Social Media Mining in Drug Development Decision Making: Prioritizing Multiple Sclerosis Patients' Unmet Medical Needs

Jonathan Koss
Witten/Herdecke University
jonathan.koss@uni-wh.de

Sabine Bohnet-Joschko
Witten/Herdecke University
sabine.bohnet-joschko@uni-wh.de

Abstract

Pharmaceutical companies increasingly must consider patients' needs in drug development. Since patients' needs are often difficult to measure, especially in rare diseases, information in drug development decision-making is limited. In the proposed study, we employ the opportunity algorithm to identify and prioritize unmet medical needs of multiple sclerosis patients shared in social media posts. Using topic modeling and sentiment analysis features of the opportunity algorithm are generated. The result implies that sensory problems, pain, mental health problems, fatigue and sleep disturbances represent the highest unmet medical needs of the samples population. The present study suggests a promising potential of this method to provide relevant insights into rare disease populations to promote patient-centered drug development.

1. Introduction

Pharmaceutical companies increasingly need to consider perspectives and needs of patients in drug development, responding to changes in the business environment. For example, patient-reported outcomes are considered by regulatory authorities in the approval, pricing, and reimbursement decision-making [1]. Furthermore, in the (Food and Drug Administration) FDA's Patient-Focused Drug Development initiative patients perspectives on treatment benefits are considered in the approval process [2]. Therefore, the patient's perceived benefit of a new treatment is increasingly important for commercial success. Thus, pharmaceutical companies need to address the unmet medical needs of a patient population, not only by increasing clinical outcomes, e.g., controlling the underlying disease progression, but improving outcomes relevant from patients' everyday perspective. Therefore, pharmaceutical companies have to consider unmet medical needs (UMNs) in their portfolio management systematically [3] as well as in overall drug development – referred to as "Patient-Centered-Drug-Development" (PCDD) [4]. The patient UMN is determined by the difference between the health status achieved with the current standard of care (SoC) and the patient's ideal health status (absence of the disease) [3]. In the context of PCDD, ideal health status is determined by the patients´ preferences [4]. However, evidence regarding UMNs and the impact on patients' life is often limited, especially in rare diseases [5]. In addition, clinicians and patients perceive UMNs differently [6]. This may lead to pursuing the improvement of less relevant therapeutic outcomes in clinical developments [6]. For example, a recent study on the quality of life of patients with multiple sclerosis (MS) demonstrated that MS patients and physicians differed in their assessment of the impact of different UMNs on quality of life [5]. In the worst case, real patient needs are disregarded in drug development and opportunities to innovate and to improve patients' quality of life are missed. Simultaneously, the development of rare diseases therapies is crucial, to improve patients' life's as well as for commercial success of pharmaceutical companies. Various health systems are financially incentivizing research of therapies, making the rare disease segment a future growth market [7]. Companies participating in the rare disease market are already associated with higher market value and profits than enterprises without rare disease treatments in their portfolio [8].

In this context, social media data could be valuable to inform drug development decision-making regarding UMNs perceived by patients [9]. Extant research revealed that patients suffering from severe diseases use social media platforms to exchange information about their diseases-trajectories [10, 11]. They exchange experiences about symptoms and therapies or use these platforms as a medium to support each other emotionally [10, 11]. The accumulated available knowledge has already been

URI: https://hdl.handle.net/10125/79702
978-0-9981331-5-7

employed for disease-specific research [12-15]. Initial studies are exploring the use of social media to explore patient needs and preferences of specific disease populations for drug development [16, 17]. Cook et al. [17] were the first to prove the use of social media posts to derive relevant items for patient preference studies in the early drug development process of dry eye disease (DED) and non-alcoholic steatohepatitis [17]. The automated extraction and analysis of social media data is referred to as social media mining [18]. In this context, we are conducting an initial feasibility study to investigate whether relevant UMNs of MS patients can be identified and prioritized exclusively by using social media mining methods. In addition to the commercial value of rare disease treatments for pharmaceutical companies, the underlying nature and symptoms associated with the disease are particularly suited to social media mining [9]. The severe physical and psychological limitations along to the rarity of the disease cause MS patients using social media as a pivotal medium for both information gathering and sharing, e.g. about treatment trajectories, as well to receive emotional support by the community [9, 10]. Specifically, we will implement the theoretical framework of the opportunity algorithm (OA) using an appropriate social media mining pipeline with several natural-language-processing methods. Generating insights, OA could be used to guide the drug development process by prioritizing therapy outcomes that are important to patients while reducing the costs incurred by analogous methods.

