consumer-generated text from online forums and social media, this approach has not been applied to and evaluated for secure messages from patient portals.

A limited number of studies have classified portal messages in primary care settings only using manual methods.[15, 16] North et al. manually classified 323 messages, demonstrating 37% were medication related, 23% were symptom related, 20% were test related, 7% were medical questions, 6% were acknowledgements, and 9% had more than one issue [16]. Haun et al. asked senders classify their messages in predefined categories and observed the following distribution, although user-assigned categories were not consistently applied accurately: 59% general (i.e., condition management/report, specialty/procedure request, correspondence request, medication refill request.), 24% appointments (i.e., confirmations, cancellations, specialty appointment requests), and 16% refill and medication inquiries [15]. As patient portal and secure messaging adoption increases, understanding the content of these messages and their implications for provider workload becomes more important.

Our research team has previously evaluated different methods of automatically classifying messages sent through a patient portal used by multiple specialties at a large, tertiary care institution. We compared the performance of basic classification and machine learning approaches to determine the major types of portal message content using a gold standard of 1,000 portal messages classified by the semantic types of communications within the message, including informational, medical, logistical and social communication types.[17] We discovered that automated methods have promise for predicting major semantic types of communications, but this work was limited because of the small data set, and it did not include an analysis of what features (e.g., words, concepts, semantic types) are most important for the machine learning classifiers or an attempt to classify messages beyond the major categories in a rich hierarchy containing many communication subtypes. In this manuscript, we expand on our previous work by comparing automated approaches to message classification using a substantially larger gold standard and the full semantic

communication type hierarchy (Figure 1), and determining the features important for classification.

2. MATERIAL AND METHODS

2.1 Setting

This study was conducted at Vanderbilt University Medical Center (VUMC), a private, nonprofit institution that provides primary and tertiary referral care to over 500,000 patients annually. The VUMC Institutional Review Board approved this study as non-human subjects research. VUMC launched a locally-developed patient portal called My Health at Vanderbilt (MHAV) in 2005, with accounts for pediatric patients added in 2007.[35, 36] MHAV functions include access to selected parts of the electronic health record (EHR), delivery of personalized health information, and secure messaging with healthcare providers. MHAV currently has over 400,000 cumulatively registered users, including more than 22,800 pediatric accounts, with almost 300,000 logins per month with over 500,000 messages generated by patients per year.

MHAV users can only send messages to healthcare providers with whom they have a prior or scheduled outpatient appointment. Clinical teams manage messages sent through MHAV such that a message may be answered directly by the healthcare provider or a staff member (e.g., nurse, administrative assistant, or allied health professional).[37, 38] MHAV messages are sent to clinical groups, which usually represent outpatient practices, but these messages may be sent while patients are in an inpatient or outpatient setting. All MHAV message content is stored in the EHR.

2.2 Taxonomy of Communication Types (Outputs)

Our research team has developed a taxonomy of consumer health-related needs and communication types (Figure 1). This taxonomy represents the semantic types of consumer health communications and has evolved from taxonomies used to describe clinical questions and organize medical information resources. The taxonomy been applied to diverse communications including questions from patient journals, patient and caregiver interviews, and messages from patient portals, and it has been evaluated with studies of interrater reliability of its applications.[17, 19, 34, 39] This taxonomy divides consumer health communications into five main types: Informational, Medical, Logistical, Social, and other (Table 1). *Informational* communications 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 [40]. *Medical* communications involve the delivery of medical care, such as the expression of a new symptom requiring management or the communication of a test result. *Logistical* communications address pragmatic information, such as the location of a clinic. *Social* communications include personal exchanges such as an expression of gratitude or a complaint. The *other* communication category covers communications that are incomplete, unintelligible, or not captured in other parts of the taxonomy. Secure portal messages can contain more than one type of communication (Figure 2). Details about the development and validation of the taxonomy are reported elsewhere [33].

2.3 Gold Standard

To train and test the classifiers, we used gold standard of 3,253 MHAV portal messages that had been manually annotated with all applicable communication types in the portal message. In creating this gold standard, we extracted all patient-generated MHAV messages from 2005 to 2014 stored in the VUMC Synthetic Derivative (SD), a database containing a de-identified copy of all hospital medical records created for research purposes. From 2.5 million such messages present in the SD, we randomly selected a set of messages, distributed equally over the 10-year period. Therefore, about 0.15% of all messages were sampled in each year. Messages were manually analyzed by three medical students who had done clinical clerkships. A 500 message training set was randomly selected from the message corpus; these messages were tagged independently then reviewed with two physicians with experience in using the taxonomy to achieve consensus. The remaining messages were annotated independently by all three students, and discrepancies were discussed with taxonomy experts to achieve consensus. Interrater reliability was measured by Mezzich's kappa, an extension of Cohen's kappa for more than two raters, and it was 0.78 for main types and 0.62 for subtypes, indicating substantial agreement. The creation of this gold standard with a detailed description of portal message content and measurement of interrater reliability is described elsewhere.[33]

2.4 Automated Classifiers

Classifiers' inputs included words from messages, natural language processing (NLP) concepts, NLP semantics, or combinations of these entities. We built two sets of classifiers to identify individual communication types in messages: a rule-based classifier (which we called the basic classifier) and three machine learning classifiers. Since a different machine learning classifier was built for each major communication type and subtypes present in greater than 10% of messages, we created a total of 11 rule-based classifiers and 231 machine learning classifiers (3 types of classifiers \times 7 feature sets \times 11 major and subtypes).

The basic classifier determined if individual communication types were present in messages through regular expressions (Table 2). Words were chosen for the basic classifier based on expert knowledge of the research team, and they were refined through iterative testing. These words represented the typical phrases or word combinations that would appear in a message of a specific communication type (e.g. pain would be seen in a medical communication type as the patient is likely describing a symptom they are having that needs the delivery of medical care). If a word was present in a message, then the basic classifier would output a 1 indicating the message belongs to that communication type, otherwise the classifier would output a 0.

