Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.JournalofSurgicalResearch.com

Association for Academic Surgery

Automating the Classification of Complexity of Medical Decision-Making in Patient-Provider Messaging in a Patient Portal

Lina Sulieman, PhD,[a,*] *Jamie R. Robinson, MD,*[a,b] *and Gretchen P. Jackson, MD, PhD*[a,b]

[a] *Vanderbilt University Medical Center, Nashville, Tennessee*
[b] *IBM Watson Health, IBM, Cambridge, Massachusetts*

ARTICLE INFO

Article history:
Received 31 January 2020
Received in revised form
9 April 2020
Accepted 5 May 2020
Available online 19 June 2020

Keywords:
Patient portals
Medical decision
Decision complexity
Machine learning
Medical billing
Consumer health

ABSTRACT

Background: Patient portals are consumer health applications that allow patients to view their health information. Portals facilitate the interactions between patients and their caregivers by offering secure messaging. Patients communicate different needs through portal messages. Medical needs contain requests for delivery of care (e.g. reporting new symptoms). Automating the classification of medical decision complexity in portal messages has not been investigated.

Materials and methods: We trained two multiclass classifiers, multinomial Naïve Bayes and random forest on 500 message threads, to quantify and label the complexity of decision-making into four classes: no decision, straightforward, low, and moderate. We compared the performance of the models to using only the number of medical terms without training a machine learning model.

Results: Our analysis demonstrated that machine learning models have better performance than the model that did not use machine learning. Moreover, machine learning models could quantify the complexity of decision-making that the messages contained with 0.59, 0.45, and 0.58 for macro, micro, and weighted precision and 0.63,0.41, and 0.63 for macro, micro, and weighted recall.

Conclusions: This study is one of the first to attempt to classify patient portal messages by whether they involve medical decision-making and the complexity of that decision-making. Machine learning classifiers trained on message content resulted in better message thread classification than classifiers that employed medical terms in the messages alone.

* *Corresponding author*. Department of Biomedical Informatics, 2525 West End - Suite 1500, Nashville, TN 37203. Tel.: 615-414-3988.
E-mail address: Sulieman.lina@gmail.com (L. Sulieman).
0022-4804/$ – see front matter © 2020 Elsevier Inc. All rights reserved.
https://doi.org/10.1016/j.jss.2020.05.039

Introduction

Patient portals are secure online applications that allow healthcare organizations to provide patients and their caregivers access to health information including medications, immunizations, and appointments.[1-3] Many patient portals offer a secure messaging function that enables patients to interact with providers through messages.[3,4] Secure messaging is one of the most popular features of patient portals, and messaging volumes are growing exponentially, and one study showed that surgery was second only to medicine in number of messages exchanged[5-7] As the use of messaging increases, techniques to automate the analysis of messages may be critical to assist with triage, message answering, or quantifying the care delivered through patient portals.

Research about portal messages has mainly focused on qualitative analyses of content, with few studies investigating automated classification.[8,9] North *et al.* analyzed messages exchanged between providers and patients from a primary care clinic at an academic medical center and found that 3.5% of messages include potential high-risk symptoms. Jackson *et al.* have developed and validated a taxonomy of consumer health communications and have applied it to questions from patients and caregivers in research and inpatient and patient portal messages.[9-15] The taxonomy comprehensively describes the semantic types of consumer health communications including informational, medical, logistical, social, and other. It has been employed to characterize the content of consumer health questions (i.e. needs) as well as the answer to those questions. Informational needs are questions that require clinical knowledge such as information about the side-effects of a drug. Medical needs are requests for delivery of medical care, such as the reporting of new symptoms that require management. Logistical needs are questions involving pragmatic issues such as the phone number of a clinic. Social needs are interpersonal communications such as emotional concerns, expressions of gratitude, or complaints. The other category covers content that does not fit into these four categories (e.g. questions that transcend categories like "how do I be a good father" or error messages). Cronin *et al.* investigated the use of machine learning to classify the content of portal messages.[15] Sulieman *et al.* trained convolutional neural networks and standard machine learning algorithms (e.g. random forest) to identify the types of needs in portal messages.[16]

Portal messages that include medical needs (e.g. time-sensitive clinical questions or information reflecting changes in patient status) are of particular importance.[17] One analysis of the content of 3253 patient portal messages from a large academic medical center showed that 72% included medical needs.[15] Answering those types of messages might involve medical decision-making such as changing a drug or ordering a test. As messaging volumes grow, the identification of messages that require medical decision-making may be essential.[6,7,18]

In this study, we trained machine learning classifiers to identify patient portal message threads that involved clinical decision-making and to classify the complexity of the decision as no decision, straightforward, low, or moderate. This study specifically focused on portal messages exchanged between surgical patients and surgeons as the majority of research on patient portals has been done in medicine and primary care. We investigated the effectiveness of machine learning by comparing the performance of our models to using a medical term extraction tool, Clamp. If effective, such automated message analysis might quantify the care delivered online or support billing for online care.

