Artificial intelligence and e-health for inflammatory bowel diseases: The quest to enhance patient experiences, outcomes and costs

Zand, A.

CHAPTER 5

An Exploration Into the Use of a Chatbot for Patients With Inflammatory Bowel Diseases: Retrospective Cohort Study

Journal of Medical Internet Research. 2020 May 26;22(5):e15589

A. Zand[1,2], A. Sharma[1], Z. Stokes[1], C. Reynolds[1], A. Montilla[3], J. Sauk[1], D.W. Hommes[1,2]

[1] Vatche and Tamar Manoukian Division of Digestive Diseases, UCLA Center for Inflammatory Bowel Diseases, David Geffen School of Medicine, University of California at Los Angeles, Los Angeles, CA, United States
[2] Department of Digestive Diseases, Leiden University Medical Center, Leiden, the Netherlands
[3] Cisco Systems Inc, Collaboration Technology Group, Dallas, TX, United States

Abstract

Background

The emergence of chatbots in health care is fast approaching. Data on the feasibility of chatbots for chronic disease management are scarce.

Objective

This study aimed to explore the feasibility of utilizing natural language processing (NLP) for the categorization of electronic dialog data of patients with inflammatory bowel diseases (IBD) for use in the development of a chatbot.

Methods

Electronic dialog data collected between 2013 and 2018 from a care management platform (*UCLA eIBD*) at a tertiary referral center for IBD at the University of California, Los Angeles, were used. Part of the data was manually reviewed, and an algorithm for categorization was created. The algorithm categorized all relevant dialogs into a set number of categories using NLP. In addition, 3 independent physicians evaluated the appropriateness of the categorization.

Results

A total of 16,453 lines of dialog were collected and analyzed. We categorized 8324 messages from 424 patients into seven categories. As there was an overlap in these categories, their frequencies were measured independently as symptoms (2033/6193, 32.83%), medications (2397/6193, 38.70%), appointments (1518/6193, 24.51%), laboratory investigations (2106/6193, 34.01%), finance or insurance (447/6193, 7.22%), communications (2161/6193, 34.89%), procedures (617/6193, 9.96%), and miscellaneous (624/6193, 10.08%). Furthermore, in 95.0% (285/300) of cases, there were minor or no differences in categorization between the algorithm and the three independent physicians.

Conclusions

With increased adaptation of electronic health technologies, chatbots could have great potential in interacting with patients, collecting data, and increasing efficiency. Our categorization showcases the feasibility of using NLP in large amounts of electronic dialog for the development of a chatbot algorithm. Chatbots could allow for the monitoring of patients beyond consultations and potentially empower and educate patients and improve clinical outcomes.

Background

Recent technological advances have allowed for artificial intelligence (AI) to successfully integrate itself into many aspects of daily life. Besides implementation in voice bots such as Amazon's Alexa and Apple's Siri, AI is also utilized to predict financial stock market changes and answer student questions in educational settings[1]. In health care, AI is expected to disrupt the role of physicians as well; however, experts predict that AI will support the intelligence and knowledge base of physicians rather than replace them entirely[2]. For instance, AI can utilize deep-learning algorithms, which function like the neural networks of the brain and distinguish patterns, to recognize certain types of brain tumors, vascular conditions, or pneumonia on imaging scans and prioritize these cases in the workflow of a radiologist[2,3]. In addition, AI can be used to quickly review patient scans and rule out certain diagnoses, thereby increasing the efficiency and accuracy of a radiologist[2].

Another significant way AI can augment health care delivery is through medical chatbots. A chatbot, or chatterbot, attempts to simulate a natural conversation with a human user[4]. Medical chatbots are already being implemented into regular practice: the Insomnobot-3000 helps insomniacs get through the night, and the Endurance bot acts as a companion for dementia patients[5]. In addition, there are significant efforts toward the development of diagnostic chatbots. Some popular ones include Your.MD, Buoy Health, Sensely, Infermedica, and Florence (Table 1)[6].

Although there are limited data on these general medical chatbots in clinical practice, some independent bodies have provided preliminary and positive results in tests with more specific medical chatbots[7,8].

Most chatbots utilize natural language processing (NLP), which can be simply defined as the use of computers for analyzing human language[9]. One application of NLP relies on human identification of key elements within an event or situation that might constitute a useful summary of a given document or dataset[10]. Recently, there have been growing trends toward the use of electronic health records (EHRs). Multiple studies have attempted to use NLP to extract useful information from EHRs. In one study, researchers used NLP to identify patients with ulcerative colitis and Crohn disease from EHR data collected from Massachusetts General Hospital and Brigham and Women's Hospital[11]. The study developed an algorithm that partly relied on recognizing keywords associated with ulcerative colitis or Crohn disease to analyze the narrative texts and was verified via comparison to a

Table 1. Overview of current medical chatbots

Name	Disease area	Objective	What does it do
Your.MD (UK[a])	General	Provide reliable information for common symptoms, recommends relevant resources	Safely advises patients based on symptoms described in an app-based messaging system
Endurance (Russia)	Dementia	Act as a companion for patients with short-term memory loss and help to identify signs of worsening patient condition	It works via voice recognition to ask questions and react to answers. It can speak on a variety of topics and pull interesting news from Google
Insomnobot-3000 (US[b])	Insomnia	Acts as a companion for insomniacs when they are awake at night.	Has conversations with patients via text
Pharmabot (Philippines)	Pediatrics	Designed to help pediatric patients get appropriate generic medicine for certain ailments	The system works in a software application that sets particular guidelines for interaction with the chatbot
Text-based healthcare chatbots on Mobile Coach (Switzerland)	Childhood obesity	Provide a peer character for obese teenagers and keep them engaged. In addition, sought to show the benefit of text-based chatbot interventions in health care	Works in a text channel within an app interface. Also, has predefined answer options for more efficient chat interactions
Molly by Sensely (US)	General	Diagnose patients with common ailments appropriately based on symptoms	Advises patients based on symptoms described in an app-based messaging system
Buoy Health (US)	General	Diagnose patients accurately based on symptoms. Harvard team developed the algorithm for this bot using 18,000 medical papers for data	Program asks a series of questions—for which there are predefined choices to choose from—to appropriately advise patient. Found on a Web-based software
Symptomate by Infermedica (Poland)	General	Attempt to increase health care provider efficiency, reduce costs, and improve patient flow by acting as a general symptom checker	Online software that collects and analyzes symptom data via predefined questions with answers to provide appropriate response
Florence (Germany)	General	Acts as a **personal nurse** that can remind patients to take prescriptions and keep track of user's health (weight, mood, etc)	Advises patients based on symptoms described in an app via Facebook messenger
Ada (international)	General	Help patients actively manage health based on common symptoms	Ada poses simple and relevant questions to patients and then compares their symptoms with thousands of similar cases to help provide possible explanations
Holly by Nimblr (US)	N/A[c]	Helps patients schedule and reschedule appointments to help prevent no shows or cancellations and improve patient experience	Interacts with patients via text and Amazon's Alexa to update electronic health records
Woebot (US)	Psychiatry	Make mental health care more accessible to people around the world	Uses methods from cognitive behavioral therapy to help patients think through situations. It also includes intelligent mood tracking

[a] UK: United Kingdom. [b] US: United States. [c] N/A: not applicable.

physician's review and classification of the same narrative texts[11]. Ultimately, the study determined that NLP of patient narrative texts provided a more accurate means of identifying patients who had ulcerative colitis and Crohn disease than previous models that had relied on reviewing billing codes[11].