The machine learning approaches to classification included Naïve Bayes, logistic regression, and random forests. Each machine learning classifier output a probability between 0 and 1 based on whether the communication type was present in the message. To create the classifiers, we used python's scikit learn package.[41] We used Bernoulli Naïve Bayes with an alpha of 0.1 and random forests with 500 trees.

Features that served as classifier inputs were words from messages or bag of words (BoW), NLP inputs such as concept unique identifiers (CUIs) and semantic types (STYs), or combinations of these entities. 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 NLP tasks, such as discovering clinical concepts in clinical text in multiple studies, with high sensitivity and specificity. [42–47] 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 (e.g., ‘if’, ‘and’, ‘or’) were removed from messages for the BoW representation.

The machine learning classifiers were trained and tested with the gold standard corpus of portal messages using 10-fold cross validation. To determine what features were important for prediction, we extracted the variable importance for the random forest with the complete feature set, measured by the decrease in impurity at the nodes using those variables. Variable importance is determined by how much the prediction error increases when data for that variable is changed while all others are left unchanged. [48] Therefore the most predictive variables (variables with the highest variable importance) are the variables, or features, that affect the classification the most.

2.5 Evaluation and Statistical Analysis

We evaluated the ability of classifiers to predict all major categories and specific subcategories with area under the receiver-operator curves (AUCs). Since classifiers may have to predict multiple communication types for a single message (Figure 2), we used the Jaccard index[49] to determine similarities between manual categorization and classifier categorization of all communication types in a single message. The Jaccard index is a measure of the similarity between two sets:

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

It has similar performance in text classification tasks as other similarity metrics such as Pearson’s correlation coefficient.[50] A Jaccard Index of 1 indicates that the sets A and B contain 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:

$$J(\text{Gold Standard}, \text{Classifier Prediction}) = \frac{|\text{Gold Standard} \cap \text{Classifier Prediction}|}{|\text{Gold Standard} \cup \text{Classifier Prediction}|}$$

We averaged the Jaccard indices across all messages to give an overall estimation of the ability to predict the set of communication types across the entire corpus of messages.

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. 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 0 and 1 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.

3. RESULTS

3.1 Gold Standard

The composition of the gold standard has been described elsewhere [33]. In summary, the gold standard contained 3,253 patient-generated messages, which were sent about 3,116 unique patients. The messages were predominantly sent about patients who were female (1,937; 62.2%) and Caucasian (2,772; 89.0%) with a median age 50 years of age (range of 1 month to 112 years). Using the gold standard, the 3,253 messages contained 2,351 (72.3%) medical, 922 (28.3%) social, 806 (24.8%) logistical, 404 (12.4%) informational, and 114 (3.5%) other communications. ‘Other’ communications were excluded from analysis as they were predominantly incomplete, unintelligible, or error messages.

3.2 Predictive Performance for Major Communication Types

AUCs for classifiers identifying major communication types ranged from the lowest of 0.604 for the Naïve Bayes classifier for social communication types to the highest of 0.925 for the random forest machine learning classifier for logistical types (Figure 3, Supplemental Table 1). The best classification approaches with highest AUCs for major communication types varied and included: medical: logistic regression 0.899 (95% Confidence Interval (CI): 0.887,0.911) BoW+STY; informational: basic (rule based) classifier 0.842 (95% CI: 0.836,0.848); social: random forest 0.875 using BoW+CUI+STY (95% CI: 0.862,0.887); and logistical: random forest 0.925 BoW+CUI (95% CI: 0.916,0.934).

The AUCs for the subtypes ranged from the lowest of 0.609 for the Naïve Bayes classifier in both Medical-Follow-up and Social-Acknowledgement to the highest of 0.963 for the random forest classifier in Logistical-Contact Information (Supplemental Figure 1, Supplemental Table 2). The classification approaches with highest AUCs for the subtypes used the random forest classifier and BoW feature..

3.3 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 (Figure 4). 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.

3.4 Features

There were a total of 9,643 words, 6,040 CUIs, and 200 STYs that were used by the machine learning classifiers. The top 10 most predictive variables, or variables that are most important in classification for random forests for each major communication type are shown in Table 3. As STYs are mapped to multiple words that are found in the text of messages, some common words associated with the ‘most important’ semantic types are described in Table 4.

For informational communications, the majority of the most predictive variables were words. However, multiple STYs were the top 10 predictive variables and mapped to words like “critical”, “high”, “lab work”, and “medication”. Logistical communications contained similar numbers of CUIs and words, with the concept “Telephone” being mapped to the word “phone” and the concept “Insurances” mapped to the word “insurance”. For social communications, the word “name”, had the third most predictive variable importance. The “Intellectual_Product” STY maps to the word “name”. In medical communications, the majority of important variables were STYs. These STYs map to words like “morning”, “day”, “evenings”, “Friday”, “appointment”, “refill”, “mg”, and “prescription”.

4. DISCUSSION

This study demonstrates promise in the ability to achieve automated classification of the content of patient portal messages into types of communications and to apply NLP techniques to consumer text written in patient portal messages. We developed and evaluated a set of methods to identify the communication types in patient messages sent through a widely deployed patient portal at an academic medical center. This research provides evidence that patient-generated messages can be automatically classified into semantic communication types with good accuracy. This work expanded upon our previous work [17] by including a larger corpus of messages with improved classifier performance. We also demonstrated the ability to accurately classify portal messages into communication subtypes, in certain cases better than for main communication types. Logistic regression and random forest classifiers performed the best for all communication types and subtypes except for informational types (Figure 3). Finally, we evaluated variables of importance and showed words to be the most predictive variables of importance for random forests for classifying messages into all major categories except for medical communications, for which semantic types were variables with best predictive performance (Table 4).

As adoption of patient portals increases, automated techniques could assist in understanding and managing growing volumes of messages. Automated classification of these messages has several important potential applications. By automatically classifying messages, the types of health-related needs expressed in patient portals can be better understood, which could inform the development of consumer health resources and design of portal functions. Classification of messages could potentially support message triage. Identification of portal messages with medical communications could help determine the volume and types of care delivered within patient portals and support compensation models for online care.