2. Background

Relevant unmet medical needs for decision-making in pharmaceutical drug development, include mortality, time suffering from disease, symptom and disease burden, side effects, treatment inconvenience and patient perception [3]. Life expectancy can be used as a mortality measure. In the MS population life expectancy is reduced by 7 to 14 life years compared to healthy population [19]. Symptom and disease burden consist of disease-related symptoms experienced by the individual patient. Green et al. [20] assessed the correlation between symptom severity and self-perceived health in MS patients by performing pearson correlations and multivariate linear regressions. They considered walking, hand function, vision, fatigue, cognition, bladder, sensory, pain, depression, and anxiety as coefficients. Pain contributed most to health perception in outpatients with multiple sclerosis, followed by walking and fatigue [20]. Roemmer et al. [21] investigated side effects of MS therapies. Infections are common while being treated with immunosuppressive drugs. Malignancies were reported in some of the pivotal studies. Other adverse effects, such as skin reactions, are usually not life threating but may lead to decreased therapy adherence [21]. Treatment inconvenience is majorly affected by route of administration [22, 23]. For instance, injectable treatments are very inconvenient. As oral treatments became available, patients started switching because of the route of administration [22, 23]. Patient perception refers to the patient's perceived health as often measured by established health-related quality of life questionnaires (QoL), such as the Short Form Questionnaire (SF-36) [3]. Gil-González [24] conducted a systematic review regarding the QoL of adults MS patients. They concluded fatigue, pain, and cognitive impairment significantly worsen quality of life and should therefore be prioritized in diseases treatment [24].

According to Christensen et al., innovation is successful if it addresses a perceived unmet need of the customer, also called "the-job-to-be-done" logic [25]. As the patients perceived benefit of a new treatment is increasingly important for commercial success, goal of drug development should be a treatment that reduces a perceived unmet medical need of patients, and therefore "gets the job done". Following the “job-to-be-done” logic, Ulwick [26] derived the opportunity algorithm (OA) to prioritize unmet needs as opportunities to innovate. The underlying assumption of OA is that successful innovation occurs when the new product addresses an important - and at the same time - relatively unsatisfied need [26]. Therefore, the OA can be used to prioritize unmet needs to identify the “job-to-be-done”, which should be completed to improve patients’ life [9]. The OA is defined as follows [27]:

Opportunity = Importance + Max (Importance − Satisfaction, 0)

Few studies accessed prioritization of innovation opportunities by analyzing social media data in means of the OA. Choi et al. [28] concluded that innovation opportunities in the smart speaker market exist in offering customized commands [28]. Jeong et al. found that innovations for smartphones should include payment services and fast charging functions [27]. Given the encouraging results in consumer markets, we evaluate the OA in identifying and prioritizing MS patients’ unmet medical needs. To the best of our knowledge, this is the first study investigating the OA for detecting innovation potential in the context of drug development.

3. Research design and method

To identify and prioritize MS patients’ unmet medical needs, we collected and analyzed posts in a

MS online forum. Data extraction was performed using web scraping. Similar to Jeong et al. [27] accessing innovation opportunities in the smartphone market [27], topic modeling is used to identify UMNs in MS patients social media posts. Furthermore, we assume that a topic that is mentioned more frequently than another topic is also more important [27] from the patient's perspective. Therefore, as a surrogate for the importance of UMN in the context of the opportunity algorithm, we use the topic frequency.

Since the frequency alone is not decisive for an unmet need, the OA additionally considers the satisfaction regarding the need [26]. Satisfaction is derived using sentiment analysis [27]. The results of the OA will be evaluated using extant literature.

3.1 Data Source and Extraction

Since unmet medical needs comprise several different dimensions [3], social media platforms with information containing comprehensive aspects of patients' disease trajectories are advantageous. Therefore, we opted for online forums, which tend to contain longer, more detailed posts [10]. Moreover, unlike leading commercial platforms employing advertising and algorithms [29] to manipulate user behavior, online forums do not use algorithms to influence discussion and contain less or even forbid advertising [9]. After inspecting samples of various MS-specific online forums, we chose a forum whose content met our criteria (detailed description of the disease trajectory and few advertisements) and whose data were accessible through web scraping. Using Parsehub, a commercial tool for web scraping, we collected 119,501 posts (consisting of original and follow-up posts) from 01-01-2013 to 31-12-2020 (Fig.1).

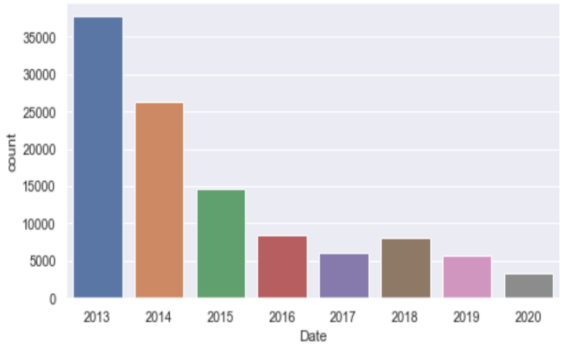

Figure 1. Number of posts collected per year

3.2 Data Preprocessing

As described above we will analyze UMNs due to the OA. Since social-media content mainly comprises unstructured text, the use of natural language processing techniques is essential during preprocessing to reduce noise and structure the data to facilitate insightful analysis [30]. As the requirements of topic model and sentiment analysis regarding the dataset are significantly different, pre-processing differs as well. Hence, we describe applied preprocessing methods separately. However, the dataset for topic modeling and sentiment analysis is prepared using deduplication, reducing data set to 110,303 unique posts.

