Web-based and machine learning approaches for identification of patient-reported outcomes in inflammatory bowel disease

Laetitia Ricci, Yannick Toussaint, Justine Becker, Hiba Najjar, Alix Renier, Myriam Choukour, Anne Buisson, Corinne Devos, Jonathan Epstein, Laurent Peyrin Biroulet, et al.

Version of Record: https://www.sciencedirect.com/science/article/pii/S159086582100774X
Manuscript_ede8c15572f1bcb78a0624e4436e850b

Web-based and machine learning approaches for identification of patient-reported outcomes in inflammatory bowel disease

Laetitia Ricci,
CHRU-Nancy, INSERM, Université de Lorraine, CIC 1433 Clinical Epidemiology, F-54000 Nancy, France
Corresponding author: l.ricci@chru-nancy.fr
+33 82 85 20 92

Yannick Toussaint,
Laboratoire lorrain de recherche en informatique et ses applications, Université de Lorraine, Nancy, France
Yannick.Toussaint@loria.fr

Justine Becker,
Ecole des mines de Nancy, Université de Lorraine, Nancy, France
justine.becker5@etu.univ-lorraine.fr

Hiba Najjar,
Ecole des mines de Nancy, Université de Lorraine, Nancy, France
hiba.najjar@etu.mines-nancy.univ-lorraine.fr

Alix Renier
Ecole des mines de Nancy, Université de Lorraine, Nancy, France
alix.renier@etu.mines-nancy.univ-lorraine.fr

Myriam Choukour,
INSERM, U1256 NGERE and gastroenterology Department, CHRU-Nancy, Université de Lorraine, Nancy, France
M.CHOUKOUR@chru-nancy.fr

Anne Buisson,
afa Crohn RCH France
anne.buisson.afa@gmail.com

Corinne Devos,
afa Crohn RCH France
corinne.afa@gmail.com

Jonathan Epstein,
CHRU-Nancy, INSERM, Université de Lorraine, CIC 1433 Clinical Epidemiology, F-54000 Nancy, France
Université de Lorraine, APEMAC, F-54000 Nancy, France
j.epstein@chru-nancy.fr

Laurent Peyrin Biroulet,
INSERM, U1256 NGERE and gastroenterology Department, CHRU-Nancy, Université de Lorraine, Nancy, France
peyrinbiroulet@gmail.com

Francis Guillemin,
CHRU-Nancy, INSERM, Université de Lorraine, CIC 1433 Clinical Epidemiology, F-54000 Nancy, France
Université de Lorraine, APEMAC, F-54000 Nancy, France
francis.guillemin@chru-nancy.fr

Word Count: 2217

This work was supported by a grant from the French Ministry of Health (CPRC 2017, 2019-A01520-57). The sponsor is Nancy University Hospital (Research and Innovation Direction).

Web-based and machine learning approaches for identification of patient-reported outcomes in inflammatory bowel disease

Abstract

Background: Messages from an Internet forum are raw material that emerges in a natural setting (i.e., non-induced by a research situation).

Aims: The FLARE-IBD project aimed at using an innovative approach consisting of collecting messages posted by patients in an Internet forum and conducting a machine-learning study (data analysis/language processing) for developing a patient-reported outcome measuring flare in inflammatory bowel disease meeting international requirements.

Methods: We used web-based and machine learning approaches, in the following steps. 1) Web-scraping to collect all available posts in an Internet forum (23 656 messages) and extracting metadata from the forum. 2) Twenty patients were randomly assigned 50 extracted messages; participants indicated whether the message corresponded or not to the flare phenomenon (labeling). If yes, participants were asked to identify excerpts from the text they considered significant flare markers (annotation). 3) The set of annotated messages underwent a vocabulary analysis.

Results: The phenomenon of flare was circumscribed with the identification of 20 surrogate flare markers classified into five dimensions with their frequency within extracted labeled data: impact on life, symptoms, extra-intestinal manifestations, drugs and environmental factors. Web-based and machine-learning approaches met international recommendations to inform the content and structure for the development of patient-reported outcomes.

Keywords: Patient-reported outcomes; World Wide Web-based approach; Machine learning

1. Introduction

Crohn's disease and ulcerative colitis, the two major forms of inflammatory bowel disease (IBD), are chronic disabling conditions characterized by flares followed by periods of remission. In the context of treat-to-target strategies and tight monitoring [1], early detection of a disease flare is the only way to change patients' lives and disease course.

Many indices of disease activity are available (Mayo Index, St Mark's Index, Lichtiger Index, Simple Clinical Colitis Activity Index, Seo Index for ulcerative colitis) [2–6]. However, these indices are applied by the physician (hetero-evaluation) and they were developed without integrating patients' perspectives.