In another study by the University of Alabama, researchers developed an algorithm that analyzed the EHRs of patients collected over 3 years and organized the EHRs into pathology clusters based on key terms[12]. This team also concluded that electronic text mining of health records, or NLP, is an effective method for analyzing large health care datasets[12]. More recent studies have even attempted to use NLP models to study the semantics and sentence flows found in clinical narrative data[13,14]. The literature shows that it is common to perform exploratory analysis on natural language data to understand the topics and vocabulary of a specific domain in health care[9-14]. This exploration is often done by grouping keywords and categorizing topics or using open-source technology such as clinical Text Analysis and Knowledge Extraction[13]. A deep initial understanding facilitates the creation and comparison of more complex, health care-focused NLP models. However, it is worth noting that certain aspects of patient consultations in clinical settings, such as electronic record style, patient behavior, and physician experience, can vary from clinic to clinic[9,14]. This variability found within patient data puts limits on what NLP can do without a large and diverse sample.

In addition, despite the extensive literature on the topic, there seems to be a lack of research into the use of NLP to analyze raw consultation dialog data of patients with specific chronic conditions such as inflammatory bowel diseases (IBD). The organization of the patient with IBD to health care provider (HCP) dialog is likely to be distinct from a general patient population due to the complex nature of the disease. Understanding how these dialogs can be organized is an important first step in assessing the feasibility of a chatbot for this population.

Chatbots that utilize NLP can help to improve the way health care is delivered in multiple ways. For one, they improve accessibility to health care for patients outside of clinics and hospitals. From kids to the elderly, patients often need care outside of inpatient consultations; lack of such support is associated with inefficiency, high health care costs, and burdened HCPs[15]. With a chatbot, these patients would have immediate and autonomous support at home.

Objectives

The primary objective of this study was to accurately categorize large datasets of electronic messages between patients with IBD and HCPs using natural language processing (NLP) to assess the feasibility of developing a medical chatbot for patients with IBD.

Methods

Design and Population

In this study, we aimed to assess the feasibility of utilizing NLP on historical electronic messaging data of patients with IBD for use in the development of a medical chatbot. As IBD is a chronic illness characterized by severe and recurring abdominal pain and diarrhea, patients require frequent contact with their physicians and care team to monitor these alternating disease states and potential relapses[16]. There is great potential here for a chatbot as patients need frequent monitoring beyond regular consultations, which is often troublesome due to the complex nature of the disease and a busy care team.

Patients enrolled in the University of California, Los Angeles (UCLA) Center for IBD electronic care management platform (UCLA eIBD) were retrospectively assessed. The UCLA eIBD platform is a care management software as a service with a Web-based platform for providers that includes treatment decision support, business intelligence, messaging functionality, and performance improvement tools. On the patient's side, there is a mobile app that includes care management insight, educational modules, surveys, and messaging (Figure 1)[16]. Retrospective dialog data between patients and their care team from 2013 until 2018 was extracted and the feasibility of applying NLP categorization algorithms was assessed.

All patients gave informed consent to participate. This study was approved by the Institutional Review Board (IRB) at UCLA with IRB protocol number 17-001208.

Data Collection and Anonymization

The dialogs were extracted from the UCLA eIBD database. The data consisted of the following: (1) a unique identifier, (2) first name, (3) last name, (4) date and time of message, (5) direction of message (HCP to patient or vice versa), (6) message content, (7) potential attachments, (8) HCP classification (urgent and nonurgent), (9) HCP action (responded yes or no), and (10) HCP response message content (Multimedia Appendix 1). The data

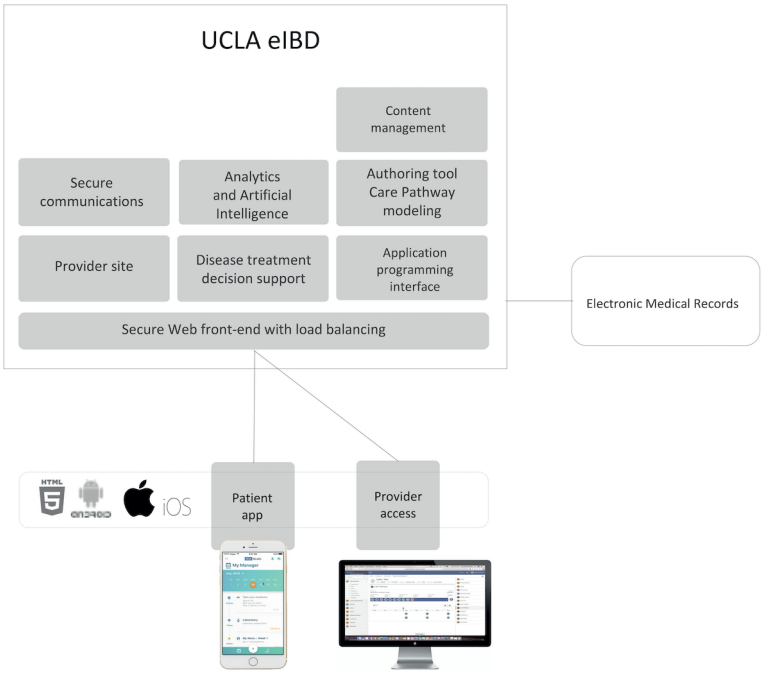

Figure 1. Overview of UCLA eIBD platform. AI: artificial intelligence; API: application programming interface.

were anonymized by removing the first and last names; for identification, we made use of the unique identifier in our analysis.