4.1 Features

The features that were most important for the random forest classifiers that used all of the features available (BoW, CUI, and STY) varied by communication type. Classifiers to identify informational communications employed more words than STYs 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 Informational-Tests/ Interpretation in our taxonomy. Examining the most predictive variables for informational communication types led to several interesting findings. The word “red” ranked seventh in importance (Table 4). In the SD version of messages from patient portal, the annotation “red” was added to laboratory results to designate an abnormal value. “Statistical_mean” was a CUI ranked sixth in importance for classifying informational communication types. This CUI maps to the words “mean” and “means”. A phrase “what does this test mean” is an informational communication type that was incorrectly mapped to the “Statistical_mean” CUI, which illustrates a homonym that was identified inappropriately. Although interpreted incorrectly, this mapping reflects a characteristic language within these messages, which may work well for identification, but not for interpretation.

For classifying logistical communication types, the word “phone” was the most predictive variable. This word was added to replace actual telephone numbers in de-identified SD portal messages, so for logistical communication types, the word “phone” (indicating the de-identification system had identified a phone number) was the most important. Without the prior step using a named entity recognition system for phone numbers (essentially a kind of “semantic type” classification), classification accuracy may not have been as high. Therefore, classifiers would likely not have similar performance characteristics on fully identified messages.

For classification of 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 significant variability of words within a single subtype, such as Medical-Problems. In this subtype, the words communicating diverse symptoms such as pain or nausea in a message would be mapped to the STY Finding.

The difference in the most predictive variables across communication types reflects the diversity of language within patient portal messages containing such communications. For the 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 highest predictive variables.

4.2 Performance of the classifiers

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. As compared to our previous work, the larger corpus led to an increase in AUCs for all major communication types by up to 0.05 (0.84 to 0.89). The good performance of the classifiers for major communication types can help triaging these messages. For example, medical communications most likely need a response from a healthcare provider, such as a nurse or physician, as these messages typically have to do with delivery of medical care. An administrative assistant or office manager may better address messages with logistical communications as they often involve pragmatic questions rather than medical knowledge.

We demonstrated that identification of most major communication types and subtypes requires a sophisticated machine learning technique because the expression of communication categories within the messages is complex. The best machine learning classifiers outperformed the basic classifiers in all subtypes and all major communication types except for the informational communication type. The basic classifier may have outperformed the machine learning classifiers for two reasons. First, informational communication types were the least likely to be present and most often occurred with another communication type. Therefore, machine learning algorithms would have more difficulty detecting these communication types because there was too much noise in the messages to detect the signal of the few most predictive words. Second, there were fewer examples for training in the dataset and more subcategories in this informational portion of the taxonomy. Finally, there may have been a few common terms that occurred in those messages that were not present in other ones.

The machine learning classifiers performed well for several communication types that have significant implications in the delivery of care, and their accurate identification could potentially expedite care. For example, the machine learning classifiers performed best in identifying the following subtypes of communications: Medical-Prescriptions and Medical-Problems. Medical-Problems communications, present in 15% of messages, contain descriptions of new or worsening symptoms. Medical-Prescription communications, also present in 15% of messages, involve the creation, renewal, or adjustment of a prescription. Managing new symptoms from a patient and adjusting prescriptions are activities traditionally done in outpatient clinic visit. When done through messaging, financial incentives are lacking.[51] Proposals for compensating online care have been developed, such as billing codes for transition of care and telehealth services, [20] 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 support the need for models for compensation.

The communication subtypes occurring most commonly in portal messages included Medical-Follow up (26% of messages) and Logistical-Contact Information (25% of messages). Logistical-Contact Information communications were identified very accurately with the best classifier performing with an AUC of 0.963 (random forest), likely due to most messages containing information about telephone and fax numbers. However, Medical-Follow up communications were less accurately classified, with the best classifier having an AUC of 0.789 (random forest). Medical-Follow up communications had a wide variety of

topics expressed within them, ranging from blood pressure readings at home to seeing if a patient was having post-operative pain. The difficulty with these predictions are likely the result of the diversity of topics involved in follow up care.

More complex machine learning methods with NLP performed better in identifying the communication types in the more complex parts of the taxonomy. Contrary to prior studies, [30] the performance improvement with a combination of BoW, CUIs, and STYs over BoW alone in this study was marginal. These findings suggest that the words within the document predict communication types well, and that adding NLP techniques do not add much. NLP tools may incorrectly identify text due to misspellings and undefined abbreviations. Patient-generated text is likely to have less formal biomedical language and thus may have fewer identifiable UMLS concepts than formal medical texts. In addition, portal messages tend to be relatively short, and this trend is supported as healthcare consumers gain experience with succinct communication through applications like Twitter, smartphone-based text messaging, and instant messaging. Clinical messages may omit inferred information, “a mass” may implicitly mean “a mass on a breast”. Better consumer health vocabularies that tackle these issues may help address these limitations. Higher order NLP methods, such as negation, also likely had little impact on content type. Additional work exploring context and semantics of portal messages is needed.

While random forests generally had the best AUCs for a single communication type, logistic regression classifiers most accurately identified 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 as the error bars did not overlap (Figure 4). The random forest classifier’s Jaccard index was only slightly, but not significantly, lower than logistic regression. 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.

4.3 Limitations

This research has several limitations. First, classification of portal messages into communication types may be the first step, but not the equivalent of understanding the meaning of the message. Fully comprehending the meaning to support question answering or triage will require additional research. Second, we used UMLS for our classifiers; however, this collection of terminologies may not be the most appropriate for capturing consumer expressions of concepts.

This study was conducted at a single institution with a locally developed patient portal. Although the communication types identified are common ones seen in other papers about patient portal messaging,[15, 52] our results may be limited by the unique policies and procedures developed for MHAV as well as dialects and communication styles of patients from the Southeastern United States. This study included data as old as 2005; some content of messages may have become antiquated based on secular trends. Our machine learning

models were 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 is robust to prevent overfitting.[53–56].