Until recently, there has been no validated patient-reported outcomes (PROs) tool to measure the phenomenon of flare in IBD. This gap is gradually being closed: in 2018, a study developed and validated a Ulcerative Colitis PROs signs and symptoms diary [7]. Another group developed a symptoms and impacts questionnaire for Crohn's disease and ulcerative colitis [8]. These two studies met international recommendations to ensure content validity stressed by the International Society for Pharmacoeconomics and Outcomes Research, the US Food and Drug Administration and the European Medicines Agency [9–12]: 1) elicitation of key concepts using focus groups and/or interviews to inform the content and structure of the new PROs and 2) assessment of the patient's understanding of the draft instrument using cognitive interviews. However, in these two studies, qualitative data were collected with experienced interviewers who used a semi-structured discussion guide. Therefore, discussions were minimally oriented by the guide and influenced by the presence of the interviewer. In other words, discussions are necessarily a biased material because they are not spontaneously generated in a natural setting.

In the last ten years, there has been an emergence in using Internet forums as a source of qualitative data to explore patients' perspectives on sensitive or intimate subjects in the field of

sexual medicine [13,14], depression symptoms [15,16], substance abuse [17–22], mental illness [23], women infertility [24], people living with HIV/AIDS and concerns about antiretroviral therapy risks, [25,26], postpartum depression [27], Munchausen by proxy [28], fathers' worries during pregnancy [29], symptomatic pelvic organ prolapse after vaginal birth [30], the experiences parents describe when administering inhaler devices to their young children [31], and suicide [32,33]. Exploration of Internet forums can also be used to try to grasp societal phenomena such as the current trend of vaccine hesitancy [34]. In these papers, messages of interest were often researched by 1) boolean queries formulated directly on search engines to mix several public Internet forums, 2) relevant posts manually selected from a set of data, or 3) selecting all the messages posted for a particular thread or for a particular periodicity. To the best of our knowledge, with the notable exception of 2 papers using web-scraping to collect all available posts [35,36], the engineering sciences were not integrated in data collection.

Our aim was to identify a PROs measuring flare in IBD via an innovative approach combining web-based and machine learning.

Machine learning is considered as a subdiscipline of artificial intelligence. Machine learning learns from examples provided in sample data (training data) and improves its performances through the use of computational algorithms [37–39].

IBD flares can occur at any time between two outpatient visits and are unpredictable. International guidelines now recommend a tight monitoring of both symptoms and intestinal inflammation to allow early detection of IBD flares and thus early intervention, with the final aim of preventing disability and disease progression (bowel damage, hospitalisations and surgeries) [40]. However, patients with IBD are seen every 3–6 months in case of active disease and every 6–12 months during the remission phases. Hence, tools allowing tight monitoring of patients with IBD outside these scheduled outpatient visits are eagerly awaited [41].

2. Materials and methods

The FLARE-IBD project (https://clinicaltrials.gov/ct2/show/NCT04180345) aimed at using an innovative approach consisting of collecting messages posted by patients in an Internet forum and conducting a machine-learning study (data analysis/language processing) for developing and validating a PROs measuring flare in IBD [41]. Most of the time, IBD is diagnosed in young people between 20 and 30 years old. Thus, collecting and analyzing perspectives of patients posted in an Internet forum seems of interest, particularly in a young population familiar with current communication media.

Methods involved three steps: data scraping and cleaning, labeling and annotation, and vocabulary analysis.

2.1.Data scraping and cleaning

Patients' testimonies were extracted from the Association François Aupetit (AFA) Internet forum (https://clinicaltrials.gov/ct2/show/NCT04180345). The AFA, with 25 000 members and supporters, is the unique French organization for IBD, recognized for its public utility.

The messages from this forum, accessible without any identification, are under current public register and thus the content can be analyzed to generate scientific knowledge. The study was conducted in full agreement with the AFA.

Messages were written in *blog-style* French. A data-scraping process was performed to extract messages posted on the AFA forum by using the Scrapy web-crawling tool combined with Splash, a JavaScript rendering service, all running under Python (Figure 1). The complete execution took less than 0.5 hour. A Readme was also available to update the dataset and cover periods after freezing, if needed.

The forum architecture consists of 8 themes (gates), namely: "Live. Move" "Treatments" "Medical examinations", "In the operating room", "Parents. Children", 'Having children", "Young", "Friends. Lovers". In the spring of 2020, a "COVID-19" theme was added.

Extracted metadata were number of posts by theme, name of the threads (plus number of posts by thread), pseudoname of contributors (plus number of posts by contributor).

Because the raw data were not well structured, we cleaned the data (Figure 2) to remove undesirable tags and characters. Then the data were stored in a spreadsheet format, assigning to each post/message an identifier.

2.2. Labeling and annotation

Twenty patients, members of the AFA, were asked by mail by the AFA research department to be annotators in the framework of an anonymous survey. They received 50 randomly assigned extracted messages. Participants should indicate whether the message corresponded to the flare phenomenon in IBD or not (labeling, yes/no). If the message positively matched the flare phenomenon (yes), the annotator identified excerpts from the text they considered significant flare markers (annotation).

A total of 1000 messages were distributed to participants (50 different messages per participant). A message could be selected for distribution if it contained the words "flare" and/or "crisis" at least once (no matter how the word was spelled). No further guidance was provided to the participants to let them freely consider all aspects they wanted about flare in IBD.

To facilitate this step and in accordance with cognitive load theory [42,43], a Visual Basic for Applications interface on Excel was created to limit input errors (Figure 3).

