#MyIBDHistory on Twitter: Identifying Disease Characteristics Using Personal Tweets

Maya Stemmer
Ben-Gurion University of the Negev, Israel
mayast@post.bgu.ac.il

Gilad Ravid
Ben-Gurion University of the Negev, Israel
rgilad@bgu.ac.il

Yisrael Parmet
Ben-Gurion University of the Negev, Israel
iparmet@bgu.ac.il

Abstract

Inflammatory bowel disease (IBD) is usually classified into Crohn's disease (CD) or ulcerative colitis (UC). Inconclusive cases are diagnosed with IBD unclassified (IBD-U). In 2018, IBD patients shared their disease history on Twitter and signed their tweets with #MyIBDHistory. In this research, we analyzed those tweets and built a logistic regression classifier that predicts patients' IBD type. We constructed tabular classification features and assessed their importance using the regression coefficients and association rules. We identified key features that distinguished CD from UC and used the classifier to predict the disease type of IBD-U patients. Our results correlated with IBD-related research. The two most prominent features that tilted the classification towards CD were suffering from fistulas or nutrient deficiencies. We identified gender differences in disease perspective prior to diagnosis. The research shows that the personal information shared by patients on Twitter can enhance existing medical knowledge regarding their disease.

Keywords: Twitter, IBD, Data Analysis, Logistic Regression, Association Rules.

1. Introduction

Inflammatory bowel disease (IBD) is a chronic inflammation condition of the digestive system characterized by flares and remission states. The two primary diseases identified with IBD, Crohn's Disease (CD) and Ulcerative Colitis (UC), are usually diagnosed in young patients (in the age range of 15-30 years) (Cosnes, Gower–Rousseau, Seksik, & Cortot, 2011; Loftus, 2004). Distinguishing between CD and UC is not trivial as their symptoms and effects may overlap. When the disease features are inconclusive and do not enable a particular diagnosis of CD or UC, patients are diagnosed with IBD Unclassified (IBD-U) (Kirschner, 2022; Winter et al., 2015; Zhou, Chen, Chen, Xu, & Li, 2011). The incidences of IBD are rapidly increasing, and it has evolved into a global disease (Kaplan, 2015).

There are no medications or surgical procedures that can cure IBD. Treatment options can only help with symptoms, affecting each patient differently. They involve prescription drugs and lifestyle-related solutions, such as diets and therapies. Symptoms include abdominal pain, diarrhea, and fatigue; severe cases may result in hospitalization or surgical interventions (Norton, Thomas, Lomax, & Dudley-Brown, 2012; Rubin et al., 2010). As chronic bowel diseases, both CD and UC require routine drug consumption and special nutrition care.

Patients describe IBD as an embarrassing disease that causes immediate disruption of daily activities. They experience difficulties adjusting to the changes IBD entails and consider themselves different from their peers. Since IBD is identified with frequent bowel movements, people do not hasten to share their disease with others (Devlen et al., 2014; Norton et al., 2012). IBD patients attribute part of the embarrassment to a lack of public awareness. Outsiders cannot see that a person's stomach hurts or his bowels are scarred. The disease is invisible, and others might doubt it exists (Frohlich, Dennis Owen, 2016; Kemp, Griffiths, & Lovell, 2012).

The embarrassment caused by IBD and the need to confide in people who undergo similar experiences help explain the creation of IBD-related communities on Twitter. By overcoming space and distance, Twitter users form a community that disregards physical boundaries or immobility. A sense of common ground can help break down barriers and enable conversation, increasing a person's willingness to share (Becker, 2013; Wiese et al., 2011). It may be easier to consult other patients who can relate and better understand the situation based on personal experience. One can identify more closely with users' stories similar to one's own and embrace their advice more easily (Paek, Hove, Ju Jeong, & Kim, 2011). When people disclose health information on Twitter, they expose themselves to various opinions and reduce uncertainty about their disease (Lin, Zhang, Song, & Omori, 2016).

The hashtag *#MyIBDHistory* was first initiated in 2018 by a Twitter account promoting IBD-related discussion called @bottomlineibd. The account's manager is the IBD patient and advocate, Rachel

Sawyer, founder of The Bottom Line (IBD). Sawyer challenged her fellow IBD patients to write their own IBD medical history in a single tweet and sign it with the hashtag *#MyIBDHistory*.

This research aimed to analyze patients' tweets containing the hashtag *#MyIBDHistory* and to determine the disease type of an IBD patient based on their symptoms and treatments. We constructed a list of classification features and used LASSO logistic regression to predict whether a patient suffered from CD or UC. We identified key features and further investigated the connections between the features using association rule learning. The results of both models correlated with IBD-related research. To adhere to ethical norms and maintain user privacy, we only publish aggregated results that do not reveal the specific users. Sawyer herself gave her informed consent to be mentioned in this study.

The rest of the paper is organized as follows: in Section 2. Related work, we explore related research regarding health and IBD on Twitter; in Section 3. Methods, we describe in detail the methods used in this research; in Section 4. Results, we review the results of our research, and in Section 5. Discussion, we discuss the implications of the results and suggest future research.

2. Related work

2.1. Health mentions on Twitter

The study of social media in the context of health and wellbeing continues to position Twitter as a new medium for disseminating health-related information. Topics of health-related tweets range from simple toothache to more severe and chronic diseases such as diabetes, asthma, or cancer (Chulis, 2016; Heaivilin, Gerbert, Page, & Gibbs, 2011). Advances in automated data processing, machine learning, and natural language processing (NLP) present the possibility of utilizing this great data source for health research (Paul et al., 2016).

During the past years, text mining and social network analysis have been used to detect mentions of health on Twitter (Luo, Wang, & Mo, 2022; Yin, Fabbri, Rosenbloom, & Malin, 2015) or to track the spread of the covid-19 pandemic and its symptoms (Jahanbin & Rahmanian, 2020; Lopreite, Panzarasa, Puliga, & Riccaboni, 2021). Regarding chronic conditions, previous research has focused on analyzing patients' tweets and uncovering their Twitter community (Beguerisse-Díaz, McLennan, Garduño-Hernández, Barahona, & Ulijaszek, 2017; Gabarron, Dorronzoro, Rivera-Romero, & Wynn, 2019; Sugawara et al., 2012). While much research was dedicated to diabetes or cancer, Twitter research on IBD is only starting to consolidate.