This study also looked only at supervised methods using a taxonomy of categories. Therefore, other associations and correlations that may be found by unsupervised learning are not captured in our study. This latter approach could lead to different types of classification and gain insight into diagnosis and treatment as well as the semantics and context of the concepts that are found in messages, and it is a topic of future research for our group.

Automatic classification of communication types in patient portals has several potentially important applications. First, it could allow routing of patient-generated messages to different members of the health care team or information resources without human intervention. However, this task is complicated by our observation that portal messages commonly have multiple communication types. Improper workflow could lead to multiple responses from different people to a single message, which might be confusing to the patient. Further work is needed to employ classification results in the task of routing. 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 [52]. Utilizing automated classifiers to detect urgent messages could prevent adverse events by prioritizing responses or alerting a provider through an alternative means of communication. However, there are also risks of misclassifying certain messages, as misclassifying a life-threatening symptom into a category deemed not important may lead to adverse consequences. 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 identify substantial volumes of uncompensated care that drive compensation model development.

5. CONCLUSIONS

This manuscript demonstrates promise in identifying the types of communication in patient-generated secure messages through patient portals. Logistic regression and random forest classifiers performed the best for almost all communication types and subtypes, but a combination of the different types of classifiers would likely be needed to classify all communication types in a message. As adoption of patient portals increases, automated techniques may be needed to assist in understanding and managing growing volumes of messages. Automated classification of messages through patient portals may aid in connecting patients to needed resources and in triaging messages. These classifiers could support consumer health informatics research to understand the nature of communications and types of care delivered within patient portals.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We are grateful to Shilo Anders, Mary Masterman, Ebone Ingram, and Jared Shenson for their assistance in creation of the gold standard for this research project. The portal message content used for the analyses was obtained from VUMC's Synthetic Derivative (SD), which is supported by institutional funding and by the Vanderbilt CTSA grant ULTR000445 from NCATS/NIH. Robert Cronin was supported by the 5T15LM007450-12 training grant from the National Library of Medicine.

References

1. Shapochka A. Providers Turn to Portals to Meet Patient Demand. Meaningful Use / Journal of AHIMA. 2012
2. Tang PC, Lansky D. The missing link: bridging the patient-provider health information gap. *Health Aff (Millwood)*. 2005; 24:1290–1295. [PubMed: 16162575]
3. Calabretta N. Consumer-driven, patient-centered health care in the age of electronic information. *J Med Libr Assoc*. 2002; 90:32–37. [PubMed: 11838457]
4. Koonce TY, Giuse DA, Beauregard JM, Giuse NB. Toward a more informed patient: bridging health care information through an interactive communication portal. *J Med Libr Assoc*. 2007; 95:77–81. [PubMed: 17252070]
5. Bussey-Smith KL, Rossen RD. A systematic review of randomized control trials evaluating the effectiveness of interactive computerized asthma patient education programs. *Ann Allergy Asthma Immunol*. 2007; 98:507–516. quiz 516, 566. [PubMed: 17601262]
6. Jackson CL, Bolen S, Brancati FL, Batts-Turner ML, Gary TL. A systematic review of interactive computer-assisted technology in diabetes care. *Interactive information technology in diabetes care. J Gen Intern Med*. 2006; 21:105–110. [PubMed: 16390512]
7. Neve M, Morgan PJ, Jones PR, Collins CE. Effectiveness of web-based interventions in achieving weight loss and weight loss maintenance in overweight and obese adults: a systematic review with meta-analysis. *Obes Rev*. 2010; 11:306–321. [PubMed: 19754633]
8. Vandelanotte C, Spathonis KM, Eakin EG, Owen N. Website-delivered physical activity interventions a review of the literature. *Am J Prev Med*. 2007; 33:54–64. [PubMed: 17572313]
9. Walters ST, Wright JA, Shegog R. A review of computer and Internet-based interventions for smoking behavior. *Addict Behav*. 2006; 31:264–277. [PubMed: 15950392]
10. Cronin R, Davis S, Shenson J, Chen Q, Rosenbloom S, Jackson G. Growth of Secure Messaging Through a Patient Portal as a Form of Outpatient Interaction across Clinical Specialties. *Appl Clin Inform*. 2015; 6:288–304. [PubMed: 26171076]
11. Shenson J, Cronin R, Davis S, Chen Q, Jackson G. Rapid growth in surgeons' use of secure messaging in a patient portal. *Surgical Endoscopy*. 2015:1–9.
12. Laccetti AL, Chen B, Cai J, Gates S, Xie Y, Lee SJ, Gerber DE. Increase in Cancer Center Staff Effort Related to Electronic Patient Portal Use. *J Oncol Pract*. 2016; 12:e981–e990. [PubMed: 27601511]
13. Dexter EN, Fields S, Rdesinski RE, Sachdeva B, Yamashita D, Marino M. Patient-Provider Communication: Does Electronic Messaging Reduce Incoming Telephone Calls? *J Am Board Fam Med*. 2016; 29:613–619. [PubMed: 27613794]
14. Carayon P HP, Cartmill R, Hassol A. Using Health Information Technology (IT) in Practice Redesign: Impact of Health IT on Workflow. *Patient-Reported Health Information Technology and Workflow*, AHRQ Publication No. 15-0043-EF. 2015
15. Haun JN, Lind JD, Shimada SL, Martin TL, Gosline RM, Antinori N, Stewart M, Simon SR. Evaluating user experiences of the secure messaging tool on the Veterans Affairs' patient portal system. *J. Med. Internet Res*. 2014; 16
16. North F, Crane SJ, Stroebel RJ, Cha SS, Edell ES, Tullledge-Scheitel SM. Patient-generated secure messages and eVisits on a patient portal: are patients at risk? *J Am Med Inform Assoc*. 2013; 20:1143–1149. [PubMed: 23703826]

