Natural Language Processing for Large-Scale Analysis of Eczema and Psoriasis Social Media Comments

Jack A. Cummins[1], Guohai Zhou[2] and Vinod E. Nambudiri[3]

Social media tools are widely used by dermatologic patients. Eczema and psoriasis, two of the most common inflammatory skin diseases, are well-represented on the social media site Reddit. We used natural language processing tools to examine comments in subreddits r/psoriasis and r/eczema (combined user base >187,000), tracking commenters' interest levels and sentiments related to common treatments for psoriasis and eczema as well as discussions of adverse drug reactions. All comments from 2014–2020 from the subreddits r/eczema (n = 196,571) and r/psoriasis (n = 123,144) were retrieved and processed using natural language processing tools. Comment volume in r/eczema related to antibacterial therapies, lifestyle changes, and prednisone decreased from 2014–2020, whereas phototherapy comments remained stable, and dupilumab comment volume increased. Comment volume in r/psoriasis for newer therapeutics (including biologics and apremilast) increased after Food and Drug Administration approval, whereas older therapies such as etanercept, adalimumab, and methotrexate decreased over time. Sentiment scores tended to decrease in the years after Food and Drug Administration approval. Among psoriasis treatments, calcipotriene and branded calcipotriene/betamethasone foam had the highest sentiment, whereas apremilast had the lowest overall sentiment score. These analyses also identified changes in patient interest levels and sentiment related to eczema and psoriasis treatments, suggesting an area for additional research.

JID Innovations (2023);**3:**100210 doi:10.1016/j.xjidi.2023.100210

INTRODUCTION

Psoriasis and eczema are common inflammatory skin diseases that can substantially impact patients' QOL (Falissard et al., 2020; Moberg et al., 2009). A wide range of topical, systemic, and nonpharmacologic therapeutic interventions are available for each of these conditions, but treatment regimens for these chronic conditions can be time consuming, difficult, or unpleasant. Studies show that noncompliance is common among patients with these conditions (Murage et al., 2018; Patel et al., 2017; Patel and Feldman, 2017). Patient counseling and frequency of visits correlate with compliance and improvement (Heaton et al., 2013), but patients frequently consult information sources outside the clinic, such as Google searches, social media, and health forums (Sunkureddi et al., 2018; Wu et al., 2020).

Data that inform clinicians about patients' perceptions of skin disease and treatments can identify patients' knowledge gaps and contribute to more effective disease management (Sunkureddi et al., 2018). Natural language processing (NLP) techniques have been used to study large quantities of unstructured patient survey response data, including the assessment of adult patients' perceptions of their atopic dermatitis (Falissard et al., 2020). Although patient perceptions have historically been researched primarily through survey studies, similar data can now be retrieved for thousands of patients through large social media communities such as Reddit and a variety of patient forums.

NLP methods can be used to efficiently and effectively process large quantities of social media data to illuminate key insights into patients' perceptions of their skin conditions and disease management. As identified by Buntinx-Krieg et al. (2017), Reddit is a viable source for patient reports and comments on their skin disease. Reddit is the sixth most popular website in the United States and is divided up into forums called subreddits. Several of these subreddits are of dermatologic interest, such as the r/eczema and r/psoriasis subreddits examined in this study. A variety of methods are available to effectively categorize and study data in the Reddit medium (Couto, 2019; Okon et al., 2020).

This study uses validated NLP tools to better understand patients' interest in psoriasis and eczema treatments, with a focus on comment volume and sentiment analysis over time, as well as patient-reported adverse drug reactions discussed in such forums. This study shows the use of NLP methodology applied to skin health topics on Reddit to better understand the level of patient community interest in relevant topics.