Materials and methods

We conducted the study at Vanderbilt University Medical Center, a private nonprofit institution with 137 outpatient locations and over two million patient visits annually. In 2004, Vanderbilt University Medical Center launched My Health At Vanderbilt (MHAV), a patient portal that offers common portal functions such as access to portions of the electronic health record, appointment scheduling, and tailored clinical information. Secure messaging is one of the commonly used features of MHAV. On average, patients send over 30,000 messages each month. Clinical care teams including administrative assistants, nurses, and physicians typically manage the messages. Clinicians can answer messages directly or delegate message answering to nurses and medical assistants.

Study data set

This study employed MHAV message threads (i.e. sets of messages exchanged between patients and surgical providers) and annotated them with communication categories from the consumer health taxonomy as well as the complexity of medical decision of the exchange. Two researchers (who were both surgeons) independently labeled 500 message threads with taxonomy categories and complexity of medical decision-making and discussed all disagreements to achieve consensus; details of the data set creation are published elsewhere.[9] The complexity of medical decision-making is one of outpatient billing elements according to the guideline defined by the Center of Medicare and Medicaid Services Evaluation and Management coding criteria.[19,20] The complexity of medical decision-making is quantified based on three factors: the amount of data reviewed, diagnoses, and risk, summarized in Table 1.[9] This data set did not contain any message threads with a high level of medical decision complexity.

Machine learning algorithms

We trained different classifiers to assign four different labels for each thread: "no decision" for messages that did not involve decision-making, straightforward, low, and moderate. To assign labels, we extracted three text features from the message threads:

1 Bag of words: We extracted the words from each message thread after removing the stop words and non-alphabetical characters. To represent the messages, we

Table 1 – Factors of medical decision-making.

Diagnosis or management options	Amount and complexity of data	Level of risk of complications	Complexity of decision-making*
Minimal	Minimal or none	Minimal	Straightforward
Limited	Limited	Low	Low
Multiple	Moderate	Moderate	Moderate
Extensive	Extensive	High	High

*Complexity of decision-making is based on the three categories listed in the first three columns. At least 2 or 3 of those factors should be met.

created a numerical vector, where each value corresponded to the number of times a word was mentioned in the thread.

2 Term frequency-inverse document frequency (TF-IDF): TF-IDF is a scoring system that assigns weights for words based on frequency in a document relative to all documents in the data set.[21] TF-IDF weights focus on term frequency and undervalued or rare words. We used Sklearn to transform the bag of words vectors into TF-IDF vectors.

3 Medical terms: We used Clamp (version 1.5.0) to extract medical terms from the message threads and used them as features.[22] We represented each message by the medical terms included in that thread.

We applied three different algorithms to predict the complexity of medical decision-making in the message.

1 Using medical terms only: We established a baseline by using only the number of medical terms in the message threads to assign the complexity label. We labeled message threads using the number of medical terms as follows:
   i. "No decision" threads included one or no medical terms
   ii. "Straightforward" decision threads included two or three medical terms
   iii. "Low" decision threads included four medical terms
   iv. "Moderate": decision threads included more than five
2 Multinomial Naïve Bayes: We trained a multiclass Multinomial Naïve Bayes model to predict the decision class for the thread using three text features: bag of words, TF-IDF, and medical terms.
3 Random forest: we trained a multiclass random forest model to predict the decision label of each thread.

Training and evaluation

We split the data set into three sets: 90% for training and validation and 10% for testing. We used 10-fold cross training-validation data set to tune the parameters and select the parameter set of the optimal model. Table 2 lists the parameter space that we searched to identify the parameters of the optimal model. We defined the optimal model as the model that had the highest evaluation metric on the validation set. Because this problem involved multiclass labeling (i.e. more than two classes), we selected two evaluation metrics to identify the optimal model: micro precision and micro recall. The micro metric calculates the precision/recall for each class and finds the weighted average precision/recall based on the number of samples per class which account for class imbalance. In our analysis, we focused on precision and recall (rather than area under the curve) because we wanted to evaluate the model for clinical implementation and thus, wanted to quantify true and false positives and negatives precisely.

False negatives represent message threads that were assigned to different decision labels and in some cases no decision. Mislabeling the thread as no decision can have a higher penalty than mislabeling the complexity of message thread as a different level of complexity. Hence, we analyzed the percentage of threads that the classifier mislabeled as "no decision" along with other mislabeling percentages. We calculated the precision for each individual label, the macro precision, and the micro precision by calculating the total true positives, false negatives and false positives, and the weighted precision by calculating the weighted average of precision by their prevalence in the evaluated data set. We calculated the same metrics for recall. Finally, we created a confusion matrix for each model to identify the percentage of mislabeled message threads for each class.

Table 2 – The parameter search space used to find optimal model for Naive Bayes and random forest.