2.3. Vocabulary analysis

2.3.1. Lemmatization of the messages

The French language has a rich morphology system. The SpaCy free open-source library for French Natural Language Processing in Python was used for data lemmatization

(details: https://spacy.io/models/fr#fr_core_news_sm). Concretely, verbs were replaced by their infinitive form. Nouns and adjectives were returned to the masculine singular.

2.3.2. Cleaning of lemmatized messages

Cleaning consisted of removing a set of stop words (uninformative words) [44], then words deemed unnecessary (greetings and salutations). Spelling was then corrected. Finally, to homogenize the results, all punctuation was removed, as were line breaks and capitalization.

Finally, messages are represented as a bag of words as shown in Figure 4.

All annotations were gathered (i.e., excerpts from the text considered by patients as flare markers). Vocabulary groups based on every lexical field that was well represented were next generated. Their representation was determined according to the root of every word. Then the counter () function was applied.

Table 1 summarizes the steps and tools we used for the web-based and machine-learning approach.

3. Results

3.1. Data scraping and cleaning

In April 2019, a total of 23 656 messages were extracted and placed in a readable Excel file enabling metadata querying even by non-computer specialists. Among the 8 themes (gates) of the forum — “Live. Move” “Treatments” “Medical examinations”, “In the operating room”, “Parents. Children”, ‘Having children”, “Young”, “Friends. Lovers” — two themes represented 70% of all posts: “Treatments” (45%) and “Live. Move” (25%).

The total number of contributors was 3645. Some patients were extremely active on the forum. For example, one contributor was the author of 2 386 messages (i.e., 10% of the total number of posts). The ten top contributors posted 32% of the messages.

3.2. Labeling and annotation

The 20 annotators were all adults with a confirmed IBD diagnosis (Table 3) and the sample was evenly distributed by age, sex and the two major forms of IBD.

A total of 4 003 of the 23 656 messages contained the words “flare” and/or “crisis” at least once. We randomly selected 1 000 messages: 47% were labeled yes (469/1 000), meaning that the message concerned a flare. Therefore, words such as “flare” or “crisis” were insufficient to define a strategy to find relevant messages. Consequently, 469 messages were annotated because patients indicated excerpts from the text they considered significant flare markers among the yes-labeled messages.

3.3.Vocabulary analysis

After lemmatization, word cleaning and spelling correction, 73% of the annotated vocabulary was retained. Twenty vocabulary groups appeared to circumscribe the phenomenon of flare in IBD (Table 4).

Crohn’s- and ulcerative colitis-related terms, as well as the words related to “flare” or “crisis” were excluded from the results presentation in Table 4. These specific terms were not informative enough about flare in IBD (crohn’s disease and ulcerative colitis are a diagnosis, and “flare” or “crisis” are general qualifiers).

Results in Table 4 provide a view of the most frequent to less frequent markers; that is, these raw data already encompassed the scope of surrogate markers for flare in IBD with, in addition, a weight notion for each one.

Moreover, raw data on vocabulary groups could be classified into five dimensions (Table 5): impact on life, symptoms, extra-intestinal manifestations, drugs and environmental factors.

4. Discussion

Our main finding is that exploring semantic groups and their frequency within extracted labeled data describing a flare phenomenon in IBD shed light on the most related markers and their degree of importance.

Using web-scraping to study patients' perspectives from a source corresponding to exchanges in an Internet forum made all posts available without being obliged to apply an artificial selection criterion such a specific thread or a fixed periodicity. Moreover, these posts emerged spontaneously in a natural context and were not generated from an artificial situation (interviews and/or focus groups) to meet research needs.

Knowledge could have been generated from metadata analysis. Metadata analysis highlights the over-representation of big contributors in writing posts. This situation raises methodological and epistemological questions in our specific case (i.e., for generating PROs items) but more broadly in all cases of studying patients' perspectives from qualitative data extracted from Internet forums. Indeed, big contributors' overload data (one contributor generated 10% of the forum messages). It is possible that those ones who write often the forum are patients with greater anxiety about their disease state or with associated functional symptoms. But we are not able to provide, at this time, a satisfactory answer on how to moderate the share of big contributors in data. Moreover, a considerable wealth of data (material inaccessible with

classical qualitative data-collection methods (i.e., interviews or focus groups) opens the way to numerous ancillary studies. This is all the more important because the general field of netnography (eth**nography** on inter**net**) consisting of a qualitative research method analyzing the free behavior of individuals on the Internet (e-forum of course, but also social media sites, YouTube and many others) is advancing [45–47].

A double evaluation for labeling and annotation process which had been achieved independently by 2 patients on the same material would have ensured an even stronger data interpretation.

The percentage of labelled yes messages was 47% (469/1000), so the presence of significant words such as “flare” and/or “crisis” was insufficient to select relevant messages to the phenomenon of interest. Hence, this result fully justifies the involvement of the engineering sciences to move away from artisanal methods and toward the systematic and automatic interrogation of available data.