Categorization Method: Use of Natural Language Processing

Once the patient to HCP dialogs were stored in a Microsoft Excel sheet, the first 400 lines within the sheet were manually analyzed to identify relevant categories for use in our NLP algorithm. To clarify that the first 400 lines were representative, an additional 400 lines were randomly generated and manually reviewed as well (by AS and ZS). The analysis consisted of reading over each line to find an intent; if a particular intent was seen to occur frequently in these first lines, it was noted as a relevant category. The rationale behind using only categories observed in the sample was to make sure that the categories coded for were relevant to what the patient sample was discussing with their HCPs. Furthermore, 2 IBD gastroenterologists reviewed the categories found from the sample and reaffirmed that each category was representative of the IBD patient conversations they had encountered through electronic channels such as email. The same first 400 lines were then used to identify which

keywords could assign a given dialog to a certain category (Multimedia Appendix 2). If a term appeared roughly 10 or more times in a given category, it was noted as a potential keyword; 2 physicians then reviewed and approved our list terms. Using these keywords, we employed a simplified, rule-based bag-of-words model to assign each line of dialog to the appropriate categories (Figure 2). The bag-of-words model essentially allows one to

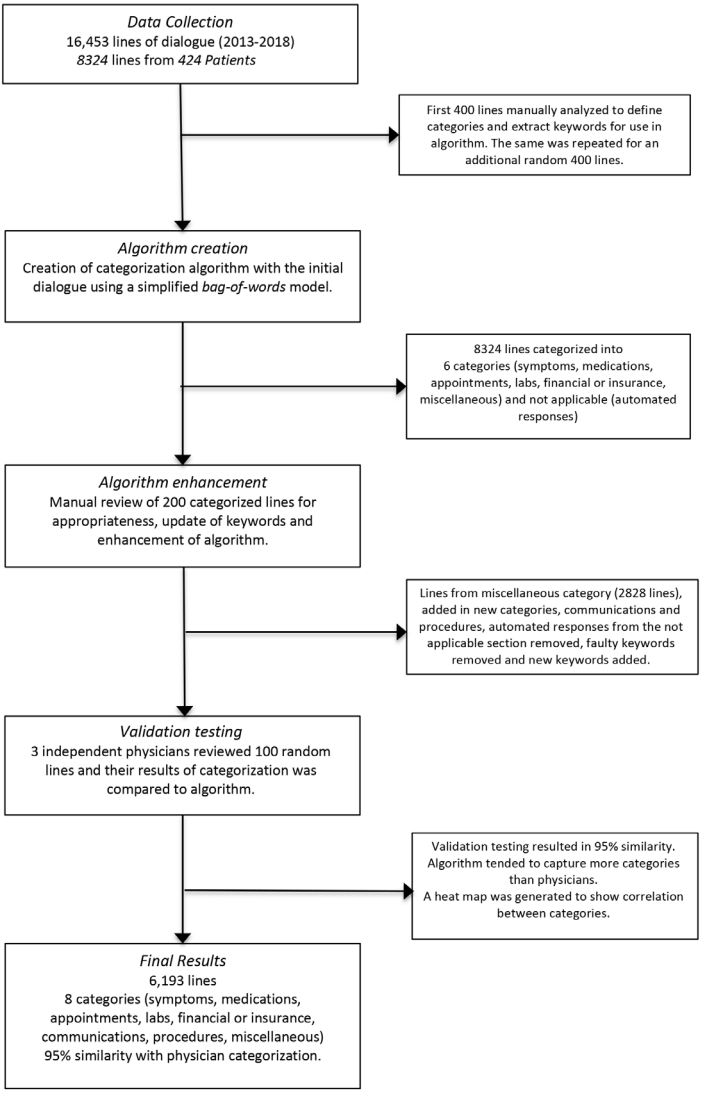

Figure 2. Flowchart of inclusion and categorization. N/A: not applicable.

extract particular features of a text, that is, keywords, and score them with relevant numbers for modeling, or in our case, categorization[17]. To be certain, each line was converted into a standard bag-of-words with a score for each word in the form of a count of the number of times it appears within the line. With stop words removed, we extract the score of each keyword from all lines and assign to each line all categories for which any one keyword has a positive score.

Enhancement and Correlation Assessment

On the basis of the preliminary results, the keywords of our initial categorization algorithm were refined, and new categories were created if necessary. If the categorization was not logical, we analyzed which keywords in the model miscategorized the dialog and made the necessary improvements. In addition, any uncategorized lines of dialogs were assigned a category, and their keywords were identified. The categorization algorithm was updated with the new, physician-approved keywords extracted from the uncategorized lines of dialog and the improvements of the existing categorization.

Once the code was refined to capture all the lines of dialog, a heat map was generated to showcase the overlap in categories, which refers to one line of dialog from a patient falling into two categories. It is worth noting that more than two categories could overlap, but there was no way to represent the higher levels of overlap in a relevant and concise diagram such as a heat map. The goal was to paint a picture of what types of questions or concerns popped up together, which is instrumental in the actual development of a chatbot and creation of multicategory scenarios.

Validation of Accuracy

The accuracy of our categorization algorithm was tested by having 3 independent physicians from the UCLA Division of Digestive Diseases (AZ, CR, and DH) evaluate the appropriateness of the categorization. Each physician was assigned to categorize 100 randomly collected lines of dialog using the defined corresponding category number. In addition, the physicians categorized each line in the same style as the algorithm: numerical order with no spaces.

Once each of the doctors had finished categorizing the lines, the results were compared with the algorithm's categorization. We showcased the extent to which the algorithm and the doctors agreed or disagreed. To do this, the number of underclassifications and overclassifications the categorization algorithm made relative to the doctors' categories was

calculated. For instance, if the algorithm missed a category that the doctor had, it would be counted as an underclassification of 1; if the category code had an extra category compared with the doctor, it would be counted as an overclassification of 1. We then created a bar chart plot based on this data. In addition, to understand the practicality of treating the doctors' assessments as ground truth, we computed the level of agreement between the three raters using Krippendorf alpha. This is a standard estimate of inter-rater reliability across ratings on a nominal scale.

To calculate a metric for the accuracy of the algorithm itself, we opted to use a nonstandard method of computing the success of the classification algorithm in an attempt to incorporate expert knowledge about the severity of misclassifications. As standard reliability measures such as Krippendorf alpha treat all disagreements between the raters and the algorithm with equal weight, we would not get a realistic view of the algorithm's strength across the spectrum of categories by following this approach. This was also done in an attempt to avoid aggregating our multiclass labels from the raters as doing so would put us at risk of destroying the variability in the ratings and inflating performance.

Software

Excel 2010 and R studio programming tool (R 3.4.0) were used for our analysis and algorithm creation (Multimedia Appendix 3).

Results

Data and Population Characteristics

Our sample consisted of 424 patients, 3 physicians, 3 nurses, and 2 administrative assistants with 16,453 lines of electronic dialog. Of the dialogs, 8324 lines were sent by 424 patients to their HCP (patient to HCP). Our analyzed patient cohort is 51.9% (220/424) female, 50.7% (215/424) have Crohn disease, and 46.9% (199/424) have ulcerative colitis with a mean disease duration of 13.4 (SD 10.4) years. The majority of the population is of the white (284/424, 67.0%) race and not of Hispanic or Latino ethnicity (386/424, 91.0%). Furthermore, most of the patients are employed (283/424, 66.7%) and have been enrolled in the care program for a mean of 4.6 (SD 1.3) years (Table 2).