Machine learning algorithm	Parameter name	Parameter possible values
Multinomial Naïve Bayes	Smoothing parameter alpha	0-1, 0 no smoothing
	Fitting prior: Learning class prior probabilities	True, false. If false, uniform prior is applied
Random forest	Number of estimators	20,30,40,50,60,70,80,90,100
	Maximum depth	None (nodes expanded until all leaves are pure),3,4,5,6,7,8,9,10
	Minimum samples split	2,3,4,5
	Maximum leaf nodes	5,6,7,8,9,10,11,12

Table 3 – The precision values of machine learning models tuned on precision value.

Class	No machine learning	Multinomial Naïve Bayes			Random forest		
	Medical term number	Bag of word	TF-IDF	Medical terms	Bag of word	TF-IDF	Medical terms
No decision	0	**0.67**	0.62	0.62	0.53	0.5	0
Straightforward	0.4	**0.67**	0.65	0.57	0.6	0.54	0.49
Low	**0.67**	0.2	0.29	0.6	0	0	0
Moderate	**0.04**	0	0	0	0	0	0
Micro average	0.14	0.57	**0.59**	0.57	0.55	0.53	0.49
Macro average	0.28	0.38	0.39	**0.45**	0.28	0.26	0.12
Weighted	0.31	**0.58**	0.57	**0.58**	0.47	0.43	0.24

Bold text represents the highest precision for the identifying the corresponding medical decision complexity class.

Results

In the collection of 500 portal message threads from surgical patients or caregivers, 339 (67.8%) threads involved medical decision-making. In those threads, 210 (62%) contained straightforward decisions; 102 (30%) threads contained low complexity decision-making; and 27 (8%) threads involved moderate complexity decision-making. Most commonly expressed medical need was scheduling of an appointment in 42% of the threads. Among those threads, 163 (32.6%) reported new or worsening problems, and 139 (27.8%) involved prescriptions, and 212 (42.4%) included need for appointment scheduling.

Tables 3 and 4 show values of the precision and recall performance metrics for the optimal machine learning models. Table 5 lists the parameters of optimal models. Table 3 summarizes the precision values for the optimal models that we trained on different text features and tuned using precision. Table 4 lists the parameters of optimal models trained in bag of words, TF-IDF, and medical terms. Using medical terms only without machine learning demonstrated generally poor performance. Predicting the type of decision-making using only the number of medical terms had a precision of 0.67 for low complexity, a precision of 0.04 for moderate complexity, and recall of 1.0 for the moderate complexity. Moreover, using the number of medical terms did not identify any of the threads that did not involve a decision.

Applying any machine learning model on the medical terms yielded higher precision for messages that did not require decision-making ranging from 0.50 to 0.67 and for straightforward decision-making from 0.49 to 0.67 *versus* not using machine learning ranging that had the values 0 and 0.4 for "no decision" and straightforward, respectively. We observed similar results for recall. Machine learning models had higher recall values than using only the number of medical terms by 0.5 to 0.62 for threads that did not involve any decision-making and by 0.31 to 0.71 for threads that involved straightforward decision-making. Moreover, applying machine learning classification models on medical terms had higher micro, macro, and weighted values for both precision and recall. The micro, macro, and weight average precision values for methods using only the number of medical terms were 0.14, 0.28, and 0.31 respectively. The micro, macro, and weighted recall values for the same model were 0.14, 0.35, and 0.14.

Applying machine learning models improved the identification of threads that involved decision-making. For the models that we tuned using precision for classifying decision-making complexity, the optimal multinomial Naïve Bayes model had higher precision values than the optimal random forest model. The precision values were 0.12, 0.14, and 0.62 when we trained the two models on TF-IDF, bag of words, and medical terms, respectively, where multinomial Naïve Bayes yielded the higher values. As shown in Table 3, the precision for detecting messages that did not require decision-making

Table 4 – The recall values of the machine learning models tuned on recall values.

Class	No machine learning	Multinomial Naïve Bayes			Random forest		
	Medical term number	Bag of word	TF-IDF	Medical terms	Bag of word	TF-IDF	Medical terms
No decision	0	**0.62**	**0.62**	**0.62**	0.5	0.5	0
Straightforward	0.17	0.67	0.71	0.57	**0.88**	0.79	1
Low	0.25	0.25	0.25	**0.6**	0.25	0.12	0
Moderate	**1**	0	0	0	0	0	0
Micro average	0.14	0.57	0.59	0.57	**0.63**	0.57	0.49
Macro average	0.35	0.39	0.4	0.45	**0.41**	0.35	0.16
Weighted	0.14	0.57	0.59	0.58	**0.63**	0.57	0.32

Bold text represents the highest precision for the identifying the corresponding medical decision complexity class.

Table 5 – The parameters of the optimal model tuned on both precision and recall.