Skills in natural language processing and knowledge discovery are still essentially research skills developed in engineering sciences with no or poor penetration in the field of health. In our study, we propose an innovative application consisting of combining a web-based and machine-learning approach in the field of PROs development. Messages were written in *blog-style* French, but substantial data are also available in English. See for example: https://inflammatoryboweldisease.net/forums/ or https://crohnsforum.com/, or https://www.crohnscolitiscommunity.org/crohns-colitis-forum or https://www.ibdsupport.org.au/community-forum/

In our study, we used web-based and machine-learning approaches that were a rich source of information because we circumscribed the phenomenon of flare in IBD with the identification

of 20 surrogate markers from patients' perspectives. These markers were classified into five dimensions with a view to allocated weight for each one.

Thus, our data from the machine-learning process are relevant material to guide the development of a PROs measuring flare in IBD for 1) the vocabulary to be contained in the items and 2) for factors (dimensions) to include. In other words, mining the web for PROs development met international recommendations stressed by the International Society for Pharmacoeconomics and Outcomes Research, the US Food and Drug Administration and the European Medicines Agency [9–12] by covering the elicitation of key concepts to inform the content and structure of the new PROs, usually covered by individual interviews or focus groups, the two predominant methods to collect the perspectives of the population of interest [48].

The development of a PROs based on patients' perspectives by using a combined web-based and machine-learning approach is an innovative method that opens great stimulating perspectives for scientific exegesis based on data from the web. For example, in the field of IBD, the study of differences in the definition of flare based on the patients' age, gender, disease activity. It also opens numerous methodological perspectives in the field of PROs development as in the completely automatized questionnaire-items generation via algorithms developed from Artificial Intelligence.

Acknowledgments

This work was supported by a grant from the French Ministry of Health (CPRC 2017, 2019-A01520-57). The sponsor is Nancy University Hospital (Research and Innovation Direction). We thank Eva-Marine Pradeau and Margaux Tornqvist for precious contributions in extracting the forum messages; Amandine Verga-Gerard for contributions in managing logistical aspects; and Andreia Carvalho for managing regulatory approvals.

We are also grateful to the RECaP Network – Perceived Health Measurement Working Group (non-author contributors: Hervé Devilliers, Philippe Martin, Hélène Mellerio, Amandine Verga-Gérard) for help with designing the study.

Declaration of competing interest

All Authors report no relevant conflicts of interests.

References

1. Peyrin-Biroulet L, Sandborn W, Sands BE, et al. Selecting Therapeutic Targets in Inflammatory Bowel Disease (STRIDE): Determining Therapeutic Goals for Treat-to-Target. Am J Gastroenterol 2015;110(9):1324–38.

2. D'Haens G, Sandborn WJ, Feagan BG, et al. A review of activity indices and efficacy end points for clinical trials of medical therapy in adults with ulcerative colitis. Gastroenterology 2007;132(2):763–86.

3. Peyrin-Biroulet L, Panés J, Sandborn WJ, et al. Defining Disease Severity in Inflammatory Bowel Diseases: Current and Future Directions. Clinical Gastroenterology and Hepatology 2016;14(3):348-354.e17.

4. Walmsley RS, Ayres RCS, Pounder RE, Allan RN. A simple clinical colitis activity index. Gut 1998;43(1):29–32.

5. Seo M, Okada M, Yao T, Ueki M, Arima S, Okumura M. An index of disease activity in patients with ulcerative colitis. Am J Gastroenterol 1992;87(8):971–6.

6. Lichtiger S, Present D. Preliminary report: cyclosporin in treatment of severe active ulcerative colitis. The Lancet 1990;336(8706):16–9.

7. Higgins PDR, Harding G, Revicki DA, et al. Development and validation of the Ulcerative Colitis patient-reported outcomes signs and symptoms (UC-pro/SS) diary. J Patient Rep Outcomes 2018;2(1):26.

8. Dulai PS, Jairath V, Khanna R, et al. Development of the symptoms and impacts questionnaire for Crohn's disease and ulcerative colitis. Alimentary Pharmacology & Therapeutics 2020;51(11):1047–66.

9. Patrick DL, Burke LB, Gwaltney CJ, et al. Content Validity—Establishing and Reporting the Evidence in Newly Developed Patient-Reported Outcomes (PRO) Instruments for Medical Product Evaluation: ISPOR PRO Good Research Practices Task Force Report: Part 1—Eliciting Concepts for a New PRO Instrument. Value in Health 2011;14(8):967–77.

10. Patrick DL, Burke LB, Gwaltney CJ, et al. Content validity--establishing and reporting the evidence in newly developed patient-reported outcomes (PRO) instruments for medical product evaluation: ISPOR PRO Good Research Practices Task Force report: part 2--assessing respondent understanding. Value Health 2011;14(8):978–88.