[1] *Manchester Essex Regional High School, Manchester, MA;* [2] *Department of Dermatology, Brigham and Women's Hospital and Harvard Medical School, Boston, MA; and* [3] *Department of Dermatology, Brigham and Women's Hospital and Harvard Medical School, Boston, MA*

Correspondence: Vinod Nambudiri, Department of Dermatology, Brigham and Women's Hospital, 221 Longwood Avenue, Boston, Massachusetts 02115, USA. E-mail: vinod.nambudiri@gmail.com

Abbreviations: ADR, adverse drug reaction; BOW, bag of words; FDA, Food and Drug Administration; NLP, natural language processing; RE, regular expression

Received 29 October 2021; revised 19 February 2023; accepted 27 February 2023; accepted manuscript published online XXX; corrected proof published online XXX

Cite this article as: JID Innovations 2023;3:100210

RESULTS

Number of comments per unique commentor

After filtering, there were 123,144 comments from the r/psoriasis subreddit and 196,571 comments from the r/eczema subreddit. There were 24,759 unique commenters in the eczema subreddit and 14,015 unique commenters in the psoriasis subreddit. In the eczema subreddit, the number of unique commenters steadily rose from 1,091 commenters in 2014 to 11,389 commenters in 2020, and in the psoriasis subreddit, the number of unique commenters also steadily rose from 1,039 commenters in 2014 to 5,939 commenters in 2020. The total number of comments, unique commenters, and comments per commenter are shown by year in Tables 1 and 2.

Proportion of comments containing keywords in subreddit r/eczema

The total comment volume for subreddits r/eczema and r/psoriasis increased each year, with an initial comment volume of 4,969 in r/eczema and 5,120 in r/psoriasis in 2014, increasing to a total volume of 70,516 in r/eczema and 33,929 in r/psoriasis in 2020. Comments containing regular expression (RE) terms related to eczema treatment were examined as a percentage of the total comment volume (Figure 1). Terms related to bacteria (staph and infection) and antibacterial treatments (cider or vinegar, bleach, and antibiotic) trend toward a lower percentage of comments over time from 2014 to 2020. Dupilumab comments increased from 2014 to 2020, whereas the comment percentage for prednisone and cyclosporine trended downward over the same period (Figure 1b). Interest in dupilumab peaked in 2018 when >4% of comments in the r/eczema subreddit contained the term dupilumab or DUPIXENT. The phototherapy comment percentage remained proportionally stable. The dashed lines show that dupilumab was an active topic for comment before Food and Drug Administration (FDA) approval in 2016. In 2017 (the year of FDA approval), the brand name DUPIXENT was introduced, and the percentage of comments containing the term DUPIXENT rose to over 3% in 2017 and over 4% in 2018. The use of the term dupilumab decreased as its brand name, and DUPIXENT became the more common term used.

Although notable but not shown, the percentage of comments containing the term steroid decreased from 12.8% in 2014 to 7.9% in 2020. The percentage of comments containing the term moisturizer or an alternative spelling of moisturiser also decreased from 5.1% of comments in 2014 to 3.5% of comments in 2020. Probiotics saw an increase in the percentage of comments in 2015 and 2016, but with this exception, there was a consistent trend toward a reduced percentage of comments for topics related to diet and stress, including reduced frequencies of the use of the terms stress, gluten, milk, and diet (Figure 1c).

Table 1. Number of Comments, Number of Unique Commenters, and Number of Comments Per Unique Commenter Per Year in r/eczema

Year	Number of Comments	Number of Unique Commenters	Number of Comments Per Unique Commenter
2014	4,969	972	5.11
2015	8,894	1,599	5.56
2016	7,858	1,825	4.31
2017	20,807	3,555	5.85
2018	32,725	5,088	6.43
2019	50,802	7,937	6.40
2020	70,516	11,387	6.19

Table 2. Number of Comments, Number of Unique Commenters, and Number of Comments Per Unique Commenter Per Year in r/psoriasis

Year	Number of Comments	Number of Unique Commenters	Number of Comments Per Unique Commenter
2014	5,120	1,038	4.93
2015	7,012	1,418	4.94
2016	8,378	1,569	5.34
2017	11,000	2,149	5.17
2018	20,707	3,190	6.49
2019	32,852	4,631	7.09
2020	33,929	5,916	5.74