Algorithm Development and Initial Results

In our manual run-through of the first 400 out of the 8324 lines of dialog, we categorized them in six newly created and distinct categories: (1) medications, (2) symptoms, (3) appointments, (4) laboratory investigations, (5) finance/insurance, and (6) miscellaneous

Table 2. Characteristics of the inclusion cohort (N=424)

Variable	Values
Age (years), mean (SD)	42 (14)
Gender, n (%)	
Female	220 (51.9)
Male	204 (48.1)
Disease type, n (%)	
Crohn's disease	215 (50.7)
Ulcerative colitis	199 (46.9)
Indeterminate colitis	10 (2.4)
Disease duration (years), mean (SD)	13.4 (10.4)
Race, n (%)	
White	284 (67.0)
Unknown	97 (22.9)
Asian	26 (6.1)
Black or African American	12 (2.8)
American Indian or Alaska Native	4 (0.9)
Native Hawaiian	1 (0.2)
Ethnicity, n (%)	
Not Hispanic or Latino	386 (91.0)
Hispanic or Latino	29 (6.8)
Unknown	9 (2.1)
Employment, n (%)	
Employed	283 (66.7)
Unemployed or unknown	141 (33.2)
Duration in program (years), mean (SD)	4.6 (1.3)

(lines that did not fall into any of the other categories). When the additional randomly generated 400 lines were reviewed for clarification, the same five relevant categories were found. At this point, we also kept a not applicable (N/A) section for automated responses produced by the mobile app itself that were in the dataset. For instance, "Patient has indicated there are no changes to medications."

We identified what keywords were relevant to each of the categories (Multimedia Appendix 2). A categorization algorithm (bags-of-words model) was created based on the keywords extracted from the dialogs in the categories and applied to categorize the remaining lines of dialog.

Out of the 8324 lines of dialogs, the algorithm initially returned symptoms (1781/8324, 21.40% lines), medications (2114/8324, 25.40% lines), appointments (1781/8324, 21.40% lines), laboratory investigations (1648/8324, 19.80% lines), finance or insurance (358/8324, 4.30% lines), miscellaneous (2830/8324, 34.00% lines), and N/A (666/8324, 8.00% lines).

Enhancement of Natural Language Processing Categorization Algorithm

The miscellaneous section (2828/8317, 34.00% lines) was manually reviewed for 200 lines. The miscellaneous section was essentially randomly generated in that it was not organized by any dialog identifier, such as medical record number or patient name; it was simply the arbitrarily leftover dialogs from our initial run of the algorithm. As the dialogs here were short and not dominated by any one patient, we found it appropriate to review the first 200 lines as an accurate representation of the larger section. On review, two additional categories were identified within it: communications and procedures. In addition, the miscellaneous category was analyzed for keywords that would improve the scope of our initial categories. For instance, there were some medications we missed in our first test, such as Tylenol, that we were able to find upon review of the miscellaneous section and add as a keyword for medications. Furthermore, we removed keywords from the algorithm that were too general and inflated certain categories, such as the keyword take for the medications category.
Finally, the categorization algorithm was enhanced to remove dialog that only contained generic greetings, such as Thank you or Hello, and the automated responses from the N/A section from the dataset so that they did not affect the final counts. After this enhancement, 2131 lines were excluded and 6193 lines of dialog were left for categorization.

Final Natural Language Processing Categorization Results

These refinements ultimately led to the algorithm yielding 32.83% (2033/6193) of the dialog relating to symptoms, 38.70% (2397/6193) to medications, 24.51% (1518/6193) to appointments, 34.01% (2106/6193) to laboratory investigations, 7.22% (447/6193) to finance or insurance, 34.89% (2161/6193) to communications, 9.96% (617/6193) to procedures, and 10.08% (624/6193) being miscellaneous (Table 3). The frequency of this overlap was measured for each possible pair combination of the categories and is displayed in a heat map (Figure 3). For instance, medications and symptoms appeared more together than they did on their own, as did communications and symptoms. Similarly, procedures and finance were very rarely brought up on their own (Figure 3).

Table 3. Final categorization results (N=6193)

Category	Percentage of total sample[a], %
Symptoms	2033 (32.83)
Medications	2397 (38.70)
Appointments	1518 (24.51)
Laboratory investigations	2106 (34.01)
Finance or insurance	447 (7.22)
Communications	2161 (34.89)
Procedures	617 (9.96)
Miscellaneous	624 (10.08)

[a]These percentages represent how frequently these categories occur in the sample of dialogs. As the categories mostly overlap in the dialogs, the percentages do not add up to 100%.

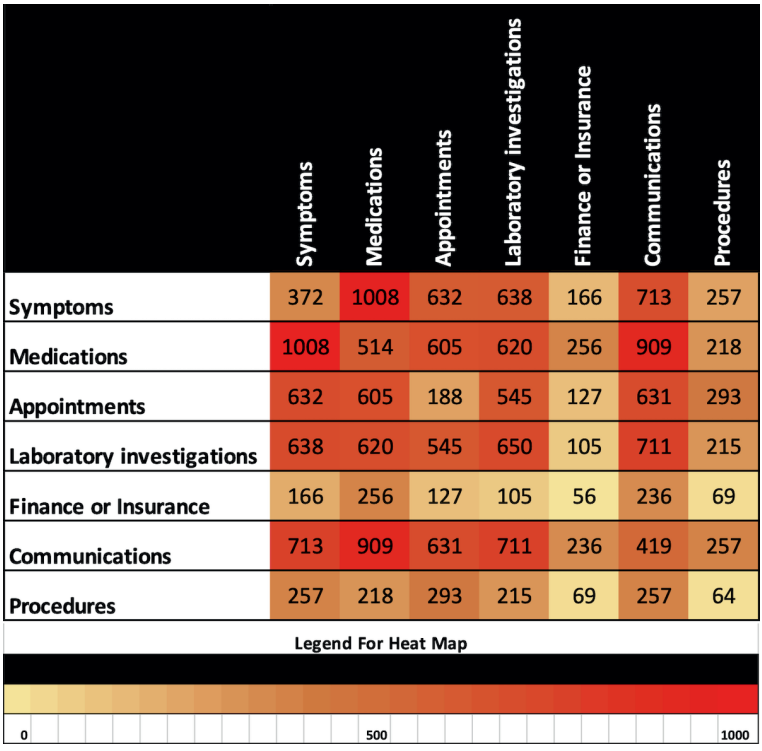

Figure 3. Heat map of category overlaps in dialog. This map shows the frequency of category overlap in pairs and how often the categories occurred by themselves out of the 6193 dialogs. Note: across the diagonal, the map is a mirror of itself.