Machine learning algorithm	Multi Naïve Bayes	Random forest
Tuning based on precision		
Bag of words	Alpha: 0.9, fit prior: True	Maximum depth: 9, maximum leaf nodes: 9, minimum samples split: 4, estimators number: 100
TF-IDF	Alpha: 0.5, fit prior: False	Maximum depth: None, maximum leaf nodes: 12, minimum samples split: 2, estimators number: 30
Medical terms	Alpha: 0.7, fit prior: True	Maximum depth: 8, maximum leaf nodes: 10, minimum samples split: 3, estimators number: 20
Tuning based on recall		
Bag of words	Alpha: 0.7, fit prior: True	Maximum depth: 6, maximum leaf nodes: 9, minimum samples split: 3, estimators number: 90
TF-IDF	Alpha: 0.5, fit prior: False	Maximum depth: 8, maximum leaf nodes: 12, minimum samples split: 2, estimators number: 20
Medical terms	Alpha: 0.9, fit prior: True	Maximum depth: 8, maximum leaf nodes: 12, minimum samples split: 2, estimators number: 20

was between 0.62 and 0.67 when we trained the multinomial Naïve Bayes model. We obtained the highest precision value for detecting those threads when we used bag of words as features. Multinomial Naïve Bayes model identified messages with low complexity decision-making with the precision values 0.20, 0.29, and 0.60, when we trained the model on Bag of words, TF-IDF, and medical terms, respectively. Both multinomial Naïve Bayes and random forest models did not identify the one thread that contained moderate decision-making. For multinomial Naïve Bayes model, using medical terms had 0.45 and 0.58 for the macro average and weighted average precision, respectively, which were higher than performance metrics for the same model trained on bag of words and TF-IDF. Training multinomial Naïve Bayes on TF-IDF had average micro precision of 0.58, which is the highest overall and the highest values for multinomial Naïve Bayes model trained on the other two features.

Tuning the models on the recall values had slightly different results. Multinomial Naïve Bayes had the highest recall overall for identifying threads that did not require decision-making with a recall of 0.62, which was higher by 0.12 than the recall yielded by random forest. Training the multinomial Naïve Bayes model to identify threads that required low decision-making had the highest recall value, which was 0.60. Random forest trained on bag of words and TF-IDF to identify threads with straightforward decision complexity had higher recall values compared with multinomial Naïve Bayes models trained on the same text features. Moreover, the random forest model trained on bag of words had the highest micro, macro, and weighted recall values.

The ability to identify the complexity of decision-making in a thread also depended on the text features used to train the models. Training the model on bag of words yielded the highest values of 0.67 for precision, and 0.62 and 0.88 for recall for identifying "No decision" and straightforward classes. Using the medical term to identify threads that contained low to moderate decision-making had the highest precision values (0.67 and 0.04), the highest recall values (0.6 and 1). For the aggregated macro and micro metrics, training the models on TF-IDF had the highest micro precision, while training the models on medical terms had the highest macro and weighted precisions. Training the models on bag of words had the highest macro, micro, and weighted recall.

Figures 1 and 2 show the confusion matrices for the models tuned on precision and recall, respectively. The confusion matrices specify the rates of true classifications and misclassifications with respect to the other classes. Each row represents the true positives and false negatives for the corresponding classes. For example, the second row details the classification of threads that involved straightforward decision-making. The first, third, and fourth columns represent the straightforward threads that the model classified as no decision, low complexity, and moderate complexity. All cells on the diagonal represent the messages that were classified correctly also known as true positive rate or recall. Darker colors correspond to higher values. The confusion matrices for the models tuned on precision are depicted in Figure 2A-F. When we tuned the model using the precision, training Multinomial naïve Bayes on bag of words or TF-IDF had the 0.62 which was the highest recall. The recall for message threads with straightforward decision-making was the highest when we trained random forest and ranged between 0.79 and 1.0, where training random forest model on medical terms achieved a complete identification of straightforward complexity decision-making. The Multinomial naïve Bayes classifier trained on medical terms identified 0.38 of threads classified as requiring low complexity decision-making correctly, which was the highest recall among all models.

The models misclassified the message threads that contained moderate complexity decision-making. For all the models trained on three different text features, the threads that did not contain any decision-making were most commonly mislabeled as straightforward with rates between 0.19 and 0.62 and random forest trained on medical terms mislabeled all of them as straightforward (percentage of labeling "no decision" as straightforward = 1, see confusion matrix in Fig. 1F, the second cell in the first row). The straightforward message threads were typically mislabeled as no decision with a rate of misclassification ranging between