2.2. Inflammatory bowel disease on Twitter

IBD patients are the most common type of users who talk about IBD on Twitter (Khan et al., 2018; Rowe, Rowe, Silverman, & Borum, 2018). They use Twitter to share their own experiences and to seek social support. They exchange thoughts about symptoms and medications and recommend treatments to one another (O'Neill, Shandro, & Poullis, 2020; Roccetti, Casari, & Marfia, 2015). By sharing their life experiences with the disease on Twitter, patients fight disease invisibility and raise public awareness about IBD (Frohlich, Dennis O. & Zmyslinski-Seelig, 2016).

Pérez-Pérez et al. explored the IBD community on Twitter and identified the types of users who talk about the disease and the key topics they discuss. They categorized users based on their Twitter profiles by analyzing their screen names and profile pictures (Pérez-Pérez, Pérez-Rodríguez, Fdez-Riverola, & Lourenço, 2019). Stemmer et al. constructed a classifier that distinguishes IBD patients from other users who tweet about IBD based on their communication patterns and the content of their tweets (Stemmer, Parmet, & Ravid, 2022). Unlike these two previous studies, in our research, we address a set of users who openly declared their IBD on Twitter and try to distinguish CD patients from UC patients.

2.3. Association rules for healthcare data

Association rule mining is a machine learning technique that aims to discover interesting patterns and relations between variables in a database of transactions. Each transaction is a binary vector defined over a set of binary attributes called items. Positive attributes within the same transaction mean they occurred together (Agrawal, Imieliński, & Swami, 1993; Kotsiantis & Kanellopoulos, 2006).

In recent years, there has been a growing interest in using association rule mining for healthcare data. When applied to health-related data, one usually tries to assess the correlation between two phenomena by considering experimental subjects as transactions (Held et al., 2016; Sarıyer & Öcal Taşar, 2020). Lakshmi and Vadivu (2017) combined association rule mining with NLP to uncover connections between diseases and their symptoms from medical transcription files (Lakshmi & Vadivu, 2017).

In this research, we extract a set of IBD characteristics from patients' tweets and use association rule mining to identify connections between them. Stilou et al. conducted a similar study on diabetes but used clinical data (a database containing records of diabetic patients) rather than social media data (Stilou, Bamidis, Maglaveras, & Pappas, 2001).

3. Methods

3.1. Data collection and preparation

On September 29th, 2021, we used Twitter academic API to collect all tweets containing the hashtag *#MyIBDHistory*. We performed a full-archive search of tweets in English and excluded retweets. 206 tweets, written by 140 different users, were collected. 125 users were IBD patients telling their IBD stories. Others were engaged spectators who did not contribute a story of their own. Patients mentioned their age at diagnosis, the medications they have tried over the years, whether they underwent any surgeries, and more. Some patients described their history in minute detail, insisting on fitting everything into several tweets; others wrote in general and focused on milestones. We were interested in transforming the heterogeneous data written by patients into fixed categorical features, so we could analyze the data using statistical algorithms.

We carefully read all patients' tweets and processed them into a tabular framework containing categorical features. We did not decide on the features in advance; we derived them from the information the patients shared in their tweets. With every new tweet we read, we added features to our framework or updated the existing ones based on the data it contained. Each author performed their own assessment, and we constructed the features upon agreement. The only feature we added that was not mentioned in the tweets was gender.

We deduced each patient's gender by manually looking into their Twitter profile and investigating their full name and profile picture. Notice that this process can be done automatically as Pérez-Pérez showed (Pérez-Pérez et al., 2019). We also investigated their user description (bio) since many users explicitly mentioned how they should be addressed (e.g., she/her) or used informative phrases (like father/husband) in their bio. The combination of full name, profile picture, and bio was enough to determine the gender of 118 patients – eighty females and thirty-eight males.

We were unable to determine the gender of seven responders. Two of them tweeted from social enterprise accounts that did not reveal personal details regarding their authors. The other five accounts were no longer available on Twitter, and we did not want to specify their gender only based on their screen name. We marked the gender of these seven users as Unknown.

The first feature we derived from the tweets was the type of the disease – whether the patient was diagnosed with CD, UC, or IBD-U. Thirty-three patients did not mention their disease type in their tweets, and we searched their Twitter profiles for the information. We determined the disease type of all thirty-three patients based on their previous tweets and Twitter bio. Seventy-six patients had CD, forty-three had UC, and six had IBD-U. They were all clinically diagnosed with IBD.

The second feature we derived from the tweets was the patient's age at diagnosis. Only sixty-seven patients mentioned their age at diagnosis, and we left the feature blank for all other patients. A logistic regression classification model ignores all records with missing values. Hence, we had to forfeit the entire feature or drop half the records in our dataset. Since IBD patterns in childhood differ from adult-onset disease and distinguishing CD from UC in children differs from the equivalent task in adults (Bousvaros et al., 2007; Day, Ledder, Leach, & Lemberg, 2012), we were unwilling to give up the age feature.

Based on previous literature (Crohn's & Colitis Foundation of America, 2014; Trivedi & Keefer, 2015), we considered three age groups that are meaningful to the outburst of IBD: under 15 years old (y/o), between 15 y/o and 35 y/o, and over 35 y/o. We transformed the continuous age feature into one categorical feature, indicating whether the patient belonged to one of the three age groups. Thirteen patients were under 15 y/o, forty-five patients were between 15 y/o and 35 y/o, and nine were over 35 y/o. The other fifty-eight patients who did not mention their age at diagnosis did not belong to any age group.