Validation of Natural Language Processing Accuracy

Three independent raters (AZ, DH, and CR) categorized 100 random lines of dialog, and their categorization was compared with our algorithms. The raters categorized in the exact style of the algorithm, so if the categories were symptoms, appointments, and medications, they would write 123. Applying Krippendorf alpha to these assessment ratings, we get an estimate of .61, indicating that there was moderate-to-high agreement between the doctors. In our underclassification and overclassification representation of the chatbot's accuracy, we found that most of the errors were pooled at one difference, suggesting that the code and the doctors had a high level of agreement on most of the dialogs. Furthermore, the graph we constructed shows that the category code tended to over classify rather than under classify the subjects of the dialogs (Table 4). As one can see from the table, there is a significant drop in the instances of two or more underclassifications, with four to five missed categories having a frequency of 0 (Table 4). When we accounted for the 1 to 2 overclassifcation differences and the one category underclassification differences as minor, we found that 285 of the 300 tests had the program and physicians reasonably agreeing on categories. This meant that our code showed minor to no differences in 95% (285/300) of cases.

Table 4. Accuracy Test Results

Number of categories added or missed by the algorithm in a given line	Instances in Sample for Overclassification	Instances in Sample for Underclassification
1	71	47
2	29	5
3	5	1
4	3	0
5	1	0

Discussion

Principal Findings

We were successful in categorizing large amounts of electronic messages between patients and providers into a reasonable number of categories (<10). Roughly 90.00% (5574/6193) of dialogs that came from patients fell into only seven categories, which shows potential for developing a chatbot with an NLP algorithm that can handle most IBD patient's

questions and concerns. In addition, our heat map gave us insight into how these categories correlate with each other in the dialogs. In terms of chatbot development, this map allows a developer to be aware of what categories or topics tend to appear together in patient with IBD to HCP dialogs. This insight would allow the developer to better prepare the chatbot's NLP algorithm to identify topic transitions in a patient conversation and respond appropriately. In addition, our accuracy test supported the reliability of this result. Most of the differences recorded in our test (100/162, 61.0%) were simply due to code over classifying with one or two categories, but it rarely missed the primary intent (Table 4). Even when it did miss a category relative to the physician, the program was not necessarily incorrect upon review. For instance, one of the dialogs in the accuracy sample had a patient describing their symptoms or medications and subtly mentioning their laboratory investigations as their previous averages. Although the doctors recognized this and appropriately categorized the line as symptoms, medications, and laboratory investigations, the algorithm categorized it as symptoms and medications only, as averages was not a keyword we had programmed for laboratory investigations. Despite this, the program correctly identified the primary intent of the dialog, which is why we considered these types of differences minor in measuring the accuracy of our program.

Limitations

One limitation of this study is that our patient sample is fairly homogenous, consisting of mostly young (mean age 42 years) and white patients, which limits the generalizability of our results to other populations. In addition, most of the patients in the study are employed, which could have potentially changed the types of questions or concerns they expressed and the overall category distribution relative to other patient populations. It is also worth noting that we used the expert opinions of 2 IBD gastroenterologists to support the validity of the categories chosen and the selected keywords. This may affect the reproducibility of our results.

Comparisons With Prior Work

The next step from collecting data to developing a chatbot is to use machine learning methods to model the relationship between questions and responses[18]. Many chatbot knowledge bases (the database from which a chatbot draws its responses from) are hand constructed, which is time consuming and reduces the algorithm's versatility[19]. For instance, Artificial Linguistic Internet Computer Entity and ELIZA, two classic chatbots, utilize hand-constructed databases to generate a response that matches a given human input[20]. As an alternative, some developers have attempted to extract high-quality dialog data from

online discussion forums to efficiently create a knowledge base for specific domain chatbots[19]. The purpose of collecting these dialog datasets is to give the chatbot a training ground to learn how to accurately respond to a specific domain of human input responses with minimal human fine tuning, or simply put: machine learning[18,21]. This machine learning approach also allows for the chatbot to continue learning through its interactions and improve its accuracy. Microsoft's Xiaoice chatbot has successfully applied this model and has already amassed a following of about 660 million online users[22]. When assessing the appropriateness of our data for actual chatbot development, our code could be distributed and tested in other centers with the same historical data without requiring much customization and would eliminate the need for hand-constructed databases.

Conclusions

Looking at the global trends of technology in health care, usage of smartphones and electronic health apps is on the rise[2,4,6]. Patient-provider communication through electronic messaging apps is becoming the standard. In our population, 25.0% (1518/6193) of messages were related to appointments. A chatbot could effectively automate requests regarding booking and cancellations or even play an instrumental part of triage, following the same guidelines as nurses, saving the provider team valuable time that could be redistributed to better patient care. The benefit is that a chatbot is available at all times, can handle tremendous amounts of conversation, and has no wait times.

Through the UCLA eIBD platform, we have already created a high-quality knowledge base of human dialogs that can be used to train an IBD chatbot using NLP. We showcased that it is feasible to categorize large amounts of electronic messaging data in one of the most complex chronic conditions into a reasonable number of categories. Given the feasibility of this categorization and the potential benefits of a chatbot, the next step would be to develop a chatbot and test it in a patient population with IBD. Further studies are required to showcase the effect on patients, providers, and costs and potential extrapolation to other chronic conditions.

References

1. Phillips A. Becoming Human: Artificial Intelligence Magazine. 2018. [2018-10-17]. 4 Industries Artificial Intelligence Is Transforming https://becominghuman.ai/4-industries-artificial-intelligence-is-transforming-fe27b750769b.

2. Kerschberg B. Forbes Magazine. 2018. [2018-10-17]. How Real-Time AI is Accelerating the Disruption of Healthcare (Interview with Nuance Communications) https://www.forbes.com/sites/benkerschberg/2018/03/19/__trashed-3/#15a28317530f.

3. Brownlee J. Machine Learning Mastery. 2019. [2018-10-17]. What is Deep Learning? https://machinelearningmastery.com/what-is-deep-learning/

4. Pearl R. Forbes Magazine. 2018. [2018-10-15]. Artificial Intelligence In Healthcare: Separating Reality From Hype https://www.forbes.com/sites/robertpearl/2018/03/13/artificial-intelligence-in-healthcare/#417555031d75.

5. Shewan D. WordStream: Online Advertising Made Easy. 2020. [2018-11-15]. 10 of the Most Innovative Chatbots on the Web https://www.wordstream.com/blog/ws/2017/10/04/chatbots.

6. Sennaar K. Emerj-Artificial Intelligence Research and Insight. 2019. [2018-10-15]. Chatbots for Healthcare–Comparing 5 Current Applications https://emerj.com/ai-application-comparisons/chatbots-for-healthcare-comparison/

7. Kowatsch T, Nißen M, Shih CH, Rüegger D, Volland D, Filler A, Künzler F, Barata F, Haug S, Büchter D, Brogle B, Heldt K, Gindrat G, Farpour-Lambert N, l'Allemand D. Text-Based Healthcare Chatbots Supporting Patient and Health Professional Teams: Preliminary Results of a Randomized Controlled Trial on Childhood Obesity. 17th International Conference on Intelligent Virtual Agents; IVA'17; August 27-30, 2017; Stockholm, Sweden. 2017.