3.2.1. Identifying UMNs and their importance

Following, the topic models' pipeline is outlined, aiming to create features for the "importance" dimension of the OA.

Word tokenization and lowercasing

Since our dataset consists of phrases and the topic model requires a feature vector consisting of separate words, we split the phrases into tokens (words). We used the function "word_tokenize()" of the NLTK library. All tokens were lower-cased.

Stop words removal

Stop words occur very frequently (e.g. articles and negations) and are meaningless in analysis. Therefore, we excluded stop words using the stop word list of the NLTK library. Furthermore, we created a custom list of stopwords, including common words such as "doctor" and "symptom".

Noun- and Verb- Phrase Extraction

Noun phrase extraction can be used to improve topic coherence [31]. We defined noun phrases as a combination of nouns and adjectives. Furthermore, we decided to extracted verb-phrases, defined as a verb followed by adjective. Verb phrases were assumed to be useful in describing symptoms. In implementation we used the "nltk.RegexpParser()" function of the ntlk python library.

Lemmatization

Since words appear in several inflected forms, eventually worsening topic model results, we used lemmatization to improve topic quality by converting words into the base form [32]. Lemmatization was conducted with the "lemmatizer()" function of the spaCy python library.

Keyword-Dictionary and Vectorization

Studies in the medical area revealed, that topic quality can be improved by defining a dictionary containing relevant keywords, considered in the topic model as features [10]. As we did not have an existing dictionary of keywords for MS related social media content we used the YAKE! keyword extraction algorithm to obtain a keyword dictionary [33]. YAKE! is an unsupervised keyword extraction method, based on statistical text features to select the most relevant keywords of a text [33]. We decided to use 4200 keywords after examining keywords. Subsequently, we cleaned the corpus of any words that did not represent keywords. To enable content analysis with Latent Dirichlet Allocation (LDA) we created a "Bag-of-Words" vector representation.

Bigram

Bigrams consist of two words occurring very frequently in conjunction with one another. This feature transformation method can be used to improve topic coherence [34]. To create bigrams we used the genism´s "models.phrases()" function.

Topic modeling

Topic modeling refers to a class of statistical algorithms [35] used to infer abstract topics in a collection of text documents [27]. In our case, the collection of documents, $D = \{d_1, d_2, \ldots, d_n\}$, are users' posts. To infer topics, Latent Dirichlet Allocation (LDA)-based topic modeling is applied, as it has already been used in similar scenarios [12, 14, 27]. In LDA topic modeling, each document consists of a mixture of topics, $T = \{t_1, t_2, \ldots, t_m\}$, described by discrete probability distributions θd [36]. Each topic consists of a mixture of vocabulary words, $W = \{w_1, w_2, \ldots, w_k\}$, described by a discrete probability distribution βt [36]. The generative process can be described as follows [35-37]:

(1) For each topic t ∈ {1, …, m},
  (a) draw a probability distribution over vocabulary words
    βt ~ Dirichlet(η).
(2) For each document d,
  (a) draw a vector of the topic probability distribution
    θd ~ Dirichlet(α).
  (b) For each word w_i in document d,
    (i) draw a topic assignment
      zi ~ Multinomial(θd);
    (ii) draw a word w_n ~ Multinomial(β_{zi})

The notations η and α represent the models' hyperparameters to determine the Dirichlet distributions [37]. The LDA outputs are a topic-document matrix and a topic-term matrix. The size of the topic-document matrix is $n \times m$, with the weight $w_{i,j}$ being the association of a document d_i and a topic t_j [36]. The topic-word matrix has the size $m \times k$, the weight $w_{i,j}$ is the association of the topic t_i and a word w_j [36]. For practical implementation we used MALLET topic model python package which uses Gibbs sampling to estimate LDA parameters [38]. The number of topics must be determined a priori. To determine the optimal number of topics, we considered topic coherence measure by Roeder et al. [39] using the topic coherence pipeline of the Genism package and evaluated it using the elbow method [40] followed by manual evaluation by the authors. We decided for 39 topics and α of 0.0105 resulting in a topic coherence of 0.701.

Computing importance

The identified topics represent the UMNs. The importance of an UMN is calculated based on the frequency with which users mention the UMN in relation to all other UMNs (topic frequency).

Therefore, the sum of the contribution probabilities of each UMN to all posts is a measure of the importance of the UMN within the collected corpus [27]. The importance scores of patient needs are scaled to a range of 0-10 obtaining values for the importance dimension of the OA [27].

3.2.2 Deriving satisfaction regarding UMNs

Following the sentiment analysis pipeline is outlined, to create features for the "satisfaction" dimension of the OA.

Unicode Transformation

Unlike in topic modeling, the corpus was preprocessed only by data transformation of the text data into UTF-8 Unicode format. For example, tokenization and vectorization using the bag-of-words model is not necessary because the used sentiments analysis algorithm process continuous text data.