Based on the diverse types of drugs known to treat IBD (Crohn's & Colitis Foundation of America, 2014; Pithadia & Jain, 2011; Wehkamp, Götz, Herrlinger, Steurer, & Stange, 2016), we created six binary features to describe the patients' medical treatments: anti-inflammatory meds, steroids, antibiotics, biologic meds, immune system suppressors, and other meds. We considered each med feature positive if the patient explicitly mentioned they had tried at least one medication from that specific drug class. A negative value meant that the patient either had tried the drug class but failed to mention it or explicitly mentioned they had not. Fifty-eight tried anti-inflammatory meds, seventy-seven tried steroids, fourteen tried antibiotics, seventy-six tried biologic meds, seventy-six tried immune system suppressors, and thirteen tried other types of medication.

We constructed another two binary features: whether the patient received a different diagnosis at first and whether they had a fistula. Twenty-two patients wrote that they were initially misdiagnosed, thought to have a different type of IBD or just irritable bowel syndrome (IBS). IBS is a gastrointestinal disorder that manifests similar symptoms to IBD but without a clinical diagnosis. Seventeen patients wrote that they had suffered from a fistula. We considered the features positive for the relevant patients and negative for the rest.

We constructed three categorical features: whether the patient underwent any weight changes, whether they changed their diet as a mandatory or preventive action, and whether it took a long time to confirm their diagnosis. Thirteen patients emphasized losing weight or being extremely underweight, while only one mentioned gaining weight with medications. Eleven patients experienced forced diet changes, resorting to a liquid diet

or even tube feeding, and four patients mentioned their diet as a way of controlling their disease. Thirty-five patients said they had suffered for a long time before eventually being diagnosed, and three patients mentioned that their diagnosis was part of emergency surgery. We considered each of the three categorical features unavailable whenever a patient did not explicitly mention one of its values.

We extracted one ordinal variable from the text: whether the patient was ever hospitalized or even had surgery. Eighty-five patients underwent at least one surgery, fifteen patients were hospitalized but have not had surgery, and six explicitly wrote they have never been hospitalized. We considered those who did not regard hospitalization in their tweets as those who said they were never hospitalized.

We transformed each categorical or ordinal feature into a set of binary features based on the number of categories it contained. Table 1 summarizes the features we gathered from the tweets by showing each feature and the presence of its values in our dataset. The right column of the table explains how we eventually used the features in our classification model.

Our study was based solely on publicly available Twitter data and did not perform any clinical intervention. The patients voluntarily provided all the information in their tweets or profile description.

3.2. Logistic regression

We wished to predict the type of IBD based on the patient's symptoms and the treatments they received. We tried several classification algorithms that showed consistent results and decided to demonstrate them using logistic regression because of the simplicity and interpretability of the model.

We considered the disease type as the dependent variable and the rest of the features as independent variables. Then, we used logistic regression to predict whether the patient suffered from CD or UC. We excluded the six patients suffering from IBD-U from our dataset and considered them unlabeled new observations. We used the seventy-six patients suffering from CD and the forty-three patients suffering from UC to train and validate our model.

We used the scilkit-learn (sklearn) package in python (Pedregosa et al., 2011) to split our dataset into training (80%) and test (20%) sets and to build a LASSO logistic regression model. We used L1 regularization since we had twenty-one explanatory variables and a relatively small dataset. The LASSO logistic regression would help eliminate unnecessary independent variables (Tibshirani, 1996). To calibrate the hyperparameter c, we used five-fold cross-validation on our training data. We evaluated different regularization values ($c \in \{0.1, 0.5, 1, 10\}$), and c=1 was the best. Then, we trained a LASSO logistic regression model with the best regularization value on the entire training set. We applied the obtained classifier to the test set to evaluate its performance and estimated feature importance by investigating the regression coefficients.

Finally, we trained our model on all 119 records of CD and UC patients and used the obtained classifier to classify the IBD-U patients.

Table 1. Classification features – description and values.

Feature Name	Description and Values	Type and Model Use
Disease type	CD: 76, UC: 43, IBD-U: 6	Binary, dependent: CD or UC IBD-U as new data
User	Unique Twitter screen name	String, unique identifier For internal use only
Gender	Females: 80, males: 38, unknowns: 7	Two binary features
Age group	Under 15: 13, 15-35: 45, over 35: 9, unknowns: 58	Three binary features
Meds: anti-inflammatory	Yes: 58, no: 67	Binary
Meds: steroids	Yes: 77, no: 48	Binary
Meds: antibiotics	Yes: 14, no: 111	Binary
Meds: biologics	Yes: 76, no: 49	Binary
Meds: immune suppressors	Yes: 76, no: 49	Binary
Meds: others	Yes: 13, no: 112	Binary
Wrong diagnosis	Yes: 21, no: 104	Binary
Fistula	Yes: 17, no: 108	Binary
Weight	Lost: 13, gained: 1, neither: 111	Two binary features
Diet	Mandatory: 11, lifestyle: 4, neither: 110	Two binary features
Pre-diagnosis (Prior to diagnosis)	Prolonged suffering: 35, Emergency surgery: 3, neither: 87	Two binary features
Hospital	Surgery: 85, hospitalized: 15, neither: 25	Two binary features

3.3. Association rule mining

We wished to use association rule mining to discover interesting connections between our IBD features. Unlike the logistic regression model, a rule model enables the exploration of connections within the entire set of features, not just between the target variable and its predictors.

To apply association rules to our data, we addressed the patients as transactions and the extracted features as items. The purpose was to identify features that patients tend to have together. We had 119 transactions – patients who suffered from CD or UC and twenty-two items – the twenty-two binary features described in

Section 3.1. Data collection and preparation, including the type of the disease (CD or UC).

We used an association rules implementation from the mlxten package in python (Raschka, 2018). Since our database was relatively small, we used minimum support of 0.1 to determine the frequent items and a confidence threshold of 0.2 to obtain the association rules. I.e., we only considered pairs of features that at least 10% of the patients mentioned together and only addressed rules where the consequent feature had at least 20% chance of accompanying the antecedent feature.