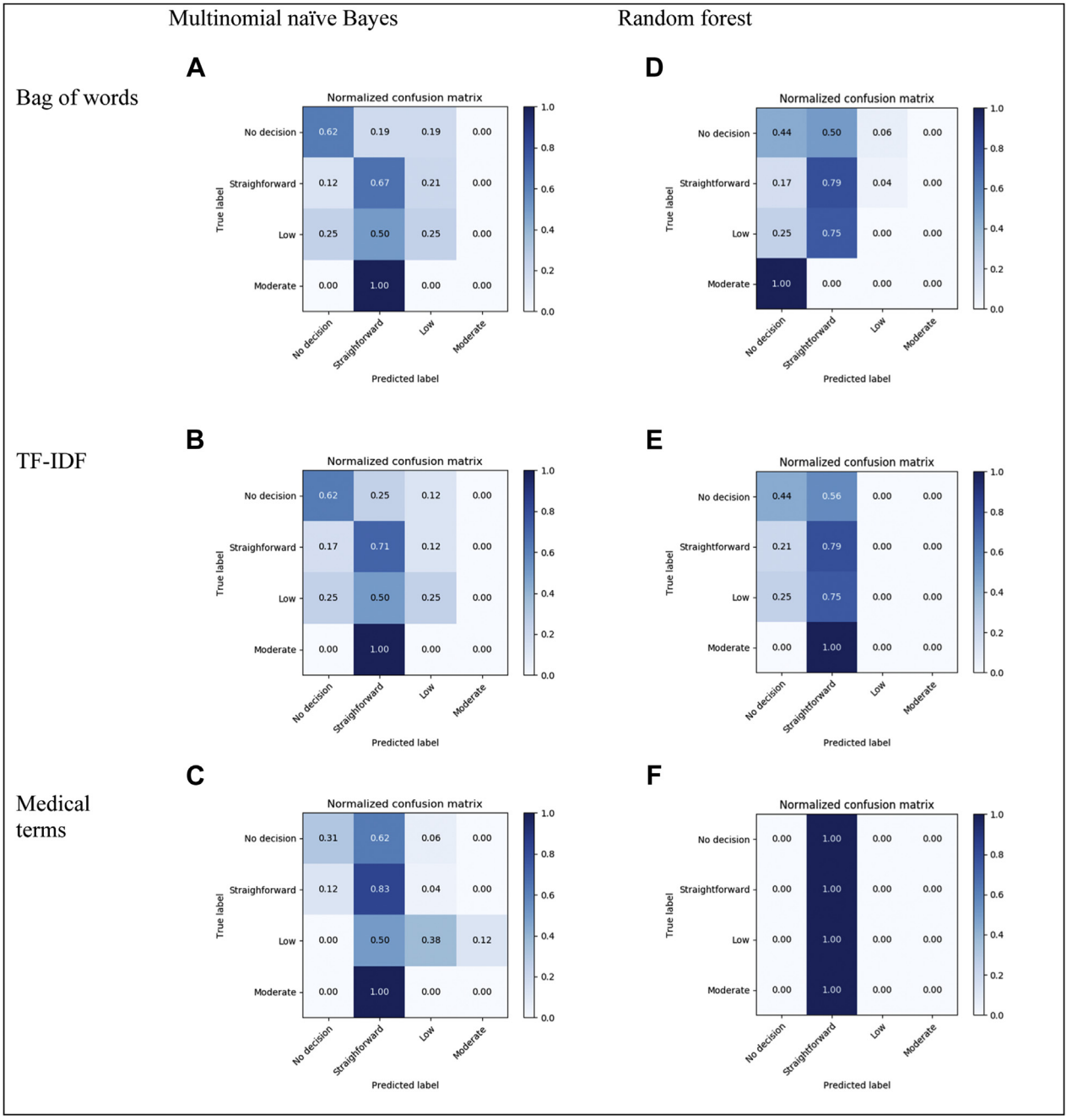

Fig. 1 – The confusion matrices for models tuned on precision. (A) confusion matrix of Naive Bayes trained on bag of words. (B) confusion matrix of Naive Bayes trained on TF-IDF. (C) confusion matrix of Naive Bayes trained on medical terms. (D) confusion matrix of random forest trained on bag of words. (E) confusion matrix of random forest trained on TF-IDF. (F) confusion matrix of random forest trained on medical terms. The x-axis is the model's predicted decision complexity class, the y-axis is the true decision complexity class. The darker color corresponds to a higher precision value. The brighter color corresponds to a lower precision value. (Color version of figure is available online.)

0.12 and 0.17 in all models except random forest trained on medical terms. The threads in the low complexity class were mainly mislabeled as straightforward with 0.5 to 0.75 except random forest trained on medical terms that mislabeled all low complexity threads as straightforward (confusion matrix Fig. 1F, third cell in the third row). Figure 2A-C depicts the confusion matrices for Multinomial Naïve Bayes trained on bag of words, TF-IDF, and medical terms. Figure 2D-F depict the confusion matrices for random forest trained on bag of words, TF-IDF, and medical terms, respectively.