17. Cronin RM, Fabbri D, Denny JC, Jackson GP. Automated Classification of Consumer Health Information Needs in Patient Portal Messages. *AMIA Annu Symp Proc.* 2015:1861–1870. [PubMed: 26958285]
18. Boffa M, Weathers A, Ouyang B. Analysis of Patient Portal Message Content in an Academic Multi-specialty Neurology Practice (S11. 005). *Neurology.* 2015; 84:S11. 005.
19. Robinson JR, Valentine A, Carney C, Fabbri D, Jackson GP. Complexity of medical decision-making in care provided by surgeons through patient portals. *Journal of Surgical Research.* 2017
20. A.A.o.F.P. (AAFP). Proposed Medicare Physician Fee Schedule - American Academy of Family Physicians (AAFP). 2014
21. Friedman JH. On bias, variance, 0/1—loss, and the curse-of-dimensionality. *Data mining and knowledge discovery.* 1997; 1:55–77.
22. Friedman N, Geiger D, Goldszmidt M. Bayesian network classifiers. *Machine learning.* 1997; 29:131–163.
23. McCallum, A., Nigam, K. AAAI-98 workshop on learning for text categorization. Citeseer; 1998. A comparison of event models for naive bayes text classification; p. 41-48.
24. Sahami, M. Learning Limited Dependence Bayesian Classifiers. *KDD;* 1996. p. 335-338.
25. Fette, I., Sadeh, N., Tomasic, A. Learning to detect phishing emails; Proceedings of the 16th international conference on World Wide Web; ACM; 2007. p. 649-656.
26. Fan R-E, Chang K-W, Hsieh C-J, Wang X-R, Lin C-J. LIBLINEAR: A library for large linear classification. *The Journal of Machine Learning Research.* 2008; 9:1871–1874.
27. Genkin A, Lewis DD, Madigan D. Large-scale Bayesian logistic regression for text categorization. *Technometrics.* 2007; 49:291–304.
28. Haas SW, Travers D, Waller A, Mahalingam D, Crouch J, Schwartz TA, Mostafa J. Emergency Medical Text Classifier: New system improves processing and classification of triage notes. *Online journal of public health informatics.* 2014; 6:e178. [PubMed: 25379126]
29. Marafino BJ, John Boscardin W, Adams Dudley R. Efficient and sparse feature selection for biomedical text classification via the elastic net: Application to ICU risk stratification from nursing notes. *J Biomed Inform.* 2015; 54:114–120. [PubMed: 25700665]
30. Sarker A, Gonzalez G. Portable automatic text classification for adverse drug reaction detection via multi-corpus training. *J Biomed Inform.* 2015; 53:196–207. [PubMed: 25451103]
31. Yang M, Kiang M, Shang W. Filtering big data from social media - Building an early warning system for adverse drug reactions. *J Biomed Inform.* 2015; 54:230–240. [PubMed: 25688695]
32. Huh J, Yetisgen-Yildiz M, Pratt W. Text classification for assisting moderators in online health communities. *J Biomed Inform.* 2013; 46:998–1005. [PubMed: 24025513]
33. Jackson GP, SJ., Ingram, E., Masterman, M., Cronin, RM. Personal Communication -A Taxonomy of Consumer Health-Related Communications for Characterizing the Content of Patient Portal Messages. Cronin, R., editor. 2017.
34. Shenson JA, Ingram E, Colon N, Jackson GP. Application of a Consumer Health Information Needs Taxonomy to Questions in Maternal-Fetal Care. *AMIA Annu Symp Proc.* 2015:1148–1156. [PubMed: 26958254]
35. Allphin M. Patient Portals 2013: On Track for Meaningful Use? KLAS research. 2013
36. Osborn CY, Rosenbloom ST, Stenner SP, Anders S, Muse S, Johnson KB, Jirjis J, Jackson GP. MyHealthAtVanderbilt: policies and procedures governing patient portal functionality. *J Am Med Inform Assoc.* 2011; 18(Suppl 1):18–23.
37. Hobbs J, Wald J, Jagannath YS, Kittler A, Pizziferri L, Volk LA, Middleton B, Bates DW. Opportunities to enhance patient and physician e-mail contact. *Int J Med Inform.* 2003; 70:1–9. [PubMed: 12706177]
38. White CB, Moyer CA, Stern DT, Katz SJ. A content analysis of e-mail communication between patients and their providers: patients get the message. *J Am Med Inform Assoc.* 2004; 11:260–267. [PubMed: 15064295]
39. Jackson GP RJ, Ingram E, Masterman M, Ivory C, Holloway D, Anders S, Cronin RM. Personal communication: A technology-based patient and family engagement consult service for the pediatric hospital setting. 2017