We obtained a set of rules indicating correlations between the features in our database and used Gephi software to visualize the rules and illustrate the connections. The features were the nodes, and each rule was a directed arc connecting the antecedent feature to the consequent feature. Each arc was colored on a scale from green to red based on the influence of the antecedent on the consequent, with green representing a strong positive correlation and red representing a strong negative correlation. The sizing of the nodes reflected the number of times the features appeared in our database: the more times they appeared, the larger their nodes were. The rules' confidence values determined the thickness of the arcs they represented.

We further investigated rules containing CD as the consequent feature. We evaluated the rules based on the lift measure, which is calculated as the ratio between the rule's confidence and the consequent's support. The lift value indicates how strongly the consequent feature is associated with the antecedent feature. A lift value of one means that the two features are independent. A lift value greater than one implies that seeing the antecedent increases the probability of seeing the consequent. A lift value smaller than one implies that seeing the antecedent decreases the probability of seeing the consequent.

4. Results

4.1. Logistic regression

Table 2 shows five classification metrics evaluating the performance of our regression model. It shows the evaluation of the model on the test set (when trained on the training set) and the evaluation of the model when trained on the entire dataset (for predicting the class of IBD-U patients). Table 3 shows the confusion matrices of both cases. We can see that while our model successfully identified the CD patients, it had difficulty identifying the UC patients.

Figure 1 demonstrates the importance of the features in our model by showing the regression coefficient of each feature. We can see that the most vital feature was Fistula tilting the classification in favor of the CD class. Indeed, in our dataset, none of the patients with a fistula had UC; sixteen had CD, and one had IBD-U. Another essential feature favoring the CD class was a mandatory diet change. Out of the eleven patients who mentioned changing their diets due to nutritional deficiencies, ten had CD, and one had IBD-U.

Table 2. Regression model evaluation results.

Evaluation Measure	Evaluation Data	
	Test Set	***Entire Dataset***
Accuracy	0.75	0.7563
Precision	0.7273	0.7527
Recall	1.0	0.9211
F1	0.8421	0.8284
ROC AUC	0.625	0.6931

Table 3. Confusion matrices for the regression model.

Evaluation Data		Predictions	
		Predicted UC	***Predicted CD***
Test Set	***True UC***	2	6
	True CD	0	16
Entire Dataset	***Ture UC***	20	23
	True CD	6	70

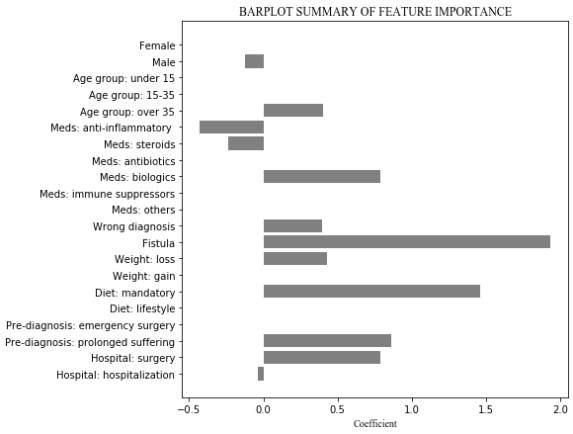

Figure 1. Barplot of feature importance based on regression coefficients.

Nine features turned out to be unimportant and were omitted from the regression model: Gender: female, Age group: under 15, Age group: 15-35, Meds: antibiotics, Meds: immune suppressors, Meds: others, Weight: gain, Diet: lifestyle, and Pre-diagnosis: emergency surgery. In Figure 1, we can see that their coefficients were shrunk to zero by the LASSO algorithm. Another insignificant feature was Hospital: hospitalization with a coefficient close to zero (0.04). Gender: male had the second smallest absolute coefficient of 0.125.

Table 4 presents the features and predictions for the six IBD-U patients. Though five of them were classified as CD patients, we can see that for only three of them, the classification probability was greater than 0.6. The classification was done with great confidence

(probability greater than 0.99) for only one IBD-U patient. The probabilities of the other three predictions were close to 0.5 (between 0.4 and 0.6), meaning that the classification between CD and UC was inconclusive. The results in Table 4 are highlighted with a grayscale based on the strength of the classification.

Table 4. Features and predictions for IBD-U patients.

Feature/ Prediction	Patient					
	IBD1	IBD2	IBD3	IBD4	IBD5	IBD6
Gender: female	0	1	1	0	1	1
Gender: male	1	0	0	1	0	0
Age group: under 15	0	0	1	0	0	0
Age group: 15-35	0	0	0	1	0	0
Age group: over 35	1	0	0	0	1	0
Meds: anti-inflammatory	1	0	0	0	1	0
Meds: steroids	1	0	1	1	1	0
Meds: antibiotics	0	0	1	0	0	0
Meds: biologics	1	0	1	1	1	1
Meds: immune suppressors	1	0	1	0	0	0
Meds: others	0	0	0	0	0	0
Wrong diagnosis	0	0	0	0	0	0
Fistula	0	0	1	0	0	0
Weight: loss	0	0	1	0	0	0
Weight: gain	0	0	0	0	0	0
Diet: mandatory	0	0	1	0	0	0
Diet: lifestyle	0	0	0	0	0	0
Pre-diagnosis: emergency surgery	0	1	0	0	0	0
Pre-diagnosis: prolonged suffering	1	0	0	0	0	0
Hospital: surgery	0	1	1	1	0	0
Hospital: hospitalization	1	0	0	0	0	0
Probability	0.622	0.567	0.994	0.603	0.428	0.589
Class	1	1	1	1	0	1

4.2. Association rule mining

The association rule learning model created ninety-eight rules overall, including inverted rules where the antecedent and the consequent were in reverse order of other existing rules. Figure 2 shows the rules' network, illustrating the connections between our features.

Table 5 presents the twenty most valuable associations ranked by descending lift value. Twelve of them contain distinct features, and the others are inverted rules. Highlighted in grey are associations containing the dependent variable of the regression model – whether the patient has CD. We can identify three features that increase the probability of having CD: Fistula, Wrong diagnosis, and Pre-diagnosis: prolonged suffering.