The correct classifications and misclassifications were slightly different for the models tuned based on recall (Table 4). The highest recall values among all models were 0.62 for the "no decision" class, 1 for straightforward, and 0.38 for low complexity when we trained multinomial Naïve Bayes on bag of words or TF-IDF, random forest on medical terms, and

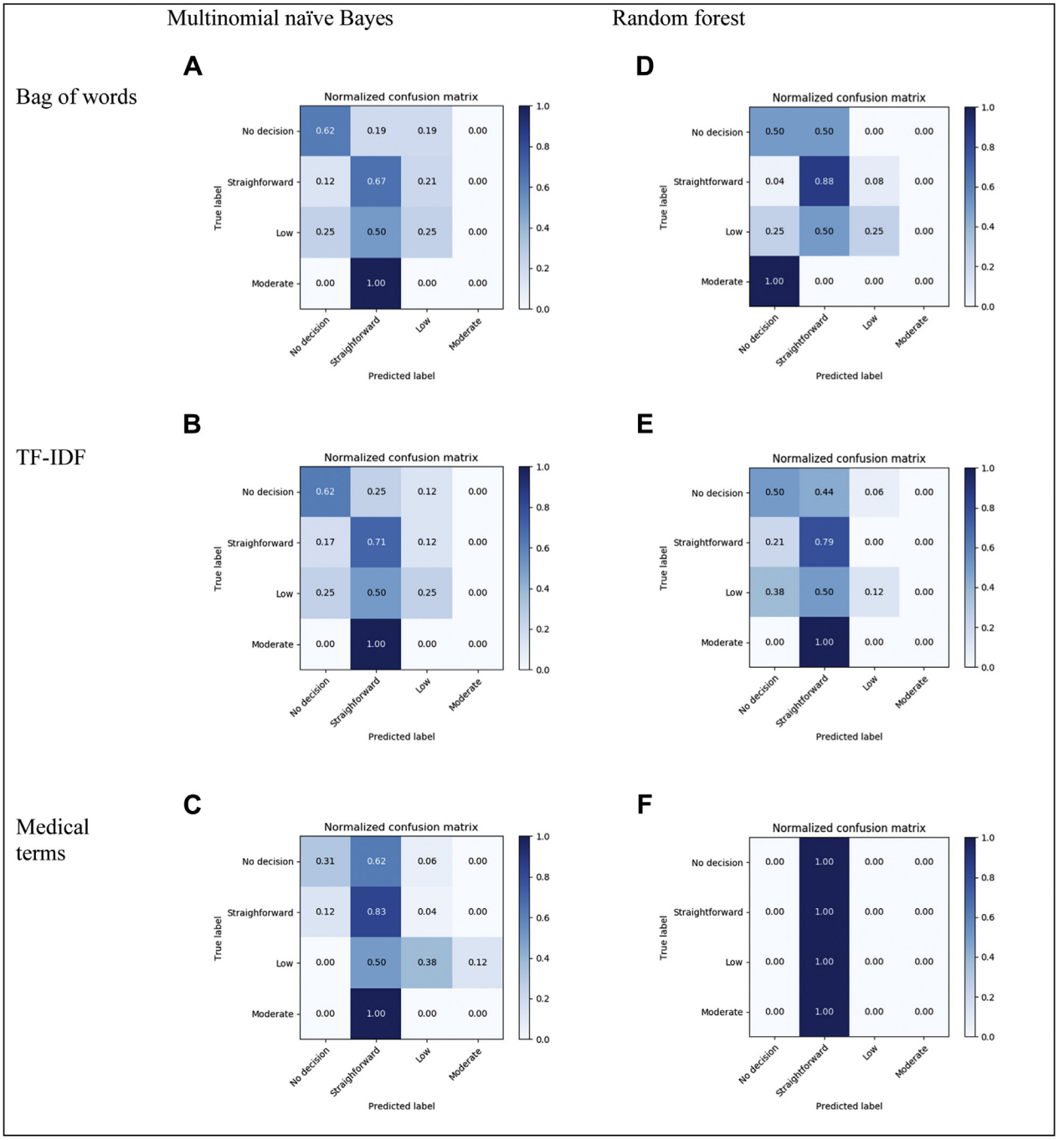

Fig. 2 – Confusion matrices for models tuned using recall. (A) confusion matrix of Naive Bayes trained on bag of words. (B) confusion matrix of Naive Bayes trained on TF-IDF. (C) confusion matrix of Naive Bayes trained on medical terms. (D) confusion matrix of random forest trained on bag of words. (E) confusion matrix of random forest trained on TF-IDF. (F) confusion matrix of random forest trained on medical terms. The x-axis is the model's predicted decision complexity class, the y-axis is the true decision complexity class. The darker color corresponds to a higher recall value. The brighter color corresponds to a lower recall value. (Color version of figure is available online.)

multinomial Naïve Bayes on medical terms, respectively. The no decisions message threads were mainly misclassified as straightforward with 0.19 as the lowest misclassification rate for multinomial naïve Bayes trained on bag of words and 1.0 as the highest misclassification rate for random forest trained on medical terms. The message threads that contained straightforward decisions had higher misclassification rates when we trained random forest, and multinomial naïve Bayes on bag of words with rates ranged between 0.08 and 0.21. While training the machine learning model on TF-IDF resulted in misclassifying straightforward message threads as "no decision" with rates 0.17 and 0.21 for multinomial naïve Bayes

and random forest. The threads in the low complexity class were misclassified as straightforward with 0.50 misclassification rate for most of the models.