Sentiment Analysis

In order to derive user satisfaction regarding a certain topic, the corresponding sentiments have to be calculated using sentiment analysis [27]. Since sentiment analysis is broadly used to analyze the polarity of sentiment statements in social media posts a variety of methods exists [41]. It can be chosen of several rule-based and machine learning methods or a combination of both approaches [42]. We chose IBM Watson sentiment analysis because of the algorithm's proven track record in analyzing user sentiments in social media data [41, 43]. The sentiment analysis

algorithm is part of the IBM natural language understanding services and can be implemented due to application programming interface in python [41]. Outputs can be retrieved as JSON files containing an array of documents and corresponding sentiment score as continuous values. Sentiment scores ranging from −1 (negative) to 0 (neutral) to +1 (positive).

Computing satisfaction

Satisfaction is derived by the sentiment score [27]. Therefore, the sentiment score associated with each document is structured in an array. To calculate the sentiment score per topic, we multiply the sentiment document score by the weight of the topic's contribution to the document. Consequently, we obtain a sentiment-topic score by assigning a sentiment to each document, weighted by the topic-document contribution. The satisfaction score regarding an UMN is the sum of sentiments of the sentiment-topic score. The satisfaction scores of patient needs are scaled to a range of 0-10 obtaining values for the satisfaction dimension of the OA [27].

3.3 Analysis of unmet medical needs using the opportunity algorithm

Applying the OA to the importance and satisfaction scores, the innovation opportunity potential regarding a patients need is calculated [27, 28]. The highest opportunities are represented by UMNs incorporating high importance and low satisfaction [27, 28]. According to the OA, such UMNs should be addressed in future drug development.

4. Results: Identification and Prioritization of Unmet Medical Needs

Following Wahbeh et al [36], the results are visualized using word clouds. The larger a given word the greater its contribution to the topic. The authors evaluated the word clouds independently to obtain the most reliable labels possible. Topics that were interpreted differently were discussed and re-evaluated. In total 13 topics of 39 topics corresponded to UMNs of MS patients, belonging to the dimensions "symptom and disease burden", "side effects" and "treatment inconvenience"[3]. Topic-document - contribution matrix revealed that 32416 posts were dominated by UMN topics. Most UMNs belonging to the dimension "symptom and disease burden" (Table 1).

Table 1. Dimensions and Topics

Symptom & disease burden	Side effects	Treatment inconvenience
Sensory problems	Flush & gastro-intestinal problems	Route of administration: injection
Pain	Cancer	
Sleep-disturbance and fatigue		
Mental health		
Bladder infection		
Word-finding-problems		
Bladder infection		
Walking impairment		
Vision problems		
Fatigue & dizziness		
Cognitive impairment		

Topics corresponding to the dimension "symptom and disease burden are "pain", "anxiety and depression", "fatigue and dizziness", "sleep-disturbance and fatigue", "cognitive impairment", "sensory problems", "bladder infection", "vision problems" and "walking impairment".

Posts belonging to the topic "pain" were dominated by words such as "pain", "muscle", "leg" and "spasm" (Fig. 2).

Figure 2. Topic "pain"

Posts corresponding to the topic "word-finding-problems" consisted of the terms "word", "remember", difficulty" and "talk" (Fig. 3).

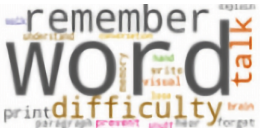

Figure 3. Topic "word-findings-problems"

The topic "mental health" included posts, dominated by the words "anxiety", "depression", "understand", "deal" and "talk" (Fig. 4).

Figure 4. Topic "mental health"

Posts belonging to the topic "fatigue and dizziness" were dominated by words such as "fatigue", "dizziness" and "seizure" (Fig. 5).

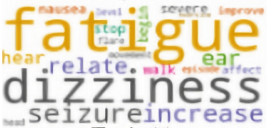

Figure 5. Topic "fatigue and dizziness"

The topic "cognitive impairment" posts are dominated by word such as "cognitive", "memory", "change" and "learn" (Fig. 6).

Figure 6. Topic "cognitive impairment"

The topic "sensory problems" represents posts, dominated by the words "leg", "hand", "foot", "sensation", "numbness" and "tingle" (Fig. 7).

Figure 7. Topic "sensory problems"

Posts belonging to the topic "bladder infection" are dominated by the terms "bladder", "infection" and antibiotic" (Fig. 8).

Figure 8. Topic "bladder infection"

The topic "vision problems" included posts, dominated by the words "eye", "vision" and "ophthalmologist" (Fig. 9).

Figure 9. Topic "vision problems"

Posts belonging to the topic "walking impairment" are dominated by the terms "walk", "exercise", "leg" and "strength" (Fig. 10).

Figure 10. Topic "walking impairment"

The topic "sleep-disturbance and fatigue" included posts, dominated by the words "sleep", "fatique" and "wake" (Fig. 11).

Figure 11. Topic "sleep-disturbance and fatigue"

UMNs belonging to the dimension "side effects" are "flush and gastrointestinal problems" and "cancer".

The topic "flush and gastrointestinal problems" includes posts, dominated by the words "eat", "flush", "stomach", "meal" and "food" (Fig. 12)

Figure 12. Topic "flush and gastrointestinal problems"

Posts belonging to the topic "cancer" are dominated by the terms "cancer", "death", "die" and "cure" (Fig. 13).