11. Research C for DE and. Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims [Internet]. U.S. Food and Drug Administration. 2020 [cited 2020 Jul 2];Available from: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/patient-reported-outcome-measures-use-medical-product-development-support-labeling-claims

12. Anonymous. Regulatory guidance for the use of health-related quality life (HRQL) measures in evaluation medicinal products [Internet]. European Medicines Agency. 2018

[cited 2020 Sep 28];Available from: https://www.ema.europa.eu/en/regulatory-guidance-use-health-related-quality-life-hrql-measures-evaluation-medicinal-products

13. Gul M, Huynh LM, El-Khatib FM, Yafi FA, Serefoglu EC. A qualitative analysis of Internet forum discussions on hard flaccid syndrome. Int J Impot Res 2019;

14. Smerecnik C, Schaalma H, Gerjo K, Meijer S, Poelman J. An exploratory study of Muslim adolescents' views on sexuality: Implications for sex education and prevention. BMC Public Health 2010;10:533.

15. Stockmann T, Odegbaro D, Timimi S, Moncrieff J. SSRI and SNRI withdrawal symptoms reported on an internet forum. Int J Risk Saf Med 2018;29(3–4):175–80.

16. Karmen C, Hsiung RC, Wetter T. Screening Internet forum participants for depression symptoms by assembling and enhancing multiple NLP methods. Comput Methods Programs Biomed 2015;120(1):27–36.

17. Chary M, Yi D, Manini AF. Candyflipping and Other Combinations: Identifying Drug-Drug Combinations from an Online Forum. Front Psychiatry 2018;9:135.

18. Kjellgren A, Henningsson H, Soussan C. Fascination and Social Togetherness-Discussions about Spice Smoking on a Swedish Internet Forum. Subst Abuse 2013;7:191–8.

19. Bilgrei OR. From "herbal highs" to the "heroin of cannabis": Exploring the evolving discourse on synthetic cannabinoid use in a Norwegian Internet drug forum. Int J Drug Policy 2016;29:1–8.

20. Bøhling F. Psychedelic pleasures: An affective understanding of the joys of tripping. Int J Drug Policy 2017;49:133–43.

21. Tighe B, Dunn M, McKay FH, Piatkowski T. Information sought, information shared: exploring performance and image enhancing drug user-facilitated harm reduction information in online forums. Harm Reduct J 2017;14(1):48.

22. Lee E, Cooper RJ. Codeine Addiction and Internet Forum Use and Support: Qualitative Netnographic Study. JMIR Ment Health 2019;6(4):e12354.

23. Widemalm M, Hjärthag F. The forum as a friend: parental mental illness and communication on open Internet forums. Soc Psychiatry Psychiatr Epidemiol 2015;50(10):1601–7.

24. Jansen NA, Saint Onge JM. An internet forum analysis of stigma power perceptions among women seeking fertility treatment in the United States. Soc Sci Med 2015;147:184–9.

25. Matza LS, Chung KC, Kim KJ, et al. Risks associated with antiretroviral treatment for human immunodeficiency virus (HIV): qualitative analysis of social media data and health state utility valuation. Qual Life Res 2017;26(7):1785–98.

26. Dudina VI, Judina DI, King EJ. Fears about antiretroviral therapy among users of the internet forum for people living with HIV/AIDS in Russia. AIDS Care 2017;29(2):268–70.

27. Kantrowitz-Gordon I. Internet confessions of postpartum depression. Issues Ment Health Nurs 2013;34(12):874–82.

28. Anderson APA, Feldman MD, Bryce J. Munchausen by Proxy: A Qualitative Investigation into Online Perceptions of Medical Child Abuse. Journal of Forensic Sciences 2018;63(3):771–5.

29. Pilkington PD, Rominov H. Fathers' Worries During Pregnancy: A Qualitative Content Analysis of Reddit. J Perinat Educ 2017;26(4):208–18.

30. Mirskaya M, Lindgren E-C, Carlsson I-M. Online reported women's experiences of symptomatic pelvic organ prolapse after vaginal birth. BMC Womens Health 2019;19(1):129.

31. Law GC, Jones CJ, Bülbül A, Smith HE. "At a loss of what to do": a qualitative analysis of parents' online discussion forums about their administration of asthma inhalers to their young children. J Asthma 2019;1–10.

32. Westerlund M, Hadlaczky G, Wasserman D. Case study of posts before and after a suicide on a Swedish internet forum. Br J Psychiatry 2015;207(6):476–82.

33. Horne J, Wiggins S. Doing being 'on the edge': managing the dilemma of being authentically suicidal in an online forum. Sociology of Health & Illness 2009;31(2):170–84.

34. Cianciara D, Szmigiel A. Posting on „Nie szczepimy („We don't vaccinate") Internet forum. Przegl Epidemiol 2019;73(1):105–15.

35. Blankers M, van der Gouwe D, van Laar M. 4-Fluoramphetamine in the Netherlands: Text-mining and sentiment analysis of internet forums. Int J Drug Policy 2019;64:34–9.

36. Kamiński M, Borger M, Prymas P, et al. Analysis of Answers to Queries among Anonymous Users with Gastroenterological Problems on an Internet Forum. Int J Environ Res Public Health [Internet] 2020 [cited 2020 Jun 30];17(3). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7037061/

37. Bini SA. Artificial Intelligence, Machine Learning, Deep Learning, and Cognitive Computing: What Do These Terms Mean and How Will They Impact Health Care? J Arthroplasty 2018;33(8):2358–61.