8. Comendador BE, Francisco BM, Medenilla JS, Nacion SM, Serac TB. Pharmabot: a pediatric generic medicine consultant chatbot. J Automation Control Eng. 2015;3(2):137–40. doi: 10.12720/joace.3.2.137-140. [CrossRef: 10.12720/joace.3.2.137-140]

9. Iroju OG, Olaleke JO. A systematic review of natural language processing in healthcare. Int J Info Technol Comput Sci. 2015 Jul 8;7(8):44–50. doi: 10.5815/ijitcs.2015.08.07. [CrossRef: 10.5815/ijitcs.2015.08.07]

10. Turner AM, Liddy ED, Bradley J, Wheatley JA. Modeling public health interventions for improved access to the gray literature. J Med Libr Assoc. 2005 Oct;93(4):487–94. http://europepmc.org/abstract/MED/16239945. [PMCID: PMC1250325] [PubMed: 16239945]

11. Ananthakrishnan AN, Cai T, Savova G, Cheng S, Chen P, Perez RG, Gainer VS, Murphy SN, Szolovits P, Xia Z, Shaw S, Churchill S, Karlson EW, Kohane I, Plenge RM, Liao KP. Improving case definition of Crohn's disease and ulcerative colitis in electronic medical records using natural language processing: a novel informatics approach. Inflamm Bowel Dis. 2013 Jun;19(7):1411–20. doi: 10.1097/MIB.0b013e31828133fd. http://europepmc.org/abstract/MED/23567779. [PMCID: PMC3665760] [PubMed: 23567779] [CrossRef: 10.1097/MIB.0b013e31828133fd]

12. Raja U, Mitchell T, Day T, Hardin JM. Text mining in healthcare. Applications and opportunities. J Healthc Inf Manag. 2008;22(3):52–6. [PubMed: 19267032]

13. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, Chute CG. Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications. J Am

Med Inform Assoc. 2010;17(5):507–13. doi: 10.1136/jamia.2009.001560. http://europepmc.org/abstract/MED/20819853. [PMCID: PMC2995668] [PubMed: 20819853] [CrossRef: 10.1136/jamia.2009.001560]

14. Kocaballi BA, Coiera E, Tong LH, White SJ, Quiroz JC, Rezazadegan F, Willcock S, Laranjo L. A network model of activities in primary care consultations. J Am Med Inform Assoc. 2019 Oct 1;26(10):1074–82. doi: 10.1093/jamia/ocz046. http://europepmc.org/abstract/MED/31329875. [PMCID: PMC6748800] [PubMed: 31329875] [CrossRef: 10.1093/jamia/ocz046]

15. Petryszyn PW, Witczak I. Costs in inflammatory bowel diseases. Prz Gastroenterol. 2016;11(1):6–13. doi: 10.5114/pg.2016.57883. http://europepmc.org/abstract/MED/27110304. [PMCID: PMC4814543] [PubMed: 27110304] [CrossRef: 10.5114/pg.2016.57883]

16. van Deen WK, van der Meulen-de Jong AE, Parekh NK, Kane E, Zand A, DiNicola CA, Hall L, Inserra EK, Choi JM, Ha CY, Esrailian E, van Oijen MG, Hommes DW. Development and validation of an inflammatory bowel diseases monitoring index for use with mobile health technologies. Clin Gastroenterol Hepatol. 2016 Dec;14(12):1742–50.e7. doi: 10.1016/j.cgh.2015.10.035. [PubMed: 26598228] [CrossRef: 10.1016/j.cgh.2015.10.035]

17. Brownlee J. Machine Learning Mastery. 2017. [2019-10-17]. A Gentle Introduction to the Bag-of-Words Model https://machinelearningmastery.com/gentle-introduction-bag-words-model/

18. Ramesh VC. Chatbots Magazine. 2016. [2019-10-17]. Unsupervised Deep Learning for Vertical Conversational Chatbots https://chatbotsmagazine.com/unsupervised-deep-learning-for-vertical-conversational-chatbots-c66f21b1e0f.

19. Huang J, Zhou M, Yang D. Extracting Chatbot Knowledge From Online Discussion Forums. Proceedings of the 20th international joint conference on Artifical Intelligence; IJCAI'07; January 6-12, 2007; Hyderabad, India. 2007. pp. 423–8. [CrossRef: 10.5555/1625275.1625342]

20. Shum H, He X, Li D. From Eliza to XiaoIce: challenges and opportunities with social chatbots. Front Inf Technol Electron Eng. 2018 Jan 8;19(1):10–26. doi: 10.1631/fitee.1700826. [CrossRef: 10.1631/fitee.1700826]

21. Yufeng G. Towards Data Science. 2017. [2019-10-17]. The 7 Steps of Machine Learning https://towardsdatascience.com/the-7-steps-of-machine-learning-2877d7e5548e.

22. Zhou L, Gao J, Li D, Shum H. arXiv. 2019. [2019-10-17]. The Design and Implementation of XiaoIce, an Empathetic Social Chatbot https://arxiv.org/abs/1812.08989.

Supplementary

Supplementary Table 1. Dialogue data content

Unique Identifier	Report Messages Received/Sent by
Report Messages Content	Report Messages Nurse Note Content
First name	
Last name	
Report Messages Patient Alert	
Report Messages Date & Time	
Report Messages Nurse Alert	

Supplementary Table 2. Keywords for Categories

Category of Dialogue	Description	Keywords
Symptoms	Patient describing characteristics of ailment/problem they are having.	"I'm noticing", "be concerned", "diagnose", "I have been" ,"breaking", "ability", "I have a", "figure out", "pale", "I haven't had", "nausea", "weight", "anemia", "restroom", "bathroom", "stomach pain", "weaken", "sore", "serious pain", "infection", "bloated", "kidney", "itch", "tendon", "sensation", "bowel movement", "sick", "BM", "discomfort", "hurts", "my disease", "pooping", "GI track", "strokes", "spots", "sleep", "ache", "recovering", "BLEEDING", "reaction", "Crohn", "effect", "affect", "symptom", "feel", "problem", "fever", "cramp", "I was experiencing", "I've been", "I've had", "rash", "inflammation", "bleeding", "depression", "anxiety", "stool", "Stool", "depressed", "having pain", "abdominal pain", "medicine"
Medications	Any mention of or changes to a patients medications.	"meds", "prescription", "drug", "treatment", "infusion", "injection", "Vaccine", "taking", "prescribe", "prescription", "refill", "take the", "tabs", "daily", "tablet", "pill", "vaccinate", "miralax", "Miralax", "laxative", "Antibiotic", "antibiotic", "steroids", "supplement", "My medication", "my medication", "vaccine", "shot", "flu shot", "oral", "Flu shot", "the medication", "Walgreens", "walgreens", "CVS", "cvs", "pharmacy", "Pharmacy", "over the counter", "mg", "miligrams", "dose", "dosage", "pro biotic", "probiotic", "Probiotic", "Entyvio", "entyvio", "6MP", "6mp"…. (Additionally, listed out about 50 different medications used by the UCLA IBD Center as keywords.)
Appointments	Patients trying to schedule appointments with provider.	"scheduling", "apt"," appointment", "see me", "see her", "see him", "see Dr", "see the", "seeing", "appt", "I can make", "schedule", "come in", "be there", "head over", "followup", "visit", "SEE OR", "meet"

Supplementary Table 2. Continued

Category of Dialogue	Description	Keywords
Labs	Any question or concerns (troubleshooting, results, etc.) the patient may have.	"lab", "Lab", "results", "blood test", "CBC", "blood panel", "draw", "result", "blood work", "Quest", "quest diagnostic", "sample", "drew blood", "tests", "CRP", "test for", "bloods", "more blood", "this test", "my blood", "Vitamin D", "vitamin D", "Vitamin d", "iron", "glucose"
Finance/Insurance	Patient discussing any questions or concerns related to monetary issues.	"insurance", "cost", "careplan", "expensive", "money", "health plan", "$", "paystub", "Blue Shield", "financial", "funds", "PPO", "HMO", "Tricare", "tricare", "medical bills", "pricing", "Remistart", "remistart", "Co-Pay", "co-pay", "Healthcare"
Communications	The patient trying to get ahold of providers or leaving their contact information.	"E-mail", "email", @gmail.com, "altour.com", "@mednet.ucla.edu", "phone", "number", "my cell", "fax", "message", "Email", "error", "call", "get a hold of", "contact", "speak", "mail", "Zip code", "located", "location", "address"
Procedures	Patient discussing any questions or concerns related to procedures.	"colonoscopy", "procedure", "scopy", "MRI", "PT scan", "Petscan", "CT", "CAT", "x-ray", "X-ray", "surgery", "biopsy", "biop", "TB test", "tuberculosis"