Figure 13. Topic "cancer"

UMN belonging to the dimension "Treatment inconvenience" is "route of administration: injection"

The topics “route of administration: injection” posts are dominated by word such as “injection”, “inject”, “shoot”, “needle” and “switch” (Fig. 14).

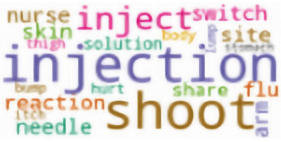

Figure 14. Topic “route of administration: injection”

4.1 Innovation opportunities for future treatments

The results indicating that future treatments in MS should address UMNs in the dimension “symptoms and disease burden”, offering the most potential for improvement. The highest opportunity score (9.86) is assigned to “sensory problems”. This topic possesses the highest topic-document contribution among the UMN topics and therefore the highest level of importance (4.93). At the same time, among all topics, it possesses the lowest satisfaction score (0) (table 2).

The following topics, ranked by the opportunity score, are “pain”, “sleep-disturbance and fatigue”, “bladder infection”, “walking impairment”, “Route of administration: injection”, “fatigue and dizziness”, “flush and gastrointestinal problems” (table 2). The results imply, that improvements in “cognitive impairment”, “vision problems” and “cancer” would not address relevant UMNs as the opportunity score results is negative, because of the low importance and relatively high satisfaction scores (table 2).

Table 2. Opportunity Score Ranking of UMN

Topics	OS-Score	Imp.	Satis.
Sensory problems	9.86	4.93	0
Pain	7.37	3.9	0.43
Sleep-disturbance and fatigue	6.85	4.51	2.16
Mental health	5.4	4.32	3.24
Bladder infection	2.71	1.82	0.94
Word-findings-problems	2.47	2.8	3.19
Walking impairment	2.05	3.39	4.74
Route of administration: injection	1.85	2.26	2.67
Fatigue & dizziness	0.36	0.78	1.2
flush and gastrointestinal problems	0.29	2.02	3.75
Vision problems	-0.7	0.74	2.2
Cancer	-1.39	0.81	3.03
Cognitive impairment	-4.15	0.23	4.6

5. Discussion

In the proposed study, we employed the opportunity algorithm to identify and prioritize unmet medical needs of multiple sclerosis patients. Using topic modeling and sentiment analysis features of the opportunity algorithm are generated. Identified UMNs were assigned to the dimensions “symptom and disease burden”, “side effects” and “treatment inconvenience”. Moreover, results are mostly consistent with findings in current MS-related research. For example, the OA results imply that pain, sensory issues, fatigue, and mental health issues (mainly consisting of the keywords “depression” and “anxiety”) represent a particularly high UMN. Green et al. [20] concluded by assessing UMNs due to a MS-specific questionnaire, that pain, fatigue and walking impairment contribute the most to MS outpatient patient perceived health, followed by depression and anxiety. The similarity of the results demonstrates this methods potential to identify key UMNs and to achieve a reasonable ranking of symptoms and disease burden comparable to analogue methods. Differences in prioritization, such as the low ranking of walking disability compared to Green et al. could be explained by the choice of model parameters used and the selection of features considered. For example, a topic was detected that consisted of posts dominated by words such as "drive," "car," and "walk" - suggesting walking disability also causes severe limitations in automobility. Given our chosen UMN definition, we excluded this topic from our analysis. However, merging the "walking impairment" topic with the “automobility” topic into a general "mobility topic" would result in a higher prioritization and therefore be even more similar to the results of Green et al. [20] One way to accomplish this would have been to reduce the number of topics, resulting in fewer UMNs extracted overall. Hence, a key challenge is the definition of the model parameters. In this regard, it seems beneficial to involve medical domain experts for ensuring that the results are meaningful. Further identified symptoms such as word-findings problems [44], cognitive impairment [45], bladder infection [46] and vision problems [47] are also consisting with extant research. In the dimension “side effects” flush and gastrointestinal issues as well as cancer could be identified, which are already observed side effects [22, 46], e.g. in treatments with dimethyl fumarate [48]. In accordance with the existing literature [22, 23], the route of administration by injection, has been indicated to be inconvenient as satisfaction score is rather low (table 2). Given the method used, information on immediate quality of life, mortality, or time of suffering in disease was not available.

Furthermore, many topics were found lying outside the conventional definition of UMNs. For example, patients discussed topics such as insurance coverage, employment, or hospitalization. Whereas these topics were not included in our analysis as they were beyond the scope of our research question, it indicates this method could be used to gain a more comprehensive understanding of patients' needs. This study highlights the potential of social media mining for analyzing patients' unmet medical needs. Given the combination of social media data and state-of-the-art analytics, as well as the need for increasing consideration of the patient perspective in drug development, we expect to witness an increasing use of similar approaches. Information could be used to align drug development with patient needs or to define endpoints in clinical trials that are relevant to patients. In addition, the proposed method could facilitate real-time tracking of patient needs as well as reduce costs associated with analog patient studies.