38. Gubatan J, Levitte S, Patel A, Balabanis T, Wei MT, Sinha SR. Artificial intelligence applications in inflammatory bowel disease: Emerging technologies and future directions. World J Gastroenterol 2021;27(17):1920–35.

39. Helm JM, Swiergosz AM, Haeberle HS, et al. Machine Learning and Artificial Intelligence: Definitions, Applications, and Future Directions. Curr Rev Musculoskelet Med 2020;13(1):69–76.

40. Danese S, Fiorino G, Peyrin-Biroulet L. Early intervention in Crohn's disease: towards disease modification trials. Gut 2017;66(12):2179–87.

41. Ricci L, Epstein J, Buisson A, et al. Flare-IBD: development and validation of a questionnaire based on patients' messages on an internet forum for early detection of flare in inflammatory bowel disease: study protocol. BMJ Open 2020;10(7):e037211.

42. Merriënboer JJGV, Sweller J. Cognitive load theory in health professional education: design principles and strategies. Medical Education 2009;1(44):85–93.

43. Rat A-C, Ricci L, Guillemin F, et al. Development of a Web-Based Formative Self-Assessment Tool for Physicians to Practice Breaking Bad News (BRADNET). JMIR Med Educ 2018;4(2):e17.

44. Gerlach M, Shi H, Amaral LAN. A universal information theoretic approach to the identification of stopwords. Nature Machine Intelligence 2019;1(12):606–12.

45. Martinez B, Dailey F, Almario CV, et al. Patient Understanding of the Risks and Benefits of Biologic Therapies in Inflammatory Bowel Disease: Insights from a Large-scale Analysis of Social Media Platforms. Inflamm Bowel Dis 2017;23(7):1057–64.

46. Schuman DL, Lawrence KA, Pope N. Broadcasting War Trauma: An Exploratory Netnography of Veterans' YouTube Vlogs. Qual Health Res 2019;29(3):357–70.

47. Tenderich A, Tenderich B, Barton T, Richards SE. What Are PWDs (People With Diabetes) Doing Online? A Netnographic Analysis. J Diabetes Sci Technol 2019;13(2):187–97.

48. Ricci L, Lanfranchi J-B, Lemetayer F, et al. Qualitative Methods Used to Generate Questionnaire Items: A Systematic Review. Qual Health Res 2018;1049732318783186.

Figure legends

Figure 1. Illustration of syntax used for data scraping

Figure 2.: Extracted data before and after cleaning

Figure 3. Visual Basic for Applications interface on Excel to facilitate annotation

Figure 4. Extract from the text file containing the lemmatized and cleaned messages

Figure 1. Illustration of syntax used for data scraping

Figure 2.: Extracted data before and after cleaning

Before Cleaning


```
                              "], "date": ["10/11/2014 \u00e0 19:64"], "pseudo":
"Anonymous"], "image": [], "theme": "Parents . Enfants", "discussion": [“\n
       De retour du S\u00e9jour des famille !\n        "], "vues_discussion": [
991"]},
“comment”: [“\n
                        Oups je vois que La fin de \"sp\u00e9cialistes\" a saut\
00e9 !", "\n", "\nPour notre part nous habitons dans le 44 et notre fille de 4,
 ans (qui a la RCH) est suivie depuis environ 1 an par Dr Ga\u00ebblle Le Henaff
\u00e0 La clinique St Augustin \u00e0 NANTES. Super p\u00e9diatre gastro-ent\u0
e9rologue tr\u00e8s pro et douce. Elle consulte m\u00eame le samedi matin, prat
que. Une semaine apr\u00e8s notre 1er rdv avec elle (apr\u00e8s 2,5 mois de t\u
0e2tonnement par le g\u00e9n\u00e9raliste au d\u00e9but des sympt\u00f4mes) not
e fille a eu une coloscopie + fibro orientant le diagnostic, confirm\u00e9 1 se
aine plus tard apr\u00e8s analyse de la biopsie. Le jour de la coloscopie notre
fille à commenc\u00e9 ses anti-inflammatoires qui ont stopp\u00e9 Les saignemen
s au bout de 10j. Nous avons stopp\u00e9 le r\u00e9gime sans l\u00e9gume ni Lai
ages 1 mois apr\u00e8s mais elle est tjs sous Pentasa \u00e0 ce jour. La crise
st officiellement termin\u00e9e depuis d\u00e9but d\u00e9cembre. Rdv de contr\u
0f4le ce samedi !\n
                             ","\n
 "], "date": ["06/06/2016 \u00e0 12:34"], "pseudo": ["auriclem"], "image": [],
theme": "Parents . Enfants", "discussion": ["\n            M\u00e9decins sp\u00
9ciali\n        "], "vues_discussion": ["807"]},
```