Discussion

This study is one of the first attempts to automatically classify patient portal message threads exchanged between surgeons and patients based on the complexity of medical making decision within the message exchanges. To our knowledge, it is the first study to implement machine learning models to identify message threads that involved medical decision-making from a healthcare provider and to classify the complexity of the decisions. It is well established that medical coding criteria are applied inconsistently in practice,[23,24] and the annotation of the data set for this study required careful analysis and discussion to achieve consensus to create a high-quality gold standard. Our analysis shows that using tools that only extract the medical terms such as KnowledgeMap, cTake, or Clamp were not efficient in quantifying medical decision-making.[22,25,26] Machine learning models improved the classification of patient portal message threads based on the complexity of medical decision-making. Automating the classification of individual patient messages may aid with triaging those messages that need the attention of a healthcare provider who can respond and deliver the appropriate care. Analyzing the content of message threads has the potential to support automated billing for online encounters. However, to realize these applications, the performance of these classifiers will need to be improved. This manuscript provides some initial evidence on which approaches may be most effective.

To evaluate the effectiveness of implementing the proposed classifier in clinical settings, we focused on precision and recall metrics. Obtaining a precise model to identify message threads that do not involve medical needs or decision-making could aid in triage to administrative assistants or allied health professionals. Our analysis demonstrated that machine learning models could accurately identify the message threads that do not involve medical decision-making or contain straightforward decisions that have minimal complexity. Such messages might be triaged to administrative assistants, nurses, or allied health professionals, allowing physicians to focus more time on messages requiring more complex medical decisions.

Developing a machine learning model with high recall could support the identification of threads with higher complexity medical decision-making, which could potentially be valuable for healthcare administrators in quantifying the care being delivered by providers online and potentially supporting automated coding of online outpatient encounters, should reimbursement be supported. Message threads that have straightforward to moderate decision-making can include new symptoms or clinical problems, which are managed. For instance, a message thread with low complexity involved the patient reporting the lack of sleep because of ache in muscle and joints despite taking Trazadone. Another message thread with straightforward complexity included a request from a patient for a referral to a dietitian after experiencing digestive problems. Our analysis demonstrated that machine learning classifiers could identify the message threads that do not contain decision-making, which could potentially facilitate appropriate triage. Moreover, our machine learning models were able to identify threads that involved straightforward and low complexity decision-making with recall higher than 0.60 and weighted recall for all classes higher than 0.55.

Although payers do not yet reimburse for delivering care through patient portals, there are various benefits to identifying message threads that involve care delivery. One important use might be quantifying the care delivered online by various types of clinical providers to plan for appropriate staffing. Documenting the volumes of care delivered online might also support the case for reimbursement of such care. Managing the low complexity issues using patient portal messaging can benefit patients, providers, and healthcare organizations. Providing online care can save patients time and money by reducing the number of unnecessary visits to clinics or hospitals, which can be a burden if the patient lives far from a medical center or if the appointments are canceled or healthcare systems transition to telehealth because of a pandemic such as COVID-19. For surgeons, online post-operative care is particularly advantageous for procedures that typically have an uncomplicated course. Further, managing low complexity care online can make available clinic appointments for higher complexity medical needs, and this availability benefits the medical center, allowing them to most effectively utilize their resources.

Our study has limitations. First, we used portal message threads from a single academic medical center data using a locally developed patient portal for analysis, and our results may not translate to other settings. Our center has transitioned to a popular commercial patient portal, so future analyses may provide a more generalizable result. Second, our data set is small and only contains message threads initiated by patients and sent to surgical providers. Although the message threads were sent to a wide variety of surgical specialties, the language used in portal messages about surgical disease might be significantly different from the content of portal messages involving other specialties. The manual annotation process is laborious, making the creation of large labeled data sets challenging. We are implementing a semi-supervised machine learning model that can leverage this data set to quantify the care delivered in other threads that have not been annotated or labeled. In our future work, when we have a larger annotated data set, we can expand our features such as combining TF-IDF and bag of words or using word embedding and deep learning classification (e.g. a convolutional neural network). Feature expansion with the existing data would risk overfitting. Third, we did not extract the lay terms that patients or caregivers might use in portal messages. In the message threads we analyzed, lay language was often restated in medical terms in the provider responses. Identifying lay or slang language in consumer health messages is a hot topic in medical informatics and natural language processing and is an area of future research for our team. Finally, we tried only three methods because of the data set size. Training and evaluating other machine learning model such as convolutional neural network another possible

approach to improve classification performance when we have a larger data set.

Conclusion

Patient portals are popular consumer health applications that allow patients and their caregivers to interact using secure messaging. The adoption of secure messaging is increasing, and studies have shown that medical care of varying complexity is delivered through patient portal message threads exchanged with surgical providers. This study is one of the first to attempt to classify patient portal messages by whether they involve medical decision-making and the complexity of that decision-making. Machine learning models that analyzed content resulted in better message thread classification than classifiers that employed medical terms in the messages alone. Further research is needed to improve the performance of these classifiers to potentially support triage of portal messages or quantification of online care to inform staffing needs or to support reimbursement for online care.