5.1. Limitations

In general, veracity of social media data for health-related analysis is limited. Since contributors such as patients are often lay persons using own definitions and terminologies the results can be biased. Comprehensive information on patient characteristics such as disease severity, concomitant diseases, site of care or medication use is not available. Moreover, data does not reveal a patient location. Therefore, we cannot conclude whether the identified UMN exists because no appropriate form of therapy is available or because the health care system does not provide access. Furthermore, MS includes different stages which is not considered in our analysis. Since results are based on social media posts from a publicly available online forum, results can be biased due to people participating in discussion but not suffering from MS, e.g., relatives of patients who searching for advice. The dataset consists of posts published during the years 2013 to 2020. Analysis was conducted using the data aggregated in its entirety. Therefore, it is uncertain whether the topic ranking reflects the current needs of patients. For instance, topic ranking could be biased by needs relevant to patients in the past, but already addressed due to pharmaceutical innovations and treatments available in the present. To enhance the distinction between past and current UMN, the opportunity algorithm might be applied using a smaller time interval. As all results were generated from the overall sample, robustness of the proposed algorithm needs to be proven. Regarding the topic model pipeline, spelling correction and the manual curation of the dictionary could improve results. The labeling of topics depends on subjective judgement. As topic models discard word order, location information is not available and causal inference is not possible, which is a significant limitation in drug development. Furthermore, a document-level sentiment analysis was used to calculate satisfaction regarding specific UMN topics. While the proposed method assigns a higher weight to relevant sentiment scores based on the document's contribution to the topic, a more detailed sentiment analysis could improve the results accuracy. For example, the more different topics are included in a document, the less accurate the document-level sentiment analysis is. In future research, we plan to apply keyword-level sentiment analysis to more accurately capture UMNs satisfaction. Since ground truth data on patient preferences is not available, limiting evaluation of the results, future studies should further assess the validity of the chosen method. For instance, the results of the proposed algorithm could be evaluated by comparing it to conventional and established methods of deriving patient needs, such as quality-of-life surveys or conjoint analysis.

5.2. Conclusion

Our study aimed to explore unmet medical needs in MS patient's posts within an online disease forum employing social media mining methods. With the use of the opportunity algorithm, we were able to identify unmet medical needs, relevant from patients' everyday perspective. Results of further automated prioritization are almost consistent with known needs of MS outpatients from extant research. The concrete results provide initial entry points for pharmaceutical innovation for MS patients in the areas of the most significant UMNs: sensory problems and pain, sleep problems and chronic fatigue, and mental health, each with an opportunity score above 5.

This feasibility study demonstrates the potential of social media mining to generate relevant insights into hard to reach, rare disease populations to foster patient-centered drug development. The observed decline in the number of posts over the years indicates fewer users are participating in the online forum analyzed. Given the rise of social media platforms and MS online forums, we suggest patients are spreading across different online platforms, resulting in a decrease of active users and posts. To gain more comprehensive and reliable insights, data from multiple social media platforms should be combined by data fusion methods.

6. References

[1] S. C. Rivera, D. G. Kyte, O. L. Aiyegbusi, A. L. Slade, C. McMullan, and M. J. Calvert, "The impact of patient-reported outcome (PRO) data from clinical trials: a systematic review and critical analysis," (in eng), Health Qual Life Outcomes, vol. 17, no. 1, p. 156, Oct 16 2019, doi: 10.1186/s12955-019-1220-z.

[2] E. A. Holmes et al., "Patient-Focused Drug Development Methods for Benefit–Risk Assessments: A Case Study Using a Discrete Choice Experiment for Antiepileptic Drugs," Clinical Pharmacology & Therapeutics, vol. 105, no. 3, pp. 672-683, 2019.

[3] M. Vennemann, V. Ruland, J. P. Kruse, C. Harloff, H. Trubel, and H. Gielen-Haertwig, "Future unmet medical need as a guiding principle for pharmaceutical R&D," Drug Discov Today, vol. 24, no. 9, pp. 1924-1929, Sep 2019, doi: 10.1016/j.drudis.2019.06.004.

[4] L. S. Crawford, G. J. Matczak, E. M. Moore, R. A. Haydar, and P. T. Coderre, "Patient-centered drug development and the Learning Health System," Learning Health Systems, vol. 1, no. 3, p. e10027, 2017, doi: https://doi.org/10.1002/lrh2.10027.

[5] M. C. Ysrraelit, M. P. Fiol, M. I. Gaitán, and J. Correale, "Quality of life assessment in multiple sclerosis: different perception between patients and neurologists," Frontiers in neurology, vol. 8, p. 729, 2018.

[6] H. Wen et al., "Comparison of expectations of physicians and patients with rheumatoid arthritis for rheumatology clinic visits: a pilot, multicenter, international study," (in eng), Int J Rheum Dis, vol. 15, no. 4, pp. 380-9, Aug 2012, doi: 10.1111/j.1756-185X.2012.01752.x.

[7] F. Lucas, "Improving market access to rare disease therapies: A worldwide perspective with recommendations to the industry," Medicine Access@ Point of Care, vol. 2, p. 2399202618810121, 2018.