After Cleaning

8	[size=4]bonjour à tous; juste ce petit mot pour parler de ce week-end de parents ! je viens de passer le week-end avec d'autres parents, qui, comme moi vivent cette maladie auprès de leurs enfants. pour dire a quel point ça fait du bien d'échanger avec d'autres qui comprennent. Pas besoin d'expliquer, on sait ! alors voilà, merci à l'afa de nous permettre ces moments. merci à Christelle et Anne qui nous ont donné leur week-end et leur chaleur humaine. Merci au Dr Hugot qui a joué le jeux de l'information et de l'écoute avec beaucoup de qualité. [/size]	16/05/2010 à 21:50	396038AR	Parents . Enfants	week-end des parents
9	bonjour, il est difficile de répondre à ce genre de question, le mieux serait probablement que vous appeliez son gastro pour lui poser directement la question. bon courage, les choses vont s'améliorer lorsque le traitement fera effet (ça prend, pour le pentasa et sauf erreur de ma part quelques semaines) Pascale	22/05/2011 à 19:41	pascale59	Parents . Enfants	mon fils de 10 ans a une RCH
10	bonjours a tous , je souhaiterais savoirs pour mon fils de 10 ans ateind d'une RCH ,si le traitement pentasa et betnesol 5mg n'est pas trop lourd pour lui car il a montrer des signe de relachement musculaires tremblement et douleurs aigus au ventre es ce le traitement ki le lui a provoquer ou la maladie?	21/05/2011 à 12:48	pupuce	Parents . Enfants	mon fils de 10 ans a une RCH
11	ma fille vient de faire 6 ans depuis peut, atteint d'une rch nous avons voulu lui faire plaisir pour sont annif malgrer quelle soit sous cortencyl a forte dose nous avons fait un ecart sur le menu (gateau au chocolat sans bonbons) dur dur pour un enfant de voir c copains ,copines feter leur annif avec toute sort de sucrerie j'en vien meme a décliner des annif je c quelle ne fera pas d'ecart ,car elle c se qu'il faut pas manger. mais je la vois triste a chaque fois quelle revien sa me fait trop mal au coeur ,j'en est marre sont etat s'arrange pas. la elle reva a la selles 6 fois pas jour surtout la nuit beaucoup de sang a chaque fois g peur pour elle ,quelle resubise tous c examens, perfusions ,etre agen pendant 10jours etre loin de chez nous (300km).Maintenant je vien a culpabiliser d'avoir fait cet ecart ,car dans 15 jour nous partons pour disney c une surprise elle le c pas pour pas quelle soit deçu si sont etat le permet pas.(car deja noel a l'hopital dur dur)	18/07/2011 à 00:09	emilie66	Parents . Enfants	marre de culpabiliser

Figure 3. Visual Basic for Applications interface on Excel to facilitate annotation

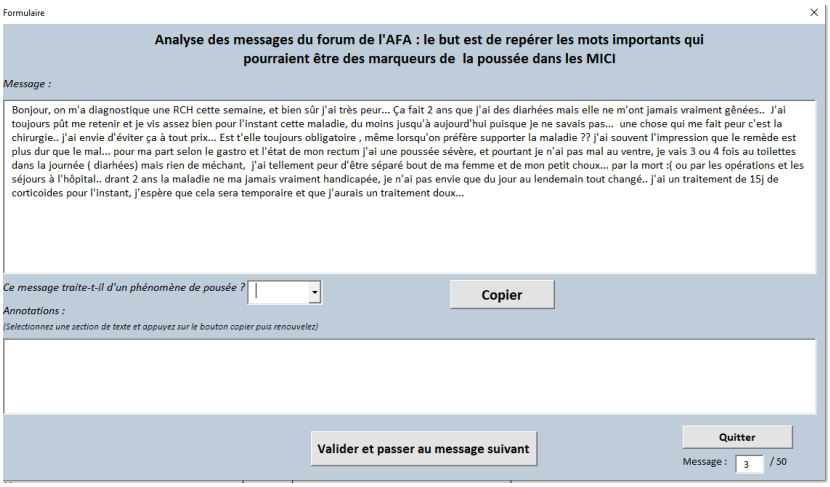
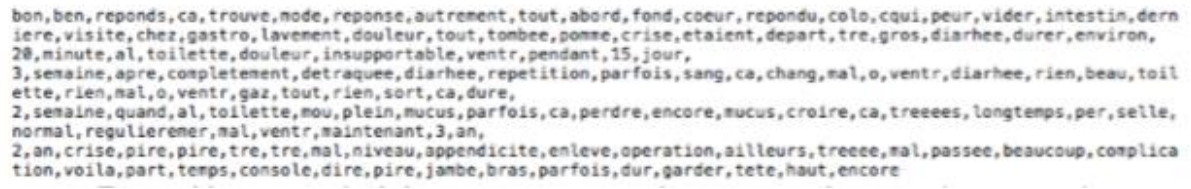


Figure 4. Extract from the text file containing the lemmatized and cleaned messages



```
bon,ben,reponds,ca,trouve,mode,reponse,autrement,tout,abord,fond,coeur,repondu,colo,cqui,peur,vider,intestin,dern
iere,visite,chez,gastro,lavement,douleur,tout,tombee,pomme,crise,etaient,depart,tre,gros,diarhee,durer,environ,
20,minute,al,toilette,douleur,insupportable,ventr,pendant,15,jour,
3,semaine,apre,completement,detraquee,diarhee,repetition,parfois,sang,ca,chang,mal,o,ventr,diarhee,rien,beau,toil
ette,rien,mal,o,ventr,gaz,tout,rien,sort,ca,dure,
2,semaine,quand,al,toilette,mou,plein,mucus,parfois,ca,perdre,encore,mucus,croire,ca,treeees,longtemps,per,selle,
normal,regulieremer,mal,ventr,maintenant,3,an,
2,an,crise,pire,pire,tre,tre,mal,niveau,appendicite,enleve,operation,ailleurs,treeee,mal,passee,beaucoup,complica
tion,voila,part,temps,console,dire,pire,jambe,bras,parfois,dur,garder,tete,haut,encore
```

Table 1. Steps and tools used for web-based and machine-learning approach.