While the regression model only captured the relations between the explanatory variables and the response variable (disease type), the association rules enabled the exploration of the correlations between one "explanatory" variable and another. The results in Table 5 indicate that females are more likely to mention that they have suffered for a long time before their diagnosis (rules 8-9) or that they were initially misdiagnosed (rules 18-19). Indeed, in our data, twenty-eight out of eighty females (35%) suffered extensively prior to their diagnosis, and sixteen females (20%) were wrongfully diagnosed. The equivalent percentages for the males were only 10.5% and 18.4%, respectively. The results also show that having a fistula correlates with surgery (rule #10) and biologic medications (rule #17).

Table 5. Twenty most significant association rules by descending lift value.

#	Antecedents	Consequents	Support	Confidence	Lift
1	Meds: antibiotics	Meds: anti-inflammatory	0.1008	0.9231	1.9271
2	Meds: anti-inflammatory	Meds: antibiotics	0.1008	0.2105	1.9271
3	Fistula	CD	0.1345	1.0000	1.5658
4	CD	Fistula	0.1345	0.2105	1.5658
5	Meds: antibiotics	Meds: biologics	0.1008	0.9231	1.5256
6	Meds: anti-inflammatory	Meds: immune suppressors	0.3866	0.8070	1.2978
7	Meds: immune suppressors	Meds: anti-inflammatory	0.3866	0.6216	1.2978
8	Pre-diagnosis: prolonged suffering	Gender: female	0.2353	0.8235	1.2895
9	Gender: female	Pre-diagnosis: prolonged suffering	0.2353	0.3684	1.2895
10	Fistula	Hospital: surgery	0.1176	0.8750	1.2698
11	Wrong diagnosis	CD	0.1429	0.8095	1.2675
12	CD	Wrong diagnosis	0.1429	0.2237	1.2675
13	Pre-diagnosis: prolonged suffering	CD	0.2269	0.7941	1.2434
14	CD	Pre-diagnosis: prolonged suffering	0.2269	0.3553	1.2434
15	Meds: anti-inflammatory	Meds: steroids	0.3697	0.7719	1.2413
16	Meds: steroids	Meds: anti-inflammatory	0.3697	0.5946	1.2413
17	Fistula	Meds: biologics	0.1008	0.7500	1.2396
18	Wrong diagnosis	Gender: female	0.1345	0.7619	1.1930
19	Gender: female	Wrong diagnosis	0.1345	0.2105	1.1930
20	Meds: anti-inflammatory	Meds: biologics	0.3445	0.7193	1.1888

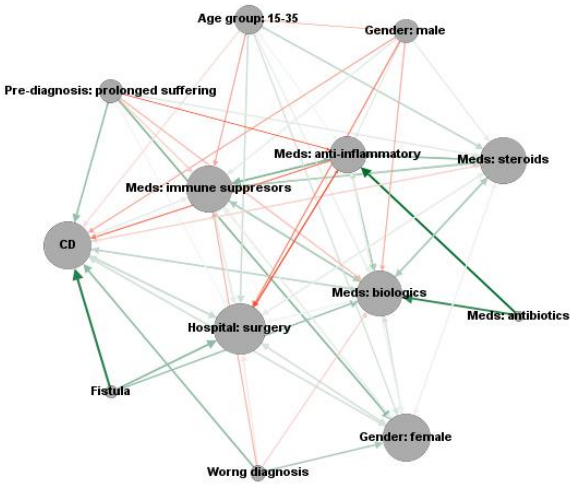

Figure 2. Visualization of feature connections.

5. Discussion

This section discusses the study's principal findings, describes its strengths and limitations, and suggests future work.

5.1. Principal findings

In this study, we collected and analyzed tweets containing the hashtag *#MyIBDHistory*, where IBD patients described their disease history in just one tweet. We transformed the natural language text of the tweets into a tabular database with binary features indicating the symptoms and the treatments the patients experienced. Then, we trained a LASSO logistic regression model that predicts the type of the disease, CD or UC, based on these binary features. We analyzed the importance of our classification features and used the classifier to predict the disease type of patients with IBD-U. We further investigated the connections between our features using association rule learning. To the best of our knowledge, this is the first study to use the *#MyIBDHistory* hashtag to scientifically draw conclusions for IBD, a vital contribution of this research.

The feature importance analysis and the prediction of disease type for IBD-U patients showed the complexity of distinguishing between CD and UC. In some cases, the distinctive characteristics of CD helped identify it. In other cases, the prediction probabilities were approximately 0.5, indicating the ambiguity of the classification.

The two key features that helped distinguish CD from UC were having a fistula and resorting to mandatory diet changes due to nutrient deficiency. These findings align with IBD-related literature since suffering from fistulas or malnutrition is common with CD but seldom occurs with UC (Ashton, Gavin, & Beattie, 2019; Cosnes et al., 2011; Forbes et al., 2017; Lichtenstein, Hanauer, Sandborn, & Practice Parameters Committee of the American College of Gastroenterology, 2009). The IBD-U patient, who was classified as a CD patient with great confidence, also suffered from a fistula and malnutrition.

The classifier ignored the female indicator entirely, and the male indicator had the second smallest absolute coefficient. Overall, our data contained more females than males, even though there are more male Twitter users than female users (Aslam, 2018; Noyes, 2021). Moreover, the prevalence of CD is similar between the two genders, and the prevalence of UC is geographically dependent (Greuter et al., 2020). Both facts explain why the gender features did not contribute to the classification.

The age features showed little, if any, contribution to the classification. Age groups "under 15" and "15-30" were omitted entirely, and the age group "over 35" had a small coefficient. Even after our efforts to maintain available age information, the age at diagnosis was still unknown for almost half the patients (58 out of 125). This fact explains the lack of importance of the age feature in the final model.

The association rule analysis emphasized the importance of the fistula feature and showed that having a fistula significantly increases the probability of having CD. The connection between Fistula and CD was one of the strongest connections in our model and the strongest one containing CD. The analysis also highlighted the connections between CD and receiving a wrong diagnosis and between CD and prolonged suffering prior to diagnosis. According to the association rules analysis, these two phenomena increased the probability of having CD. They also increased the probability of having CD in the regression model, though they did not have the strongest coefficients. Therefore, our two models showed coherent and complementary results.

The generated association rules identified gender differences in disease courses before diagnosis, reinforcing the need for gender-specific medicine. The prevalence of reporting extensive pre-diagnosis suffering or misdiagnosis was higher in females than males. The findings were consistent with IBD-related research indicating differences in coping strategies and quality of life perception between males and females (Sainsbury & Heatley, 2005; Sarid et al., 2017).

5.2 Impact on health-related research

Twitter is becoming an online space for health-related conversations where patients share personal experiences on a global scale (Pérez-Pérez et al., 2019; Yin et al., 2015). The platform is available for patients anytime, allowing them to get support from others sharing their disease. It constitutes a massive database of personal health information that can enrich traditional medical data.

Twitter research enables collecting data from substantial amounts of patients simultaneously and performing personal and aggregative analyses. Hence, such research may derive personalized insights and global comprehension regarding the disease.

This study demonstrates how findings from Twitter research on IBD patients correlate with existing medical knowledge regarding the disease. The presented methods can also help to explore other medical conditions. Therefore, further mining Twitter for health-related data may complement and enhance healthcare research.

5.3 Limitations and Future Work

We focused our research on Twitter and manually processed *#MyIBDHistory* tweets. Therefore, our patient dataset was small and contained only 125 patients. Enriching the dataset by identifying more patients on Twitter or expanding the search to other social media could improve the classifier's performance and precise classification. In a previous study, we identified 337 IBD patients who actively discussed their disease on Twitter (Stemmer, Parmet, & Ravid, 2021). In future research, we intend to add these patients to our model after determining the values of the classification features by mining the patients' Twitter timelines.

Our limited data were also imbalanced: we had 76 CD patients and only 43 UC patients. Nonetheless, we used a 0.5 classification threshold such that any probability greater than 0.5 indicated a CD prediction. This threshold could explain the bias of our model towards the CD class. Future research should consider balancing the groups by adding more UC patients or changing the classification threshold in favor of UC.

The limited dataset and its imbalance were inherent in the information available on Twitter. We did not filter the data other than excluding retweets and focusing on tweets written in English. We used all tweets containing the hashtag *#MyIBDHistory* that met these criteria. We did not oversample our data to adjust imbalanced features, such as gender. We had twice as many women in our data as men. This fact stresses the need to investigate the predilection of each gender to share personal information on social media.

When constructing the classification features, we were dependent on the availability of information provided by the patients. Features such as Fistula, Wrong diagnosis, and Meds were considered positive if the patients explicitly mentioned they had experienced them in their tweets. A negative value meant that the patients either failed to mention they had experienced them or explicitly mentioned they had not. Hence, the heterogeneity of the negative values may have impacted the results. Moreover, we only constructed binary features, representing the existence or inexistence of phenomena. We did not consider the progression of the disease or the frequency of its symptoms.

Though both having a fistula and undergoing surgery had significant coefficients, the two features were correlated since surgery is usually necessary for treating fistulas (Vavricka & Rogler, 2010). Association rule #10 captured this correlation showing how having a fistula increased the probability of having surgery. Future research should better differentiate between these features by using the surgery feature to indicate bowel surgeries or other procedures rather than fistulotomy.

In this research, we analyzed the patients' tweets but did not investigate their possible social connections. Another direction for future work would be to explore the patients' social networks on Twitter and assess their tendency to follow one another.

5.4 Conclusions

In the era of personalized medicine and patient-centered care, it is essential to derive insights reflecting patients' perspectives, as manifested in social media. Collecting and analyzing patients' data on Twitter shows that CD and UC are not easily distinguishable and highlights two key features that help identify CD from UC. It also points out insignificant features for separating the two. The findings provide an additional foundation for existing medical knowledge regarding IBD.

This research suggests that there is room for collaboration between physicians and engineers regarding understanding chronic diseases. The personal information shared by chronically ill patients on Twitter can be used to understand better the disease and how it affects patients' lives. Although such analysis should not strive to replace physicians or draw conclusions of clinical nature, it may provide complementary recommendations for healthy lifestyles based on the wisdom of the crowd.

6. Acknowledgements

This study was supported by a grant of the ERA-Net Cofund HDHL-INTIMIC (INtesTInal MICrobiomics) under the umbrella of Joint Programming Initiative "A healthy diet for a healthy life".

7. References

Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, pp. 207-216. doi:10.1145/170036.170072

Ashton, J. J., Gavin, J., & Beattie, R. M. (2019). Exclusive enteral nutrition in crohn's disease: Evidence and practicalities. Clinical Nutrition, 38(1), 80-89. doi:10.1016/j.clnu.2018.01.020

Aslam, S. (2018). Twitter by the numbers: Stats, demographics & fun facts. Omnicoreagency.Com,

Becker, K. L. (2013). Cyberhugs: Creating a voice for chronic pain sufferers through technology. Cyberpsychology, Behavior, and Social Networking, 16(2), 123-126. doi:10.1089/cyber.2012.0361

Beguerisse-Díaz, M., McLennan, A. K., Garduño-Hernández, G., Barahona, M., & Ulijaszek, S. J. (2017). The 'who'and 'what'of # diabetes on twitter. Digital Health, 3, 2055207616688841. doi:10.1177/2055207616688841

Bousvaros, A., Antonioli, D., Colletti, R., Dubinsky, M., Glickman, J., Gold, B., et al. (2007). Differentiating ulcerative colitis from crohn disease in children and young adults: Report of a working group of the north american society for pediatric gastroenterology, hepatology, and nutrition and the crohn's and colitis foundation of america. Journal of Pediatric Gastroenterology and Nutrition, 44(5), 653-674. doi:10.1097/MPG.0b013e31805563f3

Chulis, K. (2016). Data mining twitter for cancer, diabetes, and asthma insights. Purdue University.