40. Purcell GP. Surgical textbooks: past, present, and future. *Ann Surg.* 2003; 238:S34–41. [PubMed: 14703743]
41. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V. Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research.* 2011; 12:2825–2830.
42. Denny JC, Irani PR, Wehbe FH, Smithers JD, Spickard A. The KnowledgeMap project: development of a concept-based medical school curriculum database. *AMIA Annu Symp Proc.* 2003:195–199. [PubMed: 14728161]
43. Denny JC, Smithers JD, Armstrong B, Spickard A 3rd. “Where do we teach what?” Finding broad concepts in the medical school curriculum. *J Gen Intern Med.* 2005; 20:943–946. [PubMed: 16191143]
44. Denny JC, Miller RA, Waitman LR, Arrieta MA, Peterson JF. Identifying QT prolongation from ECG impressions using a general-purpose Natural Language Processor. *Int J Med Inform.* 2009; 78(Suppl 1):S34–42. [PubMed: 18938105]
45. Denny JC, Peterson JF. Identifying QT prolongation from ECG impressions using natural language processing and negation detection. *Stud Health Technol Inform.* 2007; 129:1283–1288. [PubMed: 17911921]
46. Denny JC, Peterson JF, Choma NN, Xu H, Miller RA, Bastarache L, Peterson NB. Development of a natural language processing system to identify timing and status of colonoscopy testing in electronic medical records. *AMIA Annu Symp Proc.* 2009
47. Denny JC, Spickard A, Miller RA, Schildcrout J, Darbar D, Rosenbloom ST, Peterson JF. Identifying UMLS concepts from ECG Impressions using KnowledgeMap. *AMIA Annu Symp Proc.* 2005:196–200. [PubMed: 16779029]
48. Breiman, L. Manual on setting up, using, and understanding random forests v3. 1. Vol. 1. Statistics Department University of California Berkeley, CA; USA: 2002.
49. Real R, Vargas JM. The probabilistic basis of Jaccard’s index of similarity. *Systematic biology.* 1996:380–385.
50. Huang, A. Similarity measures for text document clustering; Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch; New Zealand. 2008. p. 49-56.
51. Dixon RF. Enhancing primary care through online communication. *Health Aff (Millwood).* 2010; 29:1364–1369. [PubMed: 20606189]
52. North F, Crane SJ, Chaudhry R, Ebbert JO, Ytterberg K, Tullidge-Scheitel SM, Stroebel RJ. Impact of patient portal secure messages and electronic visits on adult primary care office visits. *Telemed J E Health.* 2014; 20:192–198. [PubMed: 24350803]
53. Grömping U. Variable importance assessment in regression: linear regression versus random forest. *The American Statistician.* 2009; 63
54. Casanova R, Saldana S, Chew EY, Danis RP, Greven CM, Ambrosius WT. Application of random forests methods to diabetic retinopathy classification analyses. *PLoS ONE.* 2014; 9
55. Liu Y, Traskin M, Lorch SA, George EI, Small D. Ensemble of trees approaches to risk adjustment for evaluating a hospital’s performance. *Health Care Manag Sci.* 2014
56. Sowa J-P, Heider D, Bechmann LP, Gerken G, Hoffmann D, Canbay A. Novel algorithm for non-invasive assessment of fibrosis in NAFLD. *PLoS ONE.* 2013; 8

Summary Table

What is known

- Patient portal secure messaging has increasing adoption leading to increased burden on healthcare providers and staff
- Prior research has demonstrated that users express diverse health-related needs in portal messages, and substantial medical care is delivered through portal interactions
- Few studies have examined the content of portal messages, mostly using manual techniques

What this study added to our knowledge

- We can accurately classify most portal messages into communication types and subtypes using random forests and logistic regression
- For random forests, important variables for classifying messages into all major communication categories were words except for medical communications, for which semantic types were the most important variables
- As adoption of patient portals increases, automated techniques could assist in understanding and managing growing volumes of messages



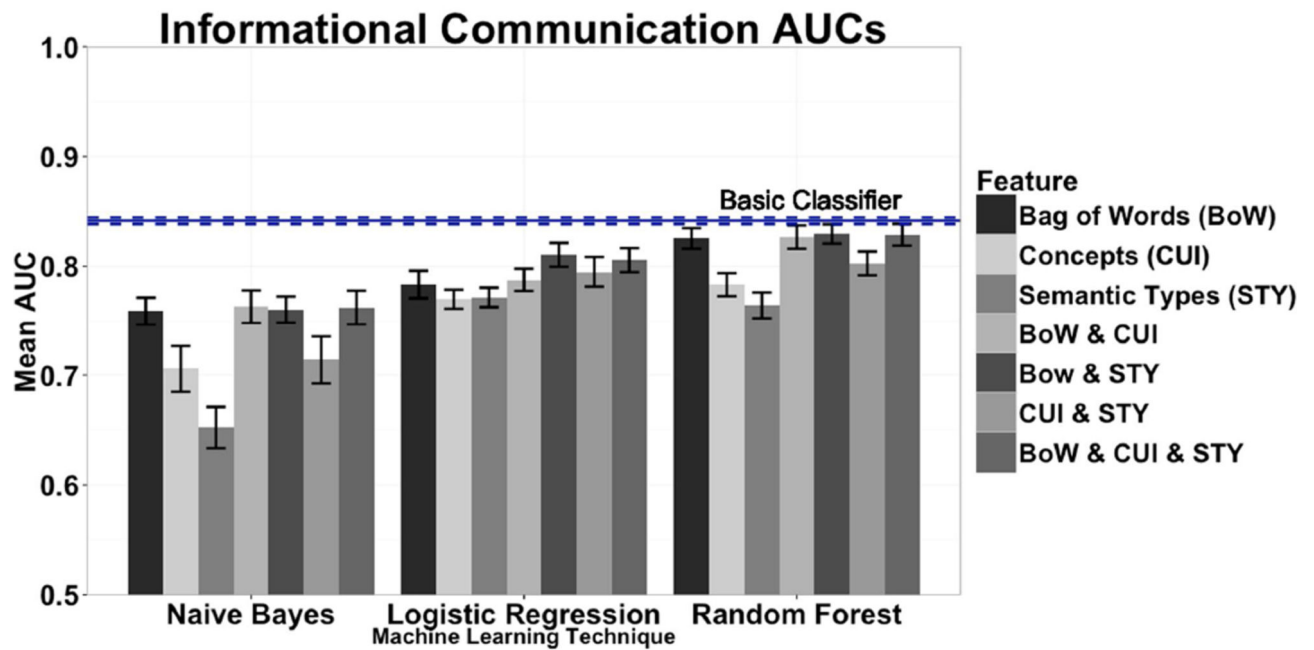
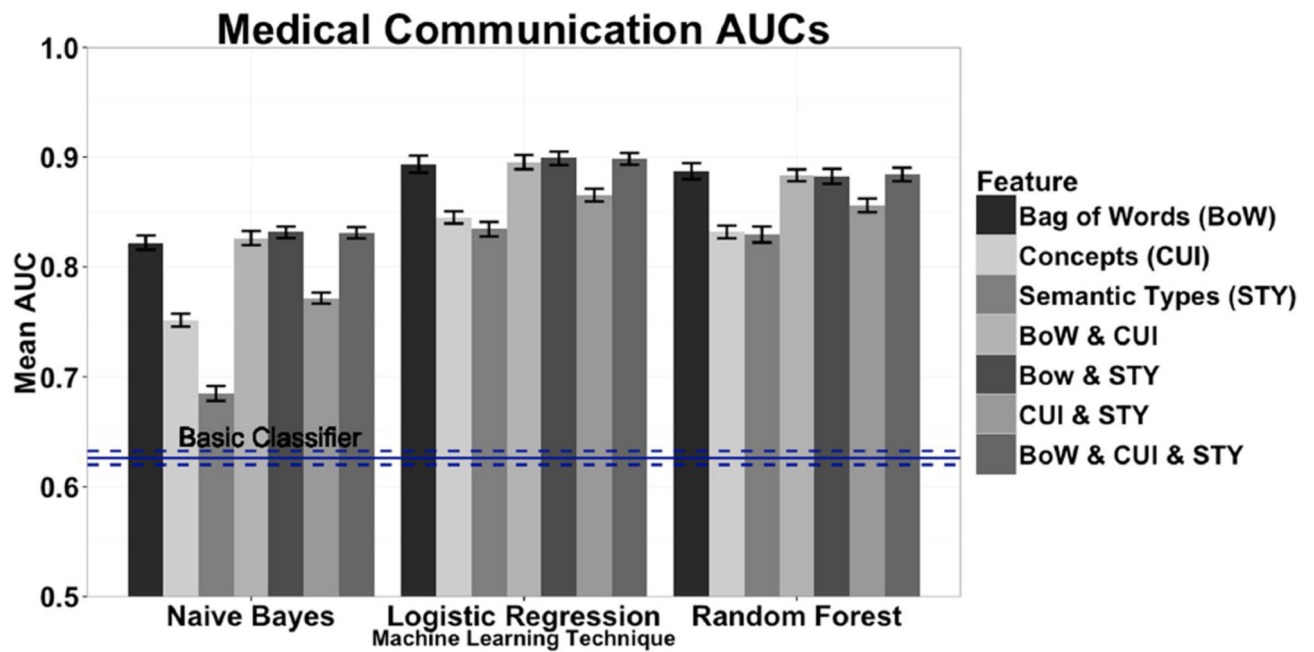
Figure 1.
The taxonomy of consumer health information communication types[17, 33, 34].