[8] D. A. Hughes and J. Poletti-Hughes, "Profitability and market value of orphan drug companies: a retrospective, propensity-matched case-control study," PLOS one, vol. 11, no. 10, p. e0164681, 2016.

[9] J. Koss, A. Rheinlaender, H. Truebel, and S. Bohnet-Joschko, "Social media mining in drug development—fundamentals and use cases," Drug Discovery Today, 2021.

[10] M. D. Tapi Nzali, S. Bringay, C. Lavergne, C. Mollevi, and T. Opitz, "What Patients Can Tell Us: Topic Analysis for Social Media on Breast Cancer," JMIR Med Inform, vol. 5, no. 3, p. e23, Jul 31 2017, doi: 10.2196/medinform.7779.

[11] F. Schafer et al., "Mapping and Modeling of Discussions Related to Gastrointestinal Discomfort in French-Speaking Online Forums: Results of a 15-Year Retrospective Infodemiology Study," J Med Internet Res, vol. 22, no. 11, p. e17247, Nov 3 2020, doi: 10.2196/17247.

[12] Y. Zhao, J. Zhang, and M. Wu, "Finding Users' Voice on Social Media: An Investigation of Online Support Groups for Autism-Affected Users on Facebook," Int J Environ Res Public Health, vol. 16, no. 23, Nov 29 2019, doi: 10.3390/ijerph16234804.

[13] D. C. Stokes, A. Andy, S. C. Guntuku, L. H. Ungar, and R. M. Merchant, "Public Priorities and Concerns Regarding COVID-19 in an Online Discussion Forum: Longitudinal Topic Modeling," J Gen Intern Med, vol. 35, no. 7, pp. 2244-2247, Jul 2020, doi: 10.1007/s11606-020-05889-w.

[14] B. Carron-Arthur, J. Reynolds, K. Bennett, A. Bennett, and K. M. Griffiths, "What's all the talk about? Topic modelling in a mental health Internet support group," BMC Psychiatry, vol. 16, no. 1, p. 367, Oct 28 2016, doi: 10.1186/s12888-016-1073-5.

[15] T. Jiang, V. Osadchiy, J. N. Mills, and S. V. Eleswarapu, "Is It All in My Head? Self-reported Psychogenic Erectile Dysfunction and Depression Are Common Among Young Men Seeking Advice on Social Media," Urology, vol. 142, pp. 133-140, Aug 2020, doi: 10.1016/j.urology.2020.04.100.

[16] N. Cook et al., "Evaluating Patient Experiences in Dry Eye Disease Through Social Media Listening Research," Ophthalmol Ther, vol. 8, no. 3, pp. 407-420, Sep 2019, doi: 10.1007/s40123-019-0188-4.

[17] N. S. Cook, J. Cave, and A. P. Holtorf, "Patient Preference Studies During Early Drug Development: Aligning Stakeholders to Ensure Development Plans Meet Patient Needs," Front Med (Lausanne), vol. 6, p. 82, 2019, doi: 10.3389/fmed.2019.00082.

[18] M. J. Paul et al., "Social media mining for public health monitoring and surveillance," in Biocomputing 2016: Proceedings of the Pacific symposium, 2016: World Scientific, pp. 468-479.

[19] A. Scalfari, V. Knappertz, G. Cutter, D. S. Goodin, R. Ashton, and G. C. Ebers, "Mortality in patients with multiple sclerosis," Neurology, vol. 81, no. 2, pp. 184-192, 2013.

[20] R. Green, G. Cutter, M. Friendly, and I. Kister, "Which symptoms contribute the most to patients' perception of health in multiple sclerosis?," Multiple Sclerosis Journal–Experimental, Translational and Clinical, vol. 3, no. 3, p. 2055217317728301, 2017.

[21] P. S. Rommer and U. K. Zettl, "Managing the side effects of multiple sclerosis therapy: pharmacotherapy options for patients," Expert opinion on pharmacotherapy, vol. 19, no. 5, pp. 483-498, 2018.

[22] G. Buard et al., "Switching for convenience from first-line injectable treatments to oral treatments in multiple sclerosis: data from a retrospective cohort study," Multiple sclerosis and related disorders, vol. 33, pp. 39-43, 2019.

[23] P. de Flon, K. Laurell, L. Söderström, M. Gunnarsson, and A. Svenningsson, "Improved treatment satisfaction after switching therapy to rituximab in relapsing–remitting MS," Multiple Sclerosis Journal, vol. 23, no. 9, pp. 1249-1257, 2017.

[24] I. Gil-González, A. Martín-Rodríguez, R. Conrad, and M. Á. Pérez-San-Gregorio, "Quality of life in adults with multiple sclerosis: a systematic review," BMJ open, vol. 10, no. 11, p. e041249, 2020.