Steps	Tools	Illustration Figures
Data scraping and cleaning	Scrapy web-crawling tool combined with Splash running under Python	Figure 1
	Data cleaning using a Python algorithm	Figure 2
Labeling and annotation	VBA interface on Excel	Figure 3
Vocabulary analysis	Lemmatization: S*paCy* free open-source library for French Natural Language Processing in Python	Figure 4
	Cleaning: removing of stop words and unnecessary words plus spelling correction	
	Vocabulary group generation based on every lexical field that was well represented	
	Application of the counter () function	

VBA, Visual Basic for Applications

Table 2. Thread with more than 100 posts (total posts = 23 656)

Thread	No. of posts	%
Vedolizumab - Experience feedback	374	1.58
How I numbed my UC	314	1.33
Pregnancy research under Humira (adalimumab)	199	0.84
Living your experiences without treatment	175	0.74
Humira (adalimumab) - efficiency / timeliness?	174	0.74
23 years old, male, CD - Very annoying anal fistula	111	0.47
Living with an ostomy	101	0.43

UC, ulcerative colitis; CD, Crohn's disease

Table 3. Description of annotators

Age	Sex	Form of IBD
Mean = 43	Female = 12	Crohn's disease = 11
		Ulcerative colitis = 8
	Male = 8	Not determined = 1

Table 4. Vocabulary groups of significant flare marker annotations (469 annotated messages)

Group	Lemmatized French vocabulary	English translation	Occurrence*	%**
Fatigue	'fatig','crev','nergi', dor'	'tire','exhaus','nergy', 'sleep'	118	25
Pain	'doul','souff'	'pain','suff'	108	23
Toilet	'toilette','wc','envie','selle'	'toilet','wc','saddle', 'ride'	98	21
Belly	'ventr','ballon','estoma','intes','colon','bide'	'belly', 'bloat','stomac', 'bowel','colon','tummy'	86	18
Diarrhea	'diar','coliq'	'diar','colic'	57	12
Blood	'sang','saign','hemo'	'blood','bleed','hemorr'	55	12
Stress	'stress','angoiss','inquiet'	'stress','anxiet','worr'	49	10
Weight loss	'pert','poids','kilo','kg'	'loss','weight','kilo','kg'	40	9
Food	'nourr', 'appet','mang', 'alim','reg','nutri','faim'	'food','appet','eat','food','diet','nutri','hunger'	41	9
Vomiting	'vomi','naus'	'vomi','naus'	25	5
Joint	'Arti'	'Join'	21	4
Mood	'humeur','moral','psy','culpa','hont','enerv','coler','depress', 'peur'	'mood','moral','psy','culpa',"'sham', 'angr', 'anger', 'depress', 'fear'	20	4
Tobacco	'tabac', 'fume', 'cigar'	'tobac', 'smok', 'cigar'	16	3
Fever	'fiev', 'tempé', 'fièv', 'temperature'	'fever', 'temperature'	14	3
Back	'dos'	'back'	10	2
Cortisone	'corti', 'penta'	'corti', 'penta'	8	2
Mucus	'glair','mucu'	'glair','mucu'	8	2
Knees	'genou'	'knee'	7	1
Legs	'jambe'	'leg'	5	1
Drugs	'med', 'trait', 'pillul'	drug', 'treatm', 'pill'	4	1

*the total of the occurrence column exceeds 469 because a message can contain several vocabulary groups at the same time

**the total of the percentage column exceeds 100% because a message can contain several vocabulary groups at the same time

Table 5. Vocabulary groups dimension (469 annotated messages)

Dimension	Occurrence*	%**
Impact on life	366	78
Fatigue	118	25
Toilet	98	21
Stress	49	10
Weight loss	40	9
Food	41	9
Mood	20	4
Symptoms	353	75
Pain	108	23
Belly	86	18
Diarrhea	57	12
Blood	55	12
Vomiting	25	5
Fever	14	3
Mucus	8	2
Extra-intestinal manifestations	43	8
Joint	21	4
Back	10	2
Knees	7	1
Legs	5	1
Environmental factors	16	3
Tobacco	16	3
Drugs	12	3
Drugs	4	1
Cortisone	8	2

*the total of the occurrence column exceeds 469 because a message can contain several vocabulary groups at the same time

**the total of the percentage column exceeds 100% because a message can contain several vocabulary groups at the same time