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:

- *Informational*
- Medical
- Social
- Logistical

Figure 2.
Example message labeled by communication types



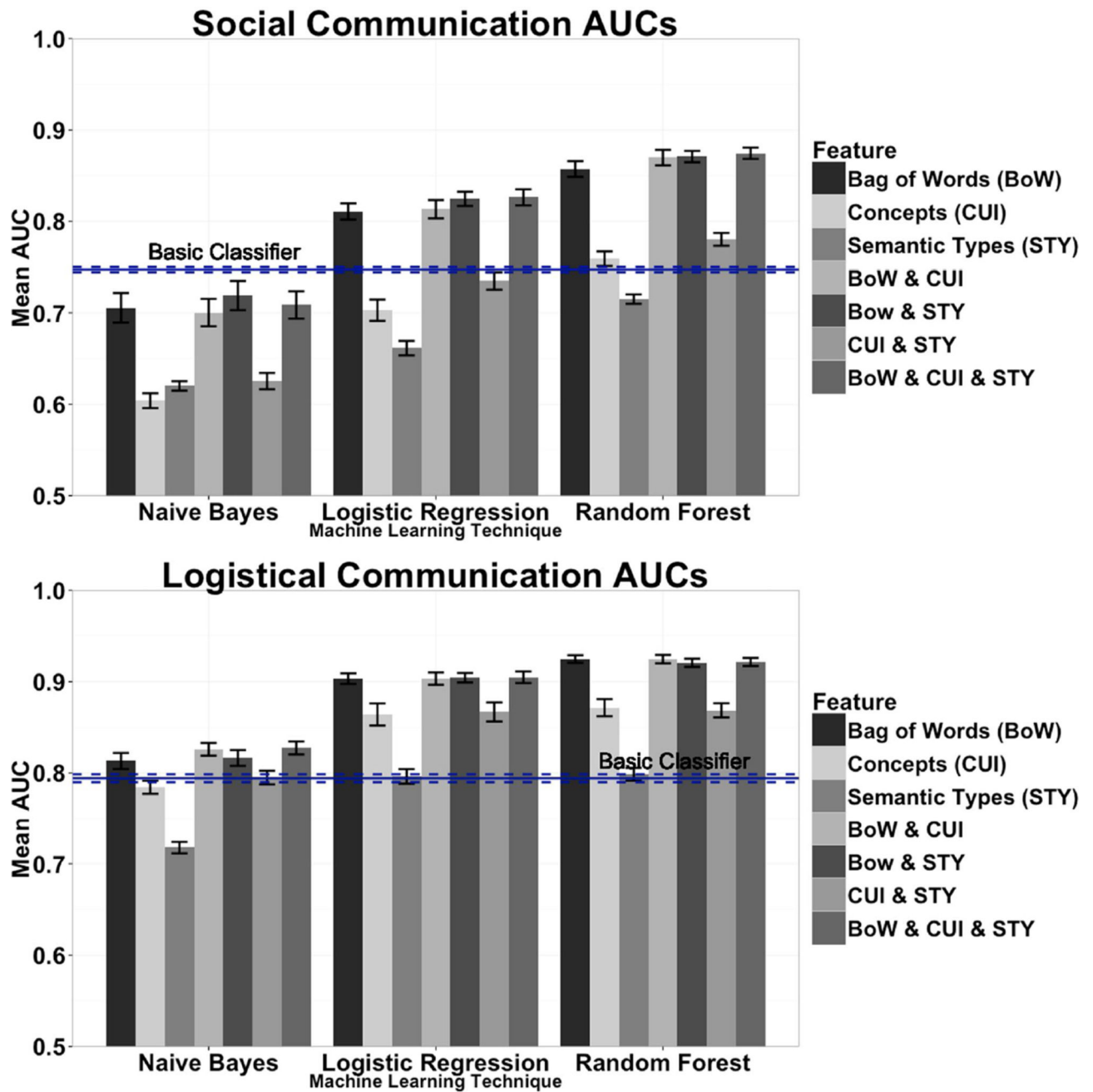


Figure 3. Area under the curve (AUC) of the different major communication types. The Basic Classifier was the Rule Based classifier. The error bars represent the 95% Confidence Interval.