Volume of comments containing keywords in subreddit r/psoriasis

Comment percentages for biologic therapeutics in r/psoriasis were determined by including the brand name and equivalent generic name to calculate the total comment percentage. Medications that have been FDA approved since 2014, such as ixekizumab, guselkumab, and secukinumab, had proportionately increased volumes, coinciding with FDA approval (Figure 2a). Drugs that were FDA approved before 2014—etanercept and adalimumab—showed a decrease in the volume of comments over time. Despite the decreased volume over time, adalimumab continued to have a higher comment percentage than any other biologic treatment for every year examined, including 2020. Ustekinumab comment volume remained stable through 2018, with a decrease in 2019 and 2020. The percentage of comments containing the name of at least one of the biologics dropped in 2019 and 2020. The percentage of comments containing the terms calcipotriol and Enstilar has increased over time, whereas the percentage of comments containing the term tacrolimus remained stable over the study period (Figure 2b). The percentage of comments containing the term steroid rose from 4.1% in 2014 to 5.3% in 2020. Apremilast comment volume increased to a peak >2% of comments in 2016 but subsequently decreased (Figure 2c). Methotrexate comment volume decreased substantially from 4.1% of comments in 2016 to 2.3% in 2017 and remained stable in 2018 and 2019, with a slight drop in 2020.

To determine whether the changing volume of Reddit comments regarding a particular medication might be a simple reflection of increased medication usage in the population, data were reviewed related to the volume of prescription sales for two eczema medications (prednisone and cyclosporine) and for two psoriasis medications (adalimumab and etanercept). Data for the total number of prescriptions of prednisone, cyclosporine, adalimumab, and etanercept from

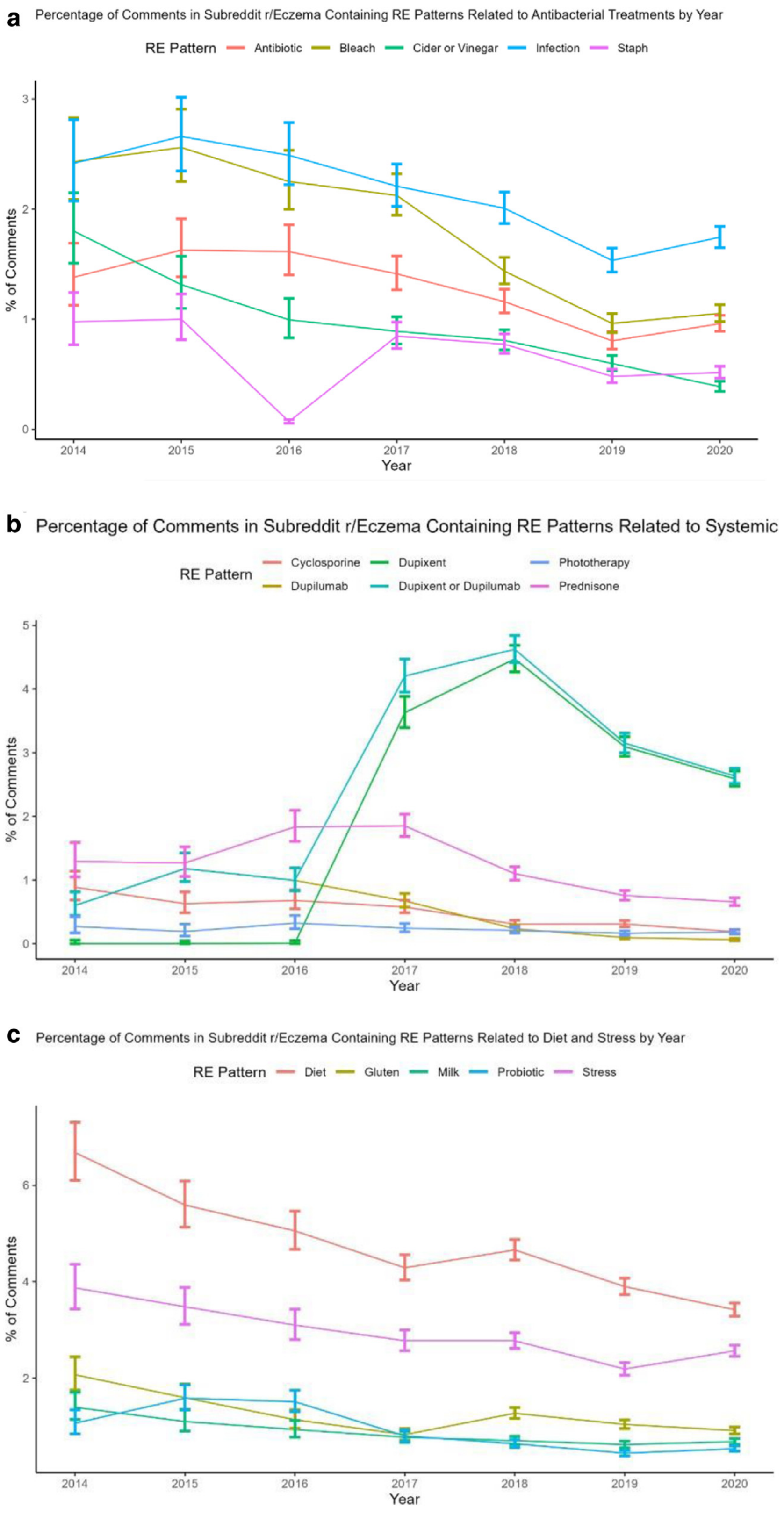


Figure 1. Percentage of comments (95% confidence interval as the vertical bar) in subreddit r/eczema containing RE patterns specified in legend by year. (**a**) Percentage of comments containing RE patterns related to antibacterial treatments (0.21% decrease over time on average, $P < 0.001$). (**b**) Percentage of comments containing RE patterns related to systemic treatments (0.49% increase over time on average, $P = 0.109$). (**c**) Percentage of comments containing RE patterns related to diet and stress (0.24% decrease over time on average, $P = 0.003$). RE, regular expression.

2014−2019 were obtained from clincalc.com. A review of the prescription sales data did not show a correlation between prescription sales and Reddit conversation volume for any of the four medications examined. Notably, these medications are used to treat multiple medical conditions, and therefore, it is not possible to determine total sales for the indications of eczema or psoriasis treatment solely on the basis of the overall sales of these medications. Moreover, many factors other than the prevalence of usage may drive Reddit conversation, including commentors' interest in

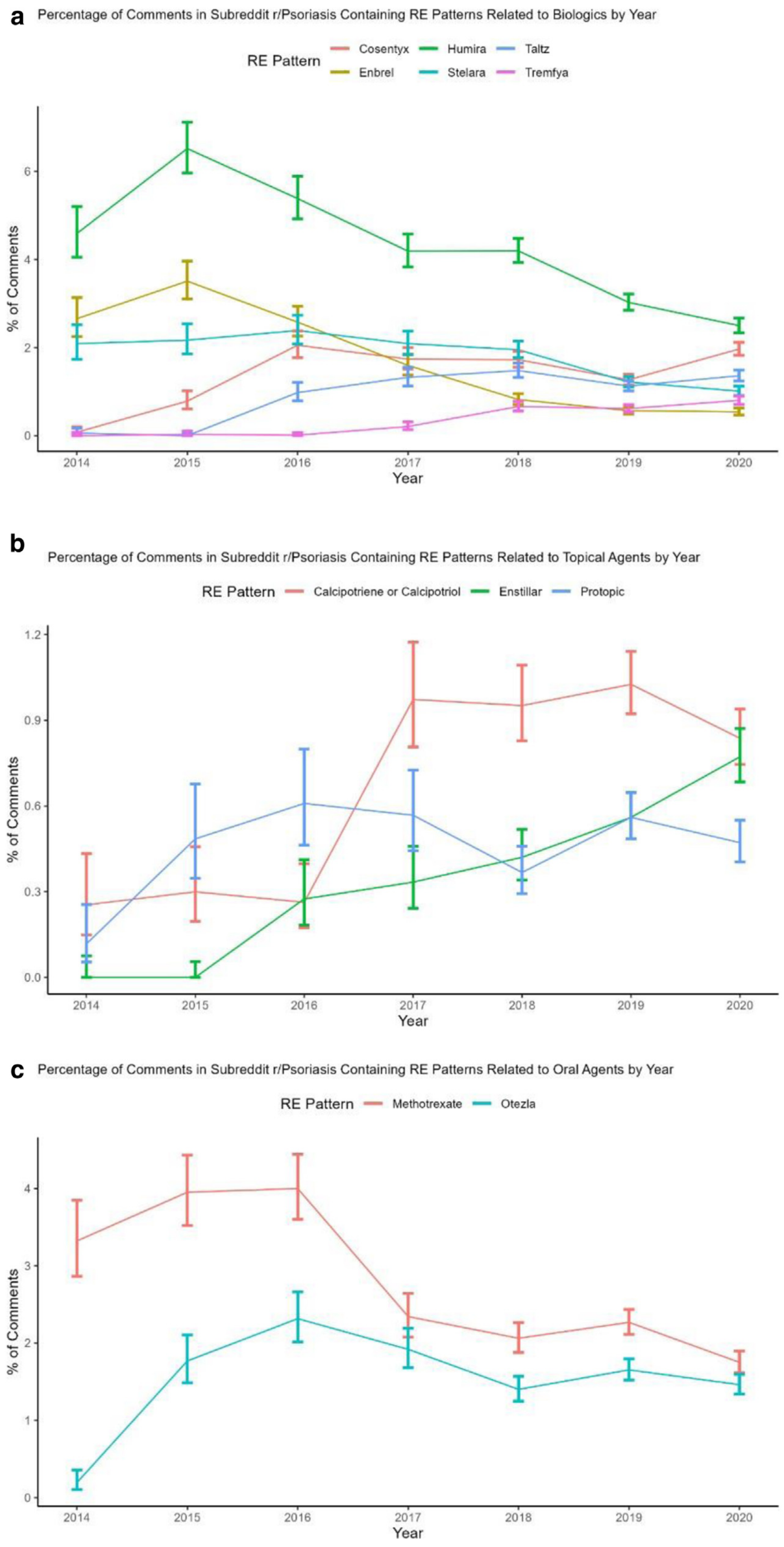
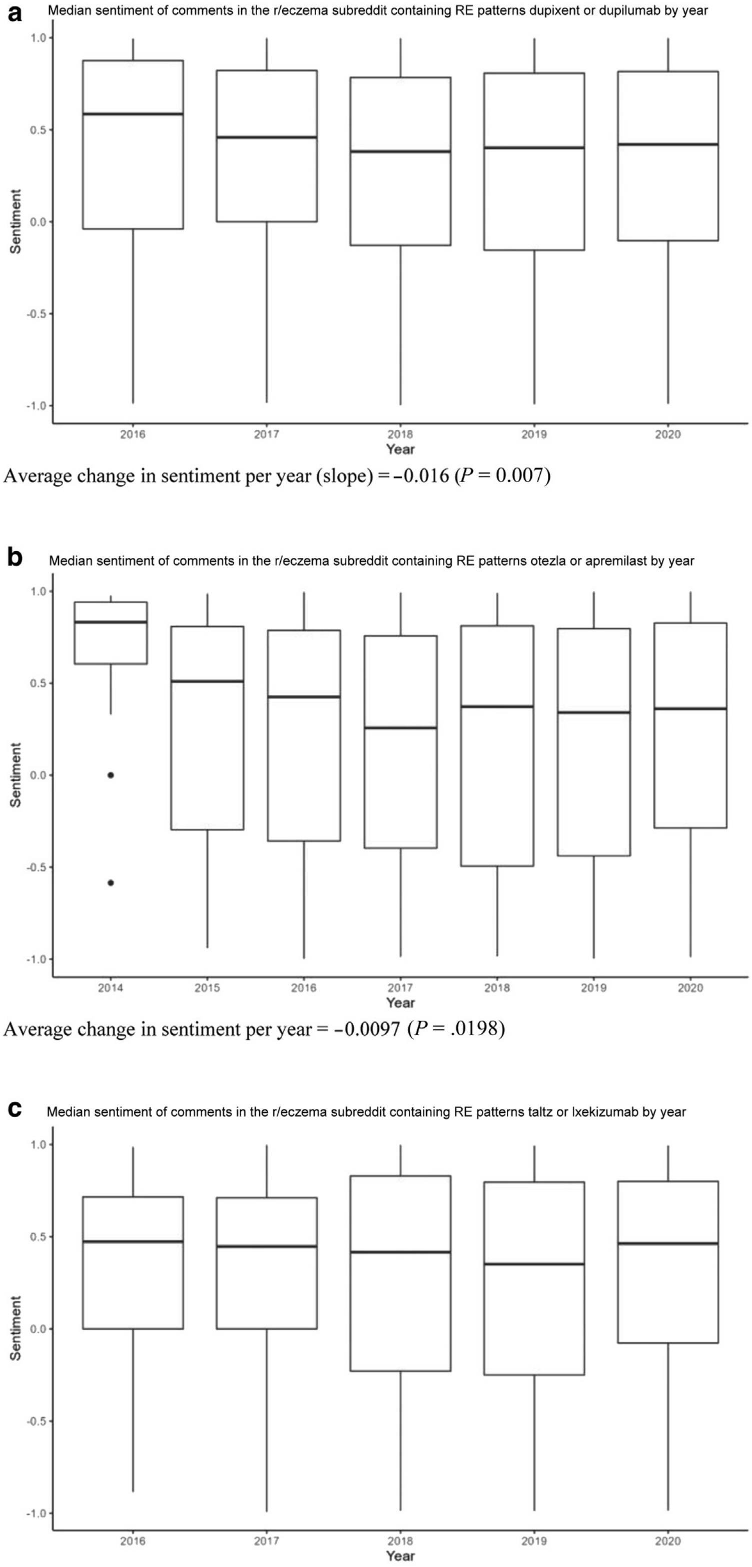


Figure 2. Percentage of comments (95% confidence interval as the vertical bar) in subreddit r/psoriasis containing RE patterns specified in legend by year. (**a**) Percentage of comments containing RE patterns related to biologics. Shaded rectangles on the line plot indicate the year of FDA approval. (**b**) Percentage of comments containing RE patterns related to topical agents. (**c**) Percentage of comments containing RE patterns related to oral agents. FDA, Food and Drug Administration; RE, regular expression.

newly available medications, newly identified side effects, or new indications for medication usage.

To determine whether comments were accurately sorted into corresponding subreddits, the most common treatments for eczema and psoriasis were examined in both r/psoriasis and r/eczema. The search terms DUPIXENT and dupilumab, which refer to a popular eczema treatment, were discussed in the eczema subreddit in 6,346 comments but were only



Average change in sentiment per year (slope) = –0.016 ($P = 0.007$)

Average change in sentiment per year = –0.0097 ($P = .0198$)

Average change in sentiment per year = 0.003406 ($P = 0.748$)

Figure 3. Median sentiment on a scale from −1 to 1 of comments in subreddit r/eczema or r/psoriasis containing specific RE patterns by year. The horizontal line within the box represents the median. The upper and lower horizontal line limits of the boxes represent the first and third quartiles. The vertical lines represent the maximum and minimum values (excluding outliers), and points outside of the vertical lines represent outliers. (**a**) The median sentiment of comments in r/eczema subreddit containing RE pattern DUPIXENT or dupilumab. (**b**) The median sentiment of comments in r/psoriasis subreddit containing RE pattern Otezla or apremilast. (**c**) The median sentiment of comments in r/psoriasis subreddit containing RE pattern Taltz or ixekizumab. RE, regular expression.

mentioned in 18 comments in the psoriasis subreddit. Likewise, the term Humira or adalimumab, which refers to a popular psoriasis treatment, was seen in 8,201 comments in the psoriasis subreddit, but only 41 comments contained these terms in the eczema subreddit. In addition, a manual review of 30 randomly selected comments from the psoriasis subreddit and 30 randomly selected comments from the eczema subreddit was conducted to determine whether most of the comments in each subreddit originated from patients with skin diseases reporting on personal experiences. From

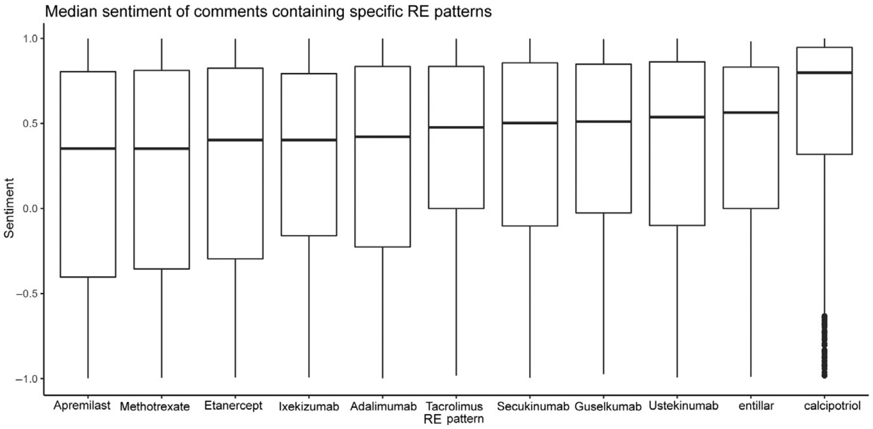


Figure 4. The median sentiment of all comments containing specific RE terms in the r/psoriasis subreddit. Both the generic and brand names were used when searching for comments. The horizontal line within the box represents the median. The upper and lower horizontal line limits of the boxes represent the first and third quartiles. The vertical lines represent the maximum and minimum values (excluding outliers), and points outside of the vertical lines represent outliers. $P < 0.005$ for all compared with apremilast (except methotrexate). RE, regular expression.

the eczema subreddit, 24 of 30 (80%, 95% confidence interval = 63−91%) comments were determined to be comments from patients describing their personal experiences. In the psoriasis subreddit, 22 of 30 (73%, 95% confidence interval = 56−86%) comments were determined to be from patients describing their personal experiences.

Sentiment analysis

The median sentiment of the comments containing the term DUPIXENT or dupilumab decreased from 0.59 in 2016, 1 year before it was FDA approved, to 0.42 in 2020 (Figure 3a). The median sentiment of comments containing the term Otezla or apremilast decreased from 0.82 in 2014 when it was FDA approved to 0.34 in 2020 (Figure 3b). The median sentiment of comments from the r/psoriasis subreddit containing the term Taltz or ixekizumab decreased gradually from 0.48 in 2016 when it was FDA approved to 0.35 in 2019 but jumped back up to 0.46 in 2020 (Figure 3c).

Comments containing the term calcipotriol or calcipotriene had the highest median sentiment score of 0.80, and comments containing apremilast or Otezla had the lowest sentiment score of 0.35 (Figure 4). The number of comments analyzed for the sentiment of each treatment ranged from 1,021 comments for guselkumab to 8,201 comments for adalimumab, with the exception of Enstillar, which had only 161 comments analyzed.

Comments related to known adverse drug reactions (ADRs) to treatments were identified by performing an RE search for comments that contained both the drug name and the known adverse reaction to the drug. The side effects examined (conjunctivitis with dupilumab use and nausea or diarrhea with apremilast use) are known side effects that are listed on the medication package insert. After the RE processing was performed, 10 comments were randomly selected for manual review for each ADR (Table 3). For each ADR, at least 4 of the 10 manually reviewed comments were self-reported. A total of 167 comments were found in the r/psoriasis subreddit containing either the term Otezla or the term apremilast and also a term related to nausea in the same comment. The RE pattern nause was queried because it is contained in multiple words related to nausea, including nausea and nauseous. In total, 73 comments were identified containing the term Otezla or apremilast and diarrhea. In the r/eczema subreddit, 82 comments were found to contain DUPIXENT or dupilumab and conjunctivitis.

Table 3. Potential ADRs detected by RE

Subreddit	Drug Name	Reaction	Number of Comments
r/psoriasis	Otezla or apremilast	Nause or vomit or throw up or threw up	167
r/psoriasis	Otezla or apremilast	Diarrhea	73
r/eczema	DUPIXENT or dupilumab	Conjunctivitis	82

Abbreviations: ADR, adverse drug reaction; RE, regular expression.
The table shows the number of comments containing any of the drug name (RE) patterns and any of the reaction RE patterns in a specific subreddit.

Nonprescription treatments

The number of comments containing RE patterns related to nonprescription treatments was analyzed by year (Figure 5). The percentage of comments containing the term aloe vera rose from 0.2% in 2014 to 0.32% in 2016, decreased to 0.24% in 2017, and then stayed stable at around 0.3% for the rest of the study period. The percentage of comments containing the term tea tree oil stayed relatively stable, with a maximum of 0.28% of comments containing the term in 2015 and a minimum of 0.15% containing it in 2019. The percentage of comments containing the term turmeric increased steadily from 0.02% in 2015 to 0.22% in 2017 and then decreased to 0.10% in 2020.