Supplementary Table 3. Algorithm Code

```
setwd("I:/IBDcenter/`STUDIES/Chat-Bot")
Chat = read.csv("Chat.csv",header=TRUE, stringsAsFactors = FALSE)
Messages = Chat[,c(5,6,7)]
Messages[,2] = 0
MessagesHP = subset(Messages, grepl(levels(factor(Messages[,1]))[1], Messages[,1]))
MessagesPH = subset(Messages, grepl(levels(factor(Messages[,1]))[2], Messages[,1]))

########Categorization CODE#########################################################################################
#####################################
one = c("I'm noticing","be concerned","diagnose","I have been","breaking","ability","I have a","figure out","pale","I haven't had","nausea","weight","
anemia","restroom","bathroom","stomach pain","weaken","sore","serious pain","infection","bloated","kidney","itch","tendon","sensation","bowel
movement","sick","BM","discomfort","hurts","my disease","pooping","GI track","strokes","spots","sleep","ache","recovering","BLEEDING","reaction",
"Crohn","effect","affect", "symptom","feel","problem","fever","cramp","I was experiencing","I've been","I've had","rash","inflammation","bleeding","
depression","anxiety","stool","Stool","depressed","having pain","abdominal pain","medicine")
two = c("meds","prescription","drug","treatment","infusion","injection","Vaccine","taking","prescribe","prescription","refill","take the","tabs","daily",
"tablet","pill","vaccinate","miralax","Miralax","laxative","Antibiotic","antibiotic","steroids","supplement","My medication","my medication","vacci-
ne","shot","flu shot","oral","Flu shot","the medication","Walgreens","walgreens","CVS","cvs","pharmacy","Pharmacy","over the counter","mg","
miligrams","dose","dosage","pro biotic","probiotic","Probiotic","Probiotic","tylenol","Entyvio","entyvio","6MP","6mp","Asprin","asprin","Apriso","
Allopurinol","Asacol","Azulfidine","azathioprine","Budesondie","Entocort","Canasa","antidepressants","Cimzia","Cipro","Creon","Colazal","
Cortenema","Cortifoam","Dipentum","Entocort","Flagyl","humira","Humira","Imuran","immodium","Immodium","Lialda","methylpredniso-
lon","Natalizumab","NyQuil","Ibuprofen","Pentasa","Prilosec","Prevacid","Aciphex","Protonix","Methotrexate","Nexium","Dexilant","Prednsione","
Phenergan","Purinethol","Remicade","Rowasa","Simponi","Solu-Medrol","Prozac","Stelara","Tylenol","Useris","vicodin","Vicodin","Zosyn")
three = c("scheduling","apt","appointment","see me","see her","see him","see Dr","see the","seeing","appt","I can make","schedule","come in","
be there","head over","followup","visit","SEE OR","meet")
four = c("lab","Lab","results","blood test","CBC","blood panel","draw","result","blood work","Quest","quest diagnostic","sample","drew blood","tests",
"CRP","test for","bloods","more blood","this test","my blood","Vitamin D","vitamin D","Vitamin d","iron","glucose")
five = c("insurance","cost","careplan","expensive","money","health plan","\\$","hemoglobin","paystub","Blue Shield","financial","funds","PPO","
HMO","Tricare","tricare","medical bills","pricing","Remistart","remistart","Co-Pay","co-pay","Healthcare")
six = c("E-mail","email","@gmail.com","altour.com","@mednet.ucla.edu","phone","number","my cell","fax","message","Email","error","call","get a hold
of","contact","speak","mail","Zip code","located","location","address")
seven = c("colonoscopy","procedure","scopy","MRI","PT scan","Petscan","CT","CAT","x-ray","X-ray","surgery","biopsy","biop","TB test","tuberculo-
sis")
eight = c("Patient has indicated there are changes to","Patient has indicated there are no changes","See attachment...")
nine = c("Thank","thank","Hi","hi","Hey","hey","Hello","hello","Ok","ok","Yes","yes","thx","Testing","testing","Merry Christmas","Good Morning",
"Good morning","Good afternoon","Good Afternoon","good afternoon","Happy New Year","Happy Thanksgiving", "Nice","Happy")

cats = list(one,two,three,four,five,six,seven,eight,nine)
for(g in 1:length(cats)){
 res = rep(0, nrow(MessagesPH))
 for(i in 1:length(cats[[g]])){
  res = res+as.numeric(grepl(cats[[g]][i], MessagesPH[,3]))
 }
 MessagesPH[which(res>0),2] = g + 10*MessagesPH[which(res>0),2]
}
for(j in 1:nrow(MessagesPH)){
 if(MessagesPH[j,2] == 9){
  if(grepl("\\?", MessagesPH[j,3])){
   MessagesPH[j,2]=0
  }
  else{
   if(length(strsplit(MessagesPH[j,3]," ")[[1]])>15){
    MessagesPH[j,2]=0
   }
  }
 }
}
for(y in 1:nrow(MessagesPH)){
 if(grepl("New medication was added on", MessagesPH[y,3])){
  MessagesPH[y,2]=8
 }
}


for(n in 1:nrow(MessagesPH)){
```

```
  if(grepl("<p>", MessagesPH[n,3])){
    MessagesPH[n,2]=8
  }
}
#Phone number searcher
for(w in 1:nrow(MessagesPH)){
  if(MessagesPH[w,2]==0){
    Test= MessagesPH[w,3]
    Test= as.numeric(strsplit(Test,"")[[1]])
    count = 0
    NAcount = 0
    for(z in 1:length(Test)){
      if(!is.na(Test[z])){
        count = count + 1
        NAcount = 0
        if(count==10){
          MessagesPH[w,2]= 6
          break
        }
      }
      else{
        if(NAcount==2 && count > 0){
          count = 0
          NAcount = 0
        }
        if(NAcount<2 && count > 0){
          NAcount = NAcount + 1
        }
      }
    }
  }
}
#####################################

########THE CLEANER: Get rid of 9's and 8's########
remove = NULL
for(w in 1:nrow(MessagesPH)){
  if((MessagesPH[w,2]-9)%%10==0){
    if(((MessagesPH[w,2]-9)/10)==0){
      remove = c(remove,w)
    }
    MessagesPH[w,2] = (MessagesPH[w,2]-9)/10
  }
  if((MessagesPH[w,2]-8)%%10==0){
    if(((MessagesPH[w,2]-8)/10)==0){
      remove = c(remove,w)
    }
  }
}
MessagesPH = MessagesPH[-unique(remove),]

#############Use wisely#################
write.csv(MessagesPH, "Categories2.csv")
#######################################

#######Category Frequency Printer###########
x = table(MessagesPH[,2])
for(h in 0:7){
  print(h)
  print(100*sum(x[which(grepl(as.character(h), rownames(x)))])/6193)

}
###########################################
```

```
#####################HEATMAP#########################
#1.create blank matrix
Heat = matrix(0,nrow = 7, ncol = 7)
#2. THE LOOP
tab = table(MessagesPH[,2])
names = row.names(tab)
for(x in 1:nrow(Heat)){
 for(y in 1:ncol(Heat)){
  for(z in 1:length(names)){
   if((x %in% as.numeric(strsplit(names[z],"")[[1]])) && (y %in% as.numeric(strsplit(names[z],"")[[1]]))){
    Heat[x,y] = Heat[x,y] + tab[z]
   }
  }
 }
}

diag(Heat) = tab[2:8]
color = heat.colors(256)
color = color[256:1]
#heatmap(Heat, main ="Overlap of Categories in Pairs", Rowv=NA, Colv=NA, labRow = c("Medications","Symptoms","Appoint-
ments","Labs","Finance/Insurance","Communications","Procedures"),labCol = c("Medications","Symptoms","Appointments","Labs","Finance/
Insurance","Communications","Procedures"), col = color, scale= "none", margins=c(5,10),symm=TRUE,revC=TRUE)
heatmap.2(Heat, main ="Overlap of Categories in Pairs", Rowv=NA, Colv=NA, labRow = c("Medications","Symptoms","Appoint-
ments","Labs","Finance/Insurance","Communications","Procedures"),labCol = c("Medications","Symptoms","Appointments","Labs","Finance/
Insurance","Communications","Procedures"), col = color, margins=c(5,10),symm=TRUE,revC=TRUE)
####Sample Test For Accuracy Creator###################
set.seed(100)
rownumber = sort(sample(1:nrow(MessagesPH),size=100, replace=FALSE))

subset = MessagesPH[rownumber,3]
subset = cbind(exam,subset)

write.csv(subset, "Catergoriestest.csv", row.names=FALSE)

subsetfull = table(MessagesPH[rownumber,2])

for(h in 0:7){
 print(h)
 print(100*sum(subsetfull[which(grepl(as.character(h), rownames(subsetfull)))])/100)

}
#########################################################

###############Accuracy Checker#########################
Testresults = read.csv("Mastertest.csv",header = TRUE)

Computer = matrix(0,nrow = 100, ncol = 9)
Dan = matrix(0,nrow = 100, ncol = 9)
Aria = matrix(0,nrow = 100, ncol = 9)
Courtney = matrix(0,nrow = 100, ncol = 9)

populate = function(frame,data){

 for(u in 1:nrow(frame)){
  for(g in 9:1){
   if((data[u]-g)%%10 == 0){
    frame[u,g] = 1
    data[u] = (data[u]-g)/10
   }

  }

 }
 return(frame)
}
```

```
Computer = populate(Computer, Testresults[,2])
Dan = populate(Dan, Testresults[,3])
Aria = populate(Aria, Testresults[,4])
Courtney = populate(Courtney, Testresults[,5])

scores = rep(0,300)
underscore = rep(0,300)
overscore = rep(0,300)

for(v in 1:nrow(Computer)){
scores[v] = sum(Computer[v,] != Dan[v,])
scores[v+100] = sum(Computer[v,] != Aria[v,])
scores[v+200] = sum(Computer[v,] != Courtney[v,])
}

for(o in 1:nrow(Computer)){
 for(x in 1:9){
  if(Computer[o,x]-Dan[o,x]<0){
    underscore[o] =  underscore[o]-1
  }
  if(Computer[o,x]-Aria[o,x]<0){
    underscore[o+100] =  underscore[o+100]-1
  }
  if(Computer[o,x]-Courtney[o,x]<0){
    underscore[o+200] =  underscore[o+200]-1
  }
 }
}

for(o in 1:nrow(Computer)){
  for(x in 1:9){
    if(Computer[o,x]-Dan[o,x]>0){
    overscore[o] = overscore[o]+1
    }
    if(Computer[o,x]-Aria[o,x]>0){
    overscore[o+100] =  overscore[o+100]+1
    }
    if(Computer[o,x]-Courtney[o,x]>0){
    overscore[o+200] =  overscore[o+200]+1
    }
  }
}
hist(scores,
     main="Histogram for Raw Differences between Program and Doctor Categorization",
     xlab ="Differences", border="blue", col="green", ylim=c(0,250))
hist(underscore,
     main="Histogram for Underestimations of Categories by Program relative to Doctor",
     xlab ="Number of Missed Categories", border="orange", col="red", ylim=c(0,250))
hist(overscore,
     main="Histogram for Overestimations of Categories by Program relative to Doctor",
     xlab ="Number of Missed Categories", border="brown", col="blue", ylim=c(0,250))
```