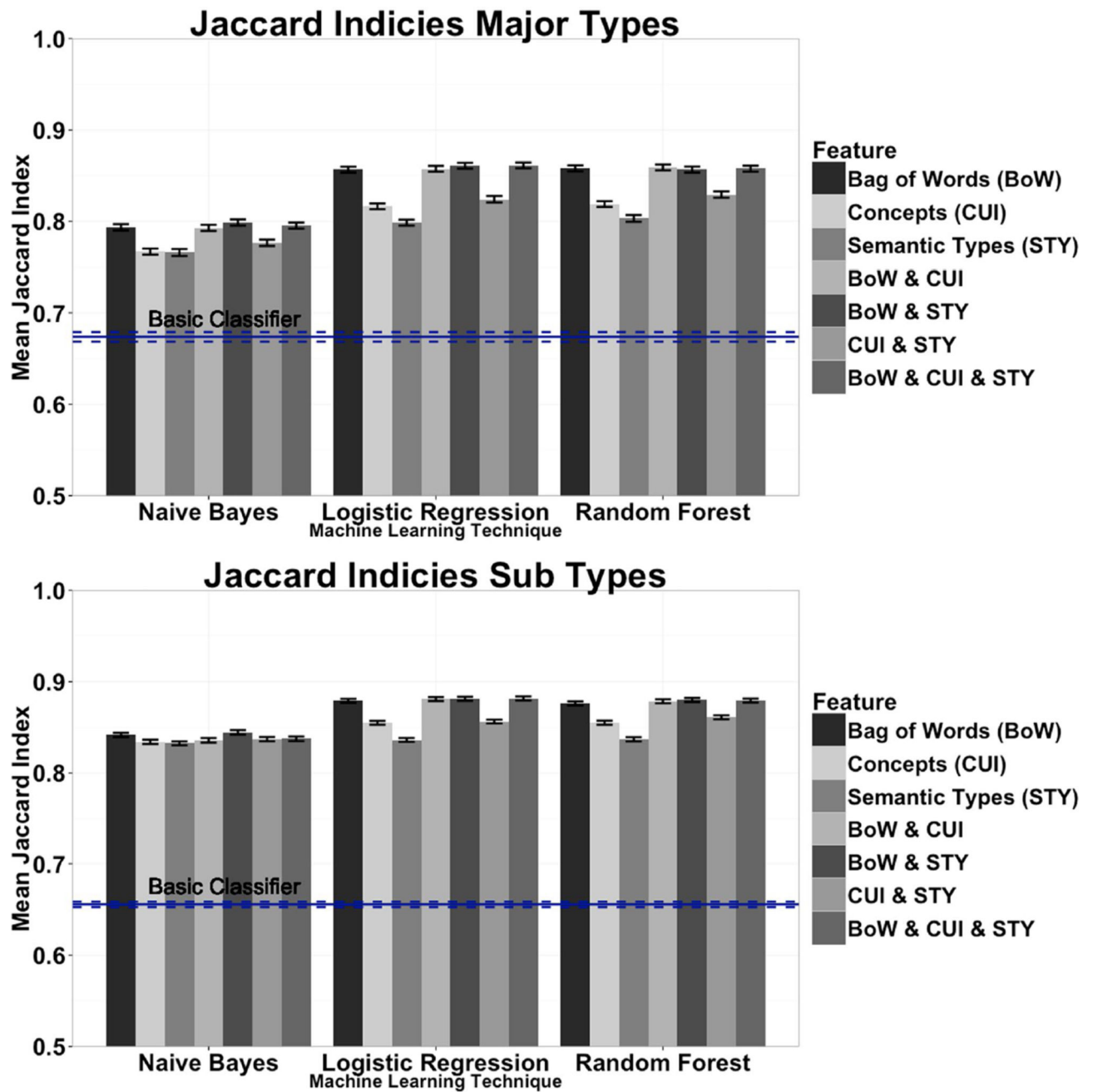


Figure 4. Bar charts of the Jaccard Indices of the different communication types. The Basic Classifier was the Rule Based classifier. The error bars represent the 95% Confidence Interval.

Table 1

Examples of communication types

Major Type	Subtype	Definition
Informational		Communications involving basic medical knowledge about problems, tests, interventions, and general management of health concerns
Logistical		Communications involving pragmatic issues
Logistical	Contact Information	Communications about 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		Communications related to social interactions or an interpersonal relationship that is not directly related to informational needs, medical needs, or logistical needs
Social	Acknowledgement	Expressions of gratitude or satisfaction or acknowledgement or agreement
Medical		Communications expressing a desire for medical care or the delivery of such care. This category is distinguished from <i>informational communications</i> 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
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	Communications 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

Table 2

Words used to determine if a message belongs to one of the major or sub communication types for the basic (rule-based) classifier.

Communication Major Type		Words
Informational		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
Major Type	Subtype	Words
Logistical	Contact information	fax, phone, telephone, cell, address, street, email
Social	Acknowledgement	appreciate, time, very much
Medical	Appointments	call me, appointment, be seen, (Mon, Tues, Wed, Thurs, Fri)-day
Medical Medical	Prescriptions Problems	refill, prescription pain, worse
Medical	Follow-up	better, follow up
Medical	Tests	lab, labs, ultrasound, CT, MRI, test

Table 3

Top 10 features for random forests. Dark grey boxes contain concepts (CUIs), and light grey 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	Informational	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

Table 4

Semantic types (STYs) and concepts (CUIs) and the words associated with them

Semantic Type (STY)	words, phrases, or abbreviations
Finding	“lab work”, “latest test results”
Qualitative_Concept	“critical”, “high”
Pharmacologic_Substance	“medication”
Conceptual_Entity	“phone”, “doctor”, “fax”
Intellectual_Product	“name”
Temporal_Concept	“morning”, “day”, “evenings”, “Friday”
Ideaor_Concept	“appointment”, “refill”
Quantitative_Concept	“mg”
Intellectual_Product	“prescription”

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript