#MyIBDHistory on Twitter

Predicting IBD Type from Personal Tweets

Maya Stemmer, Gilad Ravid, Yisrael Parmet
Department of Industrial Engineering and Management
Ben-Gurion University of the Negev
P.O.B. 653, Beer Sheva, Israel
e-mails: mayast@post.bgu.ac.il, rgilad@bgu.ac.il, iparmet@bgu.ac.il

***Abstract*—Inflammatory Bowel Disease (IBD) is a chronic inflammation condition of the digestive system that is usually classified into one of two diseases: Crohn's Disease (CD) or Ulcerative Colitis (UC). If neither one is diagnosed with certainty, the patient is diagnosed with IBD Unclassified (IBD-U). IBD patients form communities on Twitter and exchange thoughts regarding their diseases. In 2018, IBD patients shared their disease history on Twitter and signed their tweets with the hashtag *#MyIBDHistory*. In their tweets, they mentioned their age at diagnosis, the medications they have tried over the years, whether they underwent any surgeries, and more. In this research, we analyzed patients' tweets containing the hashtag *#MyIBDHistory* and built a classifier that predicts the IBD type (CD or UC) of a patient. We transformed the disease history described in the tweets into tabular classification features and assessed their importance. We identified key features that helped distinguish CD from UC and used the classifier to predict the disease type of IBD-U patients. Our results were correlated with IBD related research, as the two most prominent features that tilted the classification towards CD were having a fistula and suffering from nutrient deficiency.**

Keywords-Twitter; IBD; data analysis; logistic regression.

I. 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) [1,2]. 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 certain diagnosis of CD or UC, patients are diagnosed with IBD Unclassified (IBD-U) [3,4]. The incidences of IBD are rapidly increasing, and it has evolved into a global disease [5].

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 [6,7]. As chronic bowel diseases, both CD and UC require day-to-day care for drug consumption and special nutrition.

Patients describe IBD as an embarrassing disease that causes immediate disruption of daily activities. They experience difficulties adjusting to the changes it 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 [8,9]. IBD patients attribute part of the embarrassment to a lack of public awareness. Outsiders cannot see that a person's stomach hurts or that their bowels are scarred. The disease is invisible, and others might doubt that it exists [10,11].

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. IBD patients are the most common type of users who talk about IBD on Twitter [12]. They use Twitter for sharing their own experiences and for seeking social support. They exchange thoughts about symptoms and medications and recommend treatments to one another [13]. By sharing their life experiences with the disease on Twitter, patients fight disease invisibility and raise public awareness about IBD [14].

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 community. 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 our results 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 II we explore related research regarding health and IBD on Twitter, in Section III we describe in detail the methods used in this research, in Section IV we review the results of our research, in Section 0 we discuss the implications of the results, and in Section VI we conclude and suggest future research.

II. Related Work

Many studies have used machine learning models [15,16] and neural networks [17] for predicting whether an individual suffers from a specific illness. Nonetheless, standard logistic regression continues to hold a prominent role in performing such tasks [18]. It yields as good a performance as more complicated models while being easy to interpret [19,20]. Using a small dataset, logistic regression even shows more stable results compared to machine learning models and neural networks [21,22]. Lately, there has been a growing interest in the use of penalized regression for IBD diagnosis [23].

During the past years, text mining and social network analysis have been used to detect personal health mentions on Twitter [24,25], identify depression [26], or track the spread of the covid-19 pandemic [27]. Regarding IBD, two previous studies [28,29] have automatically identified patients with IBD among the IBD community on Twitter, but did not differentiate between CD and UC. The authors in [30] implemented a deep-learning method for extracting medical entities from social media and showed state-of-the-art results on other diseases. Nonetheless, their method failed to distinguish CD from other IBD complications because of their overlapping symptoms.

In this study, we addressed a set of users who openly declared their IBD on Twitter and tried to distinguish between CD patients and UC patients. We used a penalized logistic regression model to predict whether a patient with IBD suffered from CD or UC based on the information they shared in their tweets.

III. Methods

In this section, we describe the data we gathered for this study and the method we used to analyze the data and predict the patients' disease type.

A. 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 that was not limited to a specific timeframe. We excluded retweets and limited our search to tweets written in English. Two hundred six tweets, written by 140 different users, were collected. The earliest tweet containing the hashtag was published on July 26, 2018. The hashtag was intensively used until mid-August 2018, and sporadically used later. The latest tweet containing the hashtag was published on July 23, 2019, as a reminder of last year's discourse. Our study was based solely on those publicly available tweets and did not perform any clinical intervention.

The collected tweets were manually processed by the authors of this paper who are experts in statistics and social network analysis. One hundred twenty-five users were IBD patients telling their IBD story, while 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 information in contained. The only feature we added, though 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 et al. showed [28]. 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 twitted 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 were able to determine the disease type of all thirty-three patients based on their previous tweets and their Twitter bio. Seventy-six patients had CD, forty-three had UC, and six had IBD-U.

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. Since the logistic regression classification model ignores all records with missing values, 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 [31,32], we were unwilling to give up the age feature.

Based on previous literature [1,2], 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 when diagnosed with IBD, forty-five patients were between 15 y/o and 35 y/o, and nine were over 35 y/o. The fifty-eight patients from whom we were not able to derive their age did not belong to any of the age groups.

Based on the diverse types of drugs known to treat IBD [1,33], we created six binary features to describe the patients' medical treatments: anti-inflammatory medications (meds), steroids, antibiotics, biologic meds, immune system suppressors, and other meds. We considered each med feature positive if the patients explicitly mentioned they had tried at least one medication from that specific drug class. A negative value meant that the patients either have tried the drug class but failed to mention it or explicitly mentioned they have 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 was initially diagnosed with a different disease or a different type of IBD and whether the patient had a fistula. Twenty-two patients wrote they were misdiagnosed, and seventeen wrote they suffered from a fistula. We considered the features positive for the relevant patients and negative for the rest.

We considered 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/ordinal feature into a set of binary features based on the number of categories it contained. TABLE I. 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.

B. *Predicting Disease Type*

We wished to predict the type of IBD based on the symptoms the patients had and the treatments they received. We tried several learning algorithms that showed consistent results and decided to focus this paper on logistic regression because of its simplicity and interpretability.

TABLE I. CLASSIFICAION 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

We considered the disease type as the dependent variable and the rest of the features as independent variables and used logistic regression to predict whether the patient suffered from CD (1) or UC (0). We excluded the six patients suffering from IBD-U from our dataset and considered them as 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 [34] to split our dataset into training (80%) and test (20%) sets and to build a LASSO logistic regression model. We decided to use L1 regularization since we had twenty-one explanatory variables and a relatively small dataset. The LASSO logistic regression would help eliminate unnecessary independent variables [35].

We used five-fold cross-validation on our training data to evaluate different regularization parameter values ($c \in \{0.1, 0.5, 1, 10\}$) and select the best one ($c = 1$). Then we trained a LASSO logistic regression model with the best regularization parameter 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. This means, we trained our model on the entire dataset of labeled data to predict the classification of new observations.

IV. Results

TABLE II. shows five classification metrics evaluating the performance of our regression model. The table 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 III. 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.

TABLE II. 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
Area Under the Receiver Operating Characteristic Curve (AUC ROC)	0.625	0.6931

TABLE III. 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 demonstrates the importance of the features in our model by showing the regression coefficient of each feature. We can see that the strongest feature was fistula, tilting the classification in favor of the CD class. Indeed, in our dataset, none of the patients who mentioned suffering from a fistula had UC: sixteen of them had CD, and one had IBD-U. The second strongest feature was a mandatory diet change, again favoring the CD class. Out of the eleven patients who mentioned changing their diets due to nutritional deficiencies, ten had CD, and one had IBD-U. Then, there was a group of three features favoring a CD prediction with notable coefficients: Pre-diagnosis: prolonged suffering, Meds: biologics, and Hospital: surgery.

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, whose coefficient was close to zero (0.04). Gender: male had the second smallest absolute coefficient of 0.125.

TABLE IV. 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, and for only one of them, the classification was done with great confidence (probability greater than 0.99). 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 IV. are highlighted with a gray scale based on the strength of the classification.

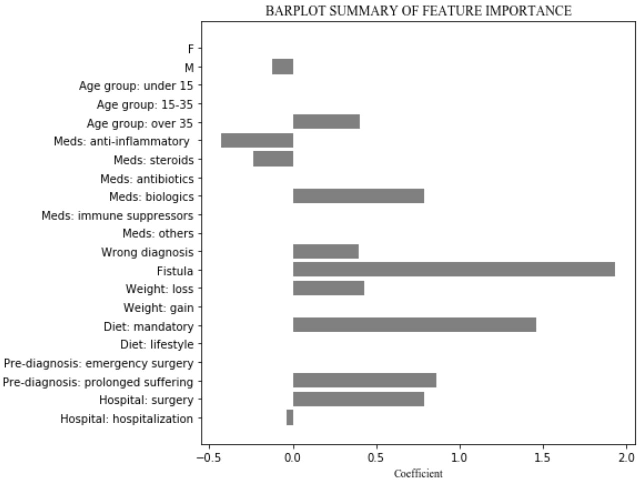

Figure 1. Barplot of feature importance based on regression coefficients.

TABLE IV. 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

The results from the feature importance analysis and the prediction for IBD-U patients correlate and demonstrate the challenge in predicting the IBD type. As can be seen in Figure 1, substantially more features favored a CD classification and the few favoring a UC classification had relatively small coefficients. The results in TABLE IV. show the influence of the features on the prediction: When the model observed one of the strong features, it confidently returned a CD prediction. When none of the strong features were present, the model returned an equivocal prediction based on weaker features.

Patient IBD3 who had a fistula and a mandatory diet change, the two strongest classification features, was unambiguously classified as a CD patient. Patients IBD1 and IBD4 both used biologic medications and showed prolonged suffering and surgery, respectively. Hence, they were both classified as CD patients, but not with the same confidence. Each of the other three patients showed only one of the medium level features, if any. Their predictions were, therefore, indecisive with probabilities close to 0.5. Patient IBD5 who used both anti-inflammatory medications and steroids, the two strongest features favoring UC, was the only one classified as a UC patient.

V. DISCUSSION

In this section, we discuss the study's principal findings, describe its strengths and limitations, and suggest future work.

A. 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 that indicated 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 did not compare the performance of our classifier with classification results from previous studies [28-30] since our prediction task differed from theirs. To the best of our knowledge, this is the first study to use social media data for differentiating between CD and UC.

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

The use of LASSO logistic regression helped to eliminate unnecessary independent variables that did not contribute to the classification. A regular logistic regression would give a small coefficient, but not zero. If we had used regular logistic regression, we had to perform a procedure of model selection such as forward selection or backward elimination to get exactly zero coefficients.

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 [36,37]. The IBD-U patient who was classified as a CD patient with great confidence also suffered from fistula and malnutrition.

The female indicator was ignored entirely by the classifier 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 Twitter users [38]. Moreover, the prevalence of CD is similar within the two genders and the prevalence of UC is geographically dependent [39]. Both facts can explain why the gender features did not contribute to the classification.

B. Impact on Health-Related Research

Twitter is becoming an online space for health-related conversations where patients share personal experiences on a global scale [25]. The platform is available for patients at any time, allowing them to get support from others sharing their disease. It constitutes a huge database of personal health information that can enrich traditional medical data.

Twitter research enables to collect data from substantial amounts of patients simultaneously and to perform both personal and aggregative analyses. Hence, such research may derive not only personalized insights but also global comprehensions regarding the disease.

This study demonstrates how findings from Twitter research on IBD patients correlate with existing medical knowledge regarding the disease. Predicting the type of IBD will help physicians when determining the right treatments for patients. Insights from such study can serve as a decision support system for physicians facing a challenging diagnosis. Therefore, further mining Twitter for health-related data may complement and enhance healthcare research.

C. Limitations

We focused our research on Twitter and manually processed tweets containing the hashtag *#MyIBDHistory*. Therefore, our patients' dataset was relatively small and contained only 125 patients. Enriching the dataset, by identifying more patients on Twitter or expanding the search to other social media, could significantly improve the classifier's performance and make the classification more precise.

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 could explain the bias of our model towards the CD class.

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. All tweets containing the hashtag *#MyIBDHistory* that met this description were used in this study.

Finally, though both having a fistula and undergoing surgery had meaningful coefficients, the two features are correlated since surgery is usually necessary for treating fistulas [37]. Our surgery feature does not differentiate between bowel surgeries and fistulotomies.

VI. Conclusion and Future Work

In the era of personalized medicine and patient-centered care, it is important to derive insights that reflect the patients' perspective, 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.

In a previous study [29], aiming to identify patients with IBD on Twitter, we trained and evaluated a classifier that distinguishes patients with IBD from other users who tweet about the disease. In future research, we intend to enrich our patients' dataset by applying the classifier and identifying more patients with IBD. Then, we can mine their Twitter timelines for the key features found in this study and enable the analysis of big data.

Future research should compare the results of such scalable analysis with those presented in this study and evaluate the contribution of collecting patients' data automatically. On the one hand, one can achieve a much larger dataset of patients. On the other hand, the dataset may contain erroneous identification of patients due to the imperfection of the classifier. It would be interesting to investigate how this trade-off affects the quality of the results.

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. The presented methods, which were applied to IBD, can also help to explore other medical conditions. 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.

Acknowledgment

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".

References

[1] Crohn’s & Colitis Foundation of America, "The facts about inflammatory bowel diseases," Inflammatory bowel diseases, vol. 2, p. 1, 2014. [PDF].

[2] I. Trivedi and L. Keefer, "The emerging adult with inflammatory bowel disease: challenges and recommendations for the adult gastroenterologist," Gastroenterology Research and Practice, 2015.

[3] B. S. Kirschner, "Inflammatory Bowel Disease Unclassified (IBD-U)/Indeterminate Colitis," Textbook of Pediatric Gastroenterology, Hepatology and Nutrition, pp. 393-399, Springer, Cham, 2022.

[4] D. A. Winter et al., "Pediatric IBD-unclassified is less common than previously reported; results of an 8-year audit of the EUROKIDS registry," Inflammatory bowel diseases, vol. 21, no. 9, pp. 2145-2153, 2015.

[5] G. G. Kaplan, "The global burden of IBD: from 2015 to 2025," Nature reviews Gastroenterology & hepatology, vol. 12, no. 12, pp. 720-727, 2015.

[6] B. Norton, R. Thomas, K. G. Lomax, and S. Dudley-Brown, "Patient perspectives on the impact of Crohn’s disease: results from group interviews," Patient preference and adherence, vol. 6, pp. 509-520, 2012.

[7] D. T. Rubin et al., "The impact of ulcerative colitis on patients’ lives compared to other chronic diseases: a patient survey," Digestive diseases and sciences, vol. 55, no. 4, pp. 1044-1052, 2010.

[8] J. Devlen et al., "The burden of inflammatory bowel disease: a patient-reported qualitative analysis and development of a conceptual model," Inflammatory bowel diseases, vol. 20, no. 3, pp. 545-552, 2014.

[9] N. J. Hall, G. P. Rubin, A. Dougall, A. Hungin, and J. Neely., "The fight for ‘health-related normality’: a qualitative study of the experiences of individuals living with established inflammatory bowel disease (IBD)," Journal of health psychology, vol. 10, no. 3, pp. 443-455, 2005.

[10] D. O. Frohlich, "The Social Construction of Inflammatory Bowel Disease Using Social Media Technologies," Health Communication, vol. 31, no. 11, pp. 1412-1420, 2016.

[11] G. G. Macdonald et al., "Patient perspectives on the challenges and responsibilities of living with chronic inflammatory diseases: qualitative study," Journal of Participatory Medicine, vol. 10, no. 4, e10815, 2018.

[12] A. Rowe, S. Rowe, A. Silverman, and M. L. Borum, " P024 Crohn’s disease messaging on twitter: who’s talking?," Gastroenterology, vol. 154, no. 1, pp. S13-S14, 2018.

[13] P. O'Neill, B. Shandro, and A. Poullis, "Patient perspectives on social-media-delivered telemedicine for inflammatory bowel disease," Future Healthcare Journal, vol. 7, no. 3, p. 241, 2020.

[14] D. O. Frohlich and A. N. Zmyslinski-Seelig, "How Uncover Ostomy challenges ostomy stigma, and encourages others to do the same," New media & society, vol. 18, no. 2, pp. 220-238, 2016.

[15] S. F. Weng, J. Reps, J. Kai, J. M. Garibaldi, and N. Qureshi, "Can machine-learning improve cardiovascular risk prediction using routine clinical data?," PloS one, vol. 12, no. 4, e0174944, 2017.

[16] I. K. A. Enriko, M. Suryanegara, and D. Gunawan, "Heart disease prediction system using k-Nearest neighbor algorithm with simplified patient's health parameters," Journal of Telecommunication, Electronic and Computer Engineering (JTEC), vol. 8, no. 12, pp. 59-65, 2016.

[17] N. H. Ismail, N. Liu, M. Du, Z. He, and X. Hu. "A deep learning approach for identifying cancer survivors living with post-traumatic stress disorder on Twitter," BMC Medical Informatics and Decision Making, vol. 20, no. 4, pp. 1-11, 2020.

[18] M. E. Shipe, S. A. Deppen, F. Farjah, and E. L. Grogan, "Developing prediction models for clinical use using logistic regression: an overview," Journal of thoracic disease, vol. 11, no. Suppl 4, pp. S574-S584, 2019.

[19] S. Nusinovici et al., "Logistic regression was as good as machine learning for predicting major chronic diseases," Journal of clinical epidemiology, vol. 122, pp. 56-69, 2020.

[20] E. Christodoulou et al., "A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models," Journal of clinical epidemiology, vol. 110, pp. 12-22, 2019.

[21] T. van der Ploeg, P. C. Austin, and E. W. Steyerberg, "Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints," BMC medical research methodology, vol. 14, no. 1, pp 1-13, 2014.

[22] S. Nalluri S, R. Vijaya Saraswathi, S. Ramasubbareddy, K. Govinda, and E. Swetha, "Chronic heart disease prediction using data mining techniques," Data engineering and communication technology, pp. 903-912, Springer, 2020.

[23] B. Stankovic et al., "Machine Learning Modeling from Omics Data as Prospective Tool for Improvement of Inflammatory Bowel Disease Diagnosis and Clinical Classifications," Genes, vol 12, no. 9, p. 1438, 2021.

[24] L. Luo, Y. Wang, and D. Y. Mo, "Identifying COVID-19 Personal Health Mentions from Tweets Using Masked Attention Model," IEEE Access, 2022.

[25] Z. Yin, D. Fabbri, S. T. Rosenbloom, and B. Malin, "A scalable framework to detect personal health mentions on Twitter," Journal of medical Internet research, vol. 17, no. 6, e4305, 2015.

[26] Y. Zhang et al., "Monitoring depression trends on twitter during the COVID-19 pandemic: observational study," JMIR infodemiology, vol. 1, no. 1, e26769, 2021.

[27] M. Lopreite, P. Panzarasa, M. Puliga, and M. Riccaboni, "Early warnings of COVID-19 outbreaks across Europe from social media," Scientific reports, vol. 11, no. 1, pp. 1-7, 2021.

[28] M. Pérez-Pérez, G. Pérez-Rodríguez, F. Fdez-Riverola, and A. Lourenço, "Using twitter to understand the human bowel disease community: exploratory analysis of key topics," Journal of medical Internet research, vol. 21, no. 8, e12610, 2019.

[29] M. Stemmer, Y. Parmet, and G. Ravid, "Identifying Patients With Inflammatory Bowel Disease on Twitter and Learning From Their Personal Experience: Retrospective Cohort Study," Journal of medical Internet research, vol. 24, no. 8, e29186, 2022.

[30] S. Scepanovic, E. Martin-Lopez, D. Quercia, and K. Baykaner. "Extracting medical entities from social media," In Proceedings of the ACM Conference on Health, Inference, and Learning, pp. 170-181, 2020.

[31] A. Bousvaros et al., "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, vol. 44, no. 5, pp. 653-674, 2007.

[32] A. S. Day, O. Ledder, S. T. Leach, and D. A. Lemberg, "Crohn’s and colitis in children and adolescents," World journal of gastroenterology: WJG, vol. 18, no. 41, p. 5862, 2012.

[33] A. B. Pithadia and S. Jain, "Treatment of inflammatory bowel disease (IBD)," Pharmacological Reports, vol. 63, no. 3, pp. 629-642, 2011.

[34] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," the Journal of machine Learning research, vol. 12, pp. 2825-2830, 2011.

[35] R. Tibshirani, "Regression shrinkage and selection via the lasso," Journal of the Royal Statistical Society: Series B (Methodological), vol. 58, no. 1, pp. 267-288, 1996.

[36] J. J. Ashton, J. Gavin, and R. M. Beattie, "Exclusive enteral nutrition in Crohn's disease: Evidence and practicalities," Clinical nutrition, vol. 38, no. 1, pp. 80-89, 2019.

[37] J. Cosnes, C. Gower–Rousseau, P. Seksik, A. Cortot, "Epidemiology and natural history of inflammatory bowel diseases," Gastroenterology, vol. 140, no. 6, pp. 1785-1794, 2011.

[38] D. Noyes, "Distribution of Twitter users worldwide as of January 2021, by gender," 2021.

[39] T. Greuter, C. Manser, V. Pittet, S. R. Vavricka, and L. Biedermann, "Gender Differences in Inflammatory Bowel Disease," Digestion, vol. 101, no. 1, pp. 98-104, 2020.