Cosnes, J., Gower–Rousseau, C., Seksik, P., & Cortot, A. (2011). Epidemiology and natural history of inflammatory bowel diseases. Gastroenterology, 140(6), 1785-1794. e4. doi:10.1053/j.gastro.2011.01.055

Crohn's & Colitis Foundation of America. (2014). The facts about inflammatory bowel diseases.

Day, A. S., Ledder, O., Leach, S. T., & Lemberg, D. A. (2012). Crohn's and colitis in children and adolescents. World Journal of Gastroenterology: WJG, 18(41), 5862. doi:10.3748/wjg.v18.i41.5862

Devlen, J., Beusterien, K., Yen, L., Ahmed, A., Cheifetz, A. S., & Moss, A. C. (2014). The burden of inflammatory bowel disease: A patient-reported qualitative analysis and development of a conceptual model. Inflammatory Bowel Diseases, 20(3), 545-552. doi:10.1097/01.MIB.0000440983.86659.81

Forbes, A., Escher, J., Hébuterne, X., Kłęk, S., Krznaric, Z., Schneider, S., et al. (2017). ESPEN guideline: Clinical nutrition in inflammatory bowel disease. Clinical Nutrition, 36(2), 321-347. doi:10.1016/j.clnu.2016.12.027

Frohlich, D. O., & Zmyslinski-Seelig, A. N. (2016). How uncover ostomy challenges ostomy stigma, and encourages others to do the same. New Media & Society, 18(2), 220-238. doi:10.1177/1461444814541943

Frohlich, D. O. (2016). The social construction of inflammatory bowel disease using social media technologies. Health Communication, 31(11), 1412-1420. doi:10.1080/10410236.2015.1077690

Gabarron, E., Dorronzoro, E., Rivera-Romero, O., & Wynn, R. (2019). Diabetes on twitter: A sentiment analysis. Journal of Diabetes Science and Technology, 13(3), 439-444. doi:10.1177/1932296818811679

Greuter, T., Manser, C., Pittet, V., Vavricka, S. R., Biedermann, L., & on behalf of Swiss IBDnet, an official working group of the Swiss Society of Gastroenterology. (2020). Gender differences in inflammatory bowel disease. Digestion, 101 Suppl 1, 98-104. doi:10.1159/000504701

Heaivilin, N., Gerbert, B., Page, J. E., & Gibbs, J. L. (2011). Public health surveillance of dental pain via twitter. Journal of Dental Research, 90(9), 1047-1051. doi:10.1177/0022034511415273

Held, F. P., Blyth, F., Gnjidic, D., Hirani, V., Naganathan, V., Waite, L. M., et al. (2016). Association rules analysis of comorbidity and multimorbidity: The concord health and aging in men project. Journals of Gerontology Series A: Biomedical Sciences and Medical Sciences, 71(5), 625-631. doi:10.1093/gerona/glv181

Jahanbin, K., & Rahmanian, V. (2020). Using twitter and web news mining to predict COVID-19 outbreak. Asian Pacific Journal of Tropical Medicine, 13(8), 378. doi:10.4103/1995-7645.279651

Kaplan, G. G. (2015). The global burden of IBD: From 2015 to 2025. Nature Reviews Gastroenterology & Hepatology, 12(12), 720-727. doi:10.1038/nrgastro.2015.150

Kemp, K., Griffiths, J., & Lovell, K. (2012). Understanding the health and social care needs of people living with IBD: A meta-synthesis of the evidence. World Journal of Gastroenterology, 18(43), 6240-6249. doi:10.3748/wjg.v18.i43.6240

Khan, A., Silverman, A., Rowe, A., Rowe, S., Tick, M., Testa, S., et al. (2018). Who is saying what about inflammatory bowel disease on twitter?

Kirschner, B. S. (2022). Inflammatory bowel disease unclassified (IBD-U)/indeterminate colitis. Textbook of pediatric gastroenterology, hepatology and nutrition (pp. 393-399) Springer. doi:10.1007/978-3-030-80068-0_29

Kotsiantis, S., & Kanellopoulos, D. (2006). Association rules mining: A recent overview. GESTS International Transactions on Computer Science and Engineering, 32(1), 71-82.

Lakshmi, K. S., & Vadivu, G. (2017). Extracting association rules from medical health records using multi-criteria decision analysis. Procedia Computer Science, 115, 290-295. doi:10.1016/j.procs.2017.09.137

Lichtenstein, G. R., Hanauer, S. B., Sandborn, W. J., & Practice Parameters Committee of the American College of Gastroenterology. (2009). Management of crohn's disease in adults. Official Journal of the American College of Gastroenterology| ACG, 104(2), 465-483. doi:10.1038/ajg.2018.27

Lin, W., Zhang, X., Song, H., & Omori, K. (2016). Health information seeking in the web 2.0 age: Trust in social media, uncertainty reduction, and self-disclosure. Computers in Human Behavior, 56, 289-294. doi:10.1016/j.chb.2015.11.055

Loftus, E. V. (2004). Clinical epidemiology of inflammatory bowel disease: Incidence, prevalence, and environmental influences. Gastroenterology, 126(6), 1504-1517. doi:10.1053/j.gastro.2004.01.063

Lopreite, M., Panzarasa, P., Puliga, M., & Riccaboni, M. (2021). Early warnings of COVID-19 outbreaks across europe from social media. Scientific Reports, 11(1), 1-7. doi:10.1038/s41598-021-81333-1

Luo, L., Wang, Y., & Mo, D. Y. (2022). Identifying COVID-19 personal health mentions from tweets using masked attention model. IEEE Access. doi:10.1109/ACCESS.2022.3179808

Norton, B., Thomas, R., Lomax, K. G., & Dudley-Brown, S. (2012). Patient perspectives on the impact of crohn's

disease: Results from group interviews. Patient Prefer Adherence, 6, 509-520. doi:10.2147/PPA.S32690

Noyes, D. (2021). Distribution of twitter users worldwide as of january 2021, by gender.