Acknowledgment

The main author (third author) works at IBM Watson Health.

Authors contributions: First author performed the analysis and wrote the paper. Second author prepared the data set and edited the paper. Third author supervised the analysis and edited the paper.

Disclosure

The first and second authors have nothing to disclose.

REFERENCES

1. HealthIT.gov. What is a patient portal?. Accessed December 1, 2019.
2. Otte-Trojel T, De Bont A, Van De Klundert J, Rundall TG. Characteristics of patient portals developed in the context of health information exchanges: early policy effects of incentives in the meaningful use program in the United States. *J Med Internet Res*. 2014;16:e258.
3. Osborn CY, Rosenbloom ST, Stenner SP, et al. MyHealthAtVanderbilt: policies and procedures governing patient portal functionality. *J Am Med Inform Assoc*. 2011;18(Suppl 1):i18–i23.
4. HealthIT.gov. What is a patient portal?. http://www.healthit.gov/providers-professionals/faqs/what-patient-portal; 2016 [accessed 05.12.16].
5. Kruse CS, Bolton K, Freriks G. The effect of patient portals on quality outcomes and its implications to meaningful use: a systematic review. *J Med Internet Res*. 2015;17:e44.
6. Cronin RM, Davis SE, Shenson JA, Chen Q, Rosenbloom ST, Jackson GP. Growth of secure messaging through a patient portal as a form of outpatient interaction across clinical specialties. *Appl Clin Inform*. 2015;6:288–304.
7. Shenson JA, Cronin RM, Davis SE, Chen Q, Jackson GP. Rapid growth in surgeons' use of secure messaging in a patient portal. *Surg Endosc Other Interv Tech*. 2016;30:1432–1440.
8. North F, Crane SJ, Stroebel RJ, Cha SS, Edell ES, Tulledge-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.
9. Robinson JR, Valentine A, Carney C, Fabbri D, Jackson GP. Complexity of medical decision-making in care provided by surgeons through patient portals. *J Surg Res*. 2017;214:93–101.
10. Cronin RM, Fabbri D, Denny JC, Rosenbloom ST, Jackson GP. A comparison of rule-based and machine learning approaches for classifying patient portal messages. *Int J Med Inform*. 2017;105:110–120.
11. 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;2015:1148–1156.
12. Lee DJ, Cronin R, Robinson J, et al. Common consumer health-related needs in the pediatric hospital setting: lessons from an engagement consultation service. *Appl Clin Inform*. 2018;9:595–603.
13. Robinson JR, Anders SH, Novak LL, Simpson CL, Holroyd LE. Consumer health-related needs of pregnant women and their caregivers. *JAMIA Open*. 2018;1:57–66.
14. Jackson GP, Robinson JR, Ingram E, et al. A technology-based patient and family engagement consult service for the pediatric hospital setting. *J Am Med Inform Assoc*. 2018;25:167–174.
15. Cronin RM, Fabbri D, Denny JC, Jackson GP. Automated classification of consumer health information needs in patient portal messages. *AMIA Annu Symp Proc*. 2015;2015:1861–1870.
16. Sulieman L, Gilmore D, French C, et al. Classifying patient portal messages using Convolutional Neural Networks. *J Biomed Inform*. 2017;74:59–70.
17. North F, Crane SJ, Stroebel RJ, Cha SS, Edell ES, Tulledge-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.
18. Masterman M, Cronin RM, Davis SE, Shenson JA, Jackson GP. Adoption of secure messaging in a patient portal across pediatric specialties. *AMIA Annu Symp Proc*. 2016;2016:1930.
19. CMS. Evaluation and management Services. Department of health and human Services, centers for Medicare and Medicaid Services. https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/PhysicianFeeSched/Evaluation-and-Management-Visits.html. Accessed December 15, 2019.
20. Staiger TO, Chew LD, Helenius I. A case-based approach to outpatient evaluation and management service coding. *Postgrad Med*. 2008;120:101–106.
21. Ramos J. Using tf-idf to determine word relevance in document queries. *Proceedings of the first instructional conference on machine learning*. 2003;242:133–142.
22. Soysal E, Wang J, Jiang M, et al. Clamp - a toolkit for efficiently building customized clinical natural language processing pipelines. *J Am Med Inform Assoc*. 2018;25:331–336.
23. Dimick C (AHIMA). Achieving coding consistency. http://library.ahima.org/doc?oid=101092#.Xo4eO4hKg2w; 2010 [accessed 01.04.20].
24. Kaplan LM, Fallon JA, Mun EC, et al. Coding and reimbursement for weight loss surgery: best practice recommendations. *Obes Res*. 2005;13:290–300.
25. Denny JC, Irani PR, Wehbe FH, Smithers JD, Spickard III A. The KnowledgeMap project: development of a concept-based medical school curriculum database. *AMIA Annu Symp Proc*. 2003;2003:195–199.
26. Savova GK, Masanz JJ, Ogren PV, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES). *J Am Med Inform Assoc*. 2010;17:507–513.