[25] C. M. Christensen, S. D. Anthony, G. Berstell, and D. Nitterhouse, "Finding the right job for your product,"

MIT Sloan Management Review, vol. 48, no. 3, p. 38, 2007.
[26] A. W. Ulwick, "What Is Outcome-Driven Innovation®(ODI)?," White Paper, 2009.
[27] B. Jeong, J. Yoon, and J.-M. Lee, "Social media mining for product planning: A product opportunity mining approach based on topic modeling and sentiment analysis," International Journal of Information Management, vol. 48, pp. 280-290, 2019, doi: 10.1016/j.ijinfomgt.2017.09.009.
[28] J. Choi, S. Oh, J. Yoon, J.-M. Lee, and B.-Y. Coh, "Identification of time-evolving product opportunities via social media mining," Technological Forecasting and Social Change, vol. 156, 2020, doi: 10.1016/j.techfore.2020.120045.
[29] N. Srnicek, "The challenges of platform capitalism: Understanding the logic of a new business model," Juncture, vol. 23, no. 4, pp. 254-257, 2017, doi: https://doi.org/10.1111/newe.12023.
[30] M. Elsayed, A. Abdelwahab, and H. Ahdelkader, "A Proposed Framework for Improving Analysis of Big Unstructured Data in Social Media," in 2019 14th International Conference on Computer Engineering and Systems (ICCES), 17-17 Dec. 2019 2019, pp. 61-65, doi: 10.1109/ICCES48960.2019.9068154.
[31] F. Martin and M. Johnson, "More efficient topic modelling through a noun only approach," in Proceedings of the Australasian Language Technology Association Workshop 2015, 2015, pp. 111-115.
[32] C. May, R. Cotterell, and B. Van Durme, "An Analysis of Lemmatization on Topic Models of Morphologically Rich Language," arXiv preprint arXiv:1608.03995, 2016.
[33] R. Campos, V. Mangaravite, A. Pasquali, A. Jorge, C. Nunes, and A. Jatowt, "YAKE! Keyword extraction from single documents using multiple local features," Information Sciences, vol. 509, pp. 257-289, 2020.
[34] J. H. Lau, T. Baldwin, and D. Newman, "On collocations and topic models," ACM Trans. Speech Lang. Process., vol. 10, no. 3, p. Article 10, 2013, doi: 10.1145/2483969.2483972.
[35] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," the Journal of machine Learning research, vol. 3, pp. 993-1022, 2003.
[36] A. Wahbeh, T. Nasralah, O. El-Gayar, M. Al-Ramahi, and A. Elnoshokaty, "Adverse Health Effects of Kratom: An Analysis of Social Media Data," in Proceedings of the 54th Hawaii International Conference on System Sciences, p. 3934.
[37] M. A. Al-Ramahi, J. Liu, and O. F. El-Gayar, "Discovering design principles for health behavioral change support systems: A text mining approach," ACM Transactions on Management Information Systems (TMIS), vol. 8, no. 2-3, pp. 1-24, 2017.
[38] H. Jelodar et al., "Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey," Multimedia Tools and Applications, vol. 78, no. 11, pp. 15169-15211, 2019.
[39] M. Röder, A. Both, and A. Hinneburg, "Exploring the Space of Topic Coherence Measures," presented at the Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, 2015.
[40] S. Syed and M. Spruit, "Full-text or abstract? examining topic coherence scores using latent dirichlet allocation," in 2017 IEEE International conference on data science and advanced analytics (DSAA), 2017: IEEE, pp. 165-174.
[41] A. Carvalho and L. Harris, "Off-the-Shelf Technologies for Sentiment Analysis of Social Media Data: Two Empirical Studies," 2020.
[42] N. K. Singh, D. S. Tomar, and A. K. Sangaiah, "Sentiment analysis: a review and comparative analysis over social media," Journal of Ambient Intelligence and Humanized Computing, vol. 11, no. 1, pp. 97-117, 2020.
[43] A. X.-L. Nguyen, X.-V. Trinh, S. Y. Wang, and A. Y. Wu, "Determination of Patient Sentiment and Emotion in Ophthalmology: Infoveillance Tutorial on Web-Based Health Forum Discussions," Journal of Medical Internet Research, vol. 23, no. 5, p. e20803, 2021.
[44] R. Brandstadter et al., "Word-finding difficulty is a prevalent disease-related deficit in early multiple sclerosis," Multiple Sclerosis Journal, vol. 26, no. 13, pp. 1752-1764, 2020.
[45] M. Á. Macías Islas and E. Ciampi, "Assessment and impact of cognitive impairment in multiple sclerosis: an overview," Biomedicines, vol. 7, no. 1, p. 22, 2019.
[46] W. L. G. de Medeiros Junior et al., "Urinary tract infection in patients with multiple sclerosis: An overview," Multiple Sclerosis and Related Disorders, vol. 46, p. 102462, 2020.
[47] B. Sanchez-Dalmau et al., "Predictors of vision impairment in multiple sclerosis," PloS one, vol. 13, no. 4, p. e0195856, 2018.
[48] G. Liang, J. Chai, and H. S. Ng, "Safety of dimethyl fumarate for multiple sclerosis: A systematic review and meta-analysis," Multiple Sclerosis and Related Disorders, p. 102566, 2020.