O'Neill, P., Shandro, B., & Poullis, A. (2020). Patient perspectives on social-media-delivered telemedicine for inflammatory bowel disease. Future Healthcare Journal, 7(3), 241. doi:10.7861/fhj.2020-0094

Paek, H., Hove, T., Ju Jeong, H., & Kim, M. (2011). Peer or expert? the persuasive impact of YouTube public service announcement producers. International Journal of Advertising, 30(1), 161-188. doi:10.2501/IJA-30-1-161-188

Paul, M. J., Sarker, A., Brownstein, J. S., Nikfarjam, A., Scotch, M., Smith, K. L., et al. (2016). Social media mining for public health monitoring and surveillance. In Biocomputing 2016: Proceedings of the Pacific Symposium, pp. 468-479. doi:10.1142/9789814749411_0043

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in python. The Journal of Machine Learning Research, 12, 2825-2830.

Pérez-Pérez, M., Pérez-Rodríguez, G., Fdez-Riverola, F., & Lourenço, A. (2019). Using twitter to understand the human bowel disease community: Exploratory analysis of key topics. Journal of Medical Internet Research, 21(8), e12610. doi:10.2196/12610

Pithadia, A. B., & Jain, S. (2011). Treatment of inflammatory bowel disease (IBD). Pharmacological Reports, 63(3), 629-642. doi:10.1016/s1734-1140(11)70575-8

Raschka, S. (2018). MLxtend: Providing machine learning and data science utilities and extensions to python's scientific computing stack. Journal of Open Source Software, 3(24), 638. doi:10.21105/joss.00638

Roccetti, M., Casari, A., & Marfia, G. (2015). Inside chronic autoimmune disease communities: A social networks perspective to Crohn's patient behavior and medical information. In 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 1089-1096. IEEE. doi:10.1145/2808797.2808813

Rowe, A., Rowe, S., Silverman, A., & Borum, M. L. (2018). P024 crohn's disease messaging on twitter: Who's talking? Gastroenterology, 154(1), S13-S14.

Rubin, D. T., Dubinsky, M. C., Panaccione, R., Siegel, C. A., Binion, D. G., Kane, S. V., et al. (2010). The impact of ulcerative colitis on patients' lives compared to other chronic diseases: A patient survey. Digestive Diseases and Sciences, 55(4), 1044-1052. doi:10.1007/s10620-009-0953-7

Sainsbury, A., & Heatley, R. (2005). Psychosocial factors in the quality of life of patients with inflammatory bowel disease. Alimentary Pharmacology & Therapeutics, 21(5), 499-508. doi:10.1111/j.1365-2036.2005.02380.x

Sarid, O., Slonim-Nevo, V., Pereg, A., Friger, M., Sergienko, R., Schwartz, D., et al. (2017). Coping strategies, satisfaction with life, and quality of life in crohn's disease: A gender perspective using structural equation modeling analysis. PLoS One, 12(2), e0172779. doi:10.1371/journal.pone.0172779

Sarıyer, G., & Öcal Taşar, C. (2020). Highlighting the rules between diagnosis types and laboratory diagnostic tests for patients of an emergency department: Use of association rule mining. Health Informatics Journal, 26(2), 1177-1193. doi:10.1177/1460458219871135

Stemmer, M., Parmet, Y., & Ravid, G. (2021). What Are IBD Patients Talking About on Twitter?. In International Conference on ICT for Health, Accessibility and Wellbeing (pp. 206-220). Springer, Cham. doi:10.1007/978-3-030-94209-0_18.

Stemmer, M., Parmet, Y., & Ravid, G. (2022). Identifying Patients With Inflammatory Bowel Disease on Twitter and Learning From Their Personal Experience: Retrospective Cohort Study, Journal of medical Internet research, 24(8), e29186. doi:10.2196/29186

Stilou, S., Bamidis, P. D., Maglaveras, N., & Pappas, C. (2001). Mining association rules from clinical databases: An intelligent diagnostic process in healthcare. Studies in Health Technology and Informatics, (2), 1399-1403.

Sugawara, Y., Narimatsu, H., Hozawa, A., Shao, L., Otani, K., & Fukao, A. (2012). Cancer patients on twitter: A novel patient community on social media. BMC Research Notes, 5(1), 1-9. doi:10.1186/1756-0500-5-699

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267-288.

Trivedi, I., & Keefer, L. (2015). The emerging adult with inflammatory bowel disease: Challenges and recommendations for the adult gastroenterologist. Gastroenterology Research and Practice, 2015, 260807. doi:10.1155/2015/260807

Vavricka, S. R., & Rogler, G. (2010). Fistula treatment: The unresolved challenge. Digestive Diseases, 28(3), 556-564. doi:10.1159/000320416

Wehkamp, J., Götz, M., Herrlinger, K., Steurer, W., & Stange, E. F. (2016). Inflammatory bowel disease: Crohn's disease and ulcerative colitis. Deutsches Ärzteblatt International, 113(5), 72. doi:10.3238/arztebl.2016.0072

Wiese, J., Kelley, P. G., Cranor, L. F., Dabbish, L., Hong, J. I., & Zimmerman, J. (2011). Are you close with me? are you nearby?: Investigating social groups, closeness, and willingness to share. In Proceedings of the 13th International Conference on Ubiquitous Computing, pp. 197-206. doi:10.1145/2030112.2030140

Winter, D. A., Karolewska-Bochenek, K., Lazowska-Przeorek, I., Lionetti, P., Mearin, M. L., Chong, S. K., et al. (2015). Pediatric IBD-unclassified is less common than previously reported; results of an 8-year audit of the EUROKIDS registry. Inflammatory Bowel Diseases, 21(9), 2145-2153. doi:10.1097/MIB.0000000000000483

Yin, Z., Fabbri, D., Rosenbloom, S. T., & Malin, B. (2015). A scalable framework to detect personal health mentions on twitter. Journal of Medical Internet Research, 17(6), e4305. doi:10.2196/jmir.4305

Zhou, N., Chen, W., Chen, S., Xu, C., & Li, Y. (2011). Inflammatory bowel disease unclassified. Journal of Zhejiang University Science B, 12(4), 280-286. doi:10.1631/jzus.B1000172