The average sentiment of comments containing words related to nonprescription treatments was analyzed. The median sentiment for apple cider vinegar or ACV was 0.47 in 766 comments in the r/eczema subreddit. The median sentiment for comments containing the term bleach bath was 0.40 in 1,876 comments. By comparison, the median sentiment for comments containing the term steroid was 0.36 in 2,730 comments, and the median sentiment for comments containing the term dupilumab/DUPIXENT was 0.42 in 5,941 comments.

The adverse effects of two commonly used nonprescription treatments, apple cider vinegar and calendula, were also analyzed in the r/eczema subreddit. Of 480 manually reviewed comments that contained the term apple cider

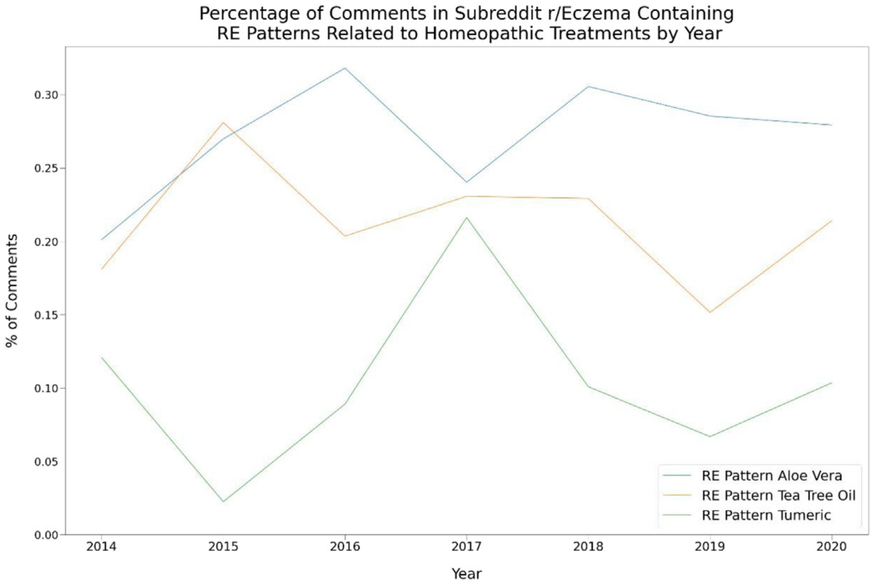


Figure 5. Percentage of comments in the r/eczema subreddit containing RE patterns specified in legend by year. RE, regular expression.

vinegar, 28 mentioned at least a mild adverse effect such as irritation, and 3 described an actual burn with lasting symptoms and pain. A total of 102 comments containing the term calendula were reviewed. Only two reports were identified describing mild stinging, and no other adverse events were observed.

Insight into the patient experience

The frequency of comments containing patterns related to the patient experience (itch, pain, sleep) was counted in both the eczema and psoriasis subreddits. Words such as painful and sleepy were also detected with a pattern search because it just looks for a pattern of letters, not a word. In the eczema subreddit, 8.39% of comments contained the pattern itch, compared with 3.82% of comments in the psoriasis subreddit. A total of 3.19% of comments in the eczema subreddit contained the pattern pain, compared with 4.27% of comments containing the pattern pain in the psoriasis subreddit. The pattern sleep was found in 2.72% of comments in the eczema subreddit, compared with 1.18% in the psoriasis subreddit.

DISCUSSION

This study provides insight into the interest level and sentiment related to treatment options for eczema and psoriasis as well as self-reported side effects. The results suggest that Reddit users who comment on r/eczema and r/psoriasis are knowledgeable about a range of treatments, including anticipated and newly FDA-approved medications. In many cases, increases and decreases in comment volume appear to parallel the degree to which these treatments are favored among dermatologists on the basis of available data and consensus.

Keyword searches related to biologic and oral medications in the r/psoriasis subreddit showed that when drugs are FDA approved, the percentage of comments containing the name of the drug tends to rise rapidly in the initial years and then plateau. By contrast, drugs that were FDA approved many years ago decrease in comment volume over time as newer medications become available, suggesting that the r/psoriasis forum participants may rapidly acquire knowledge about new drugs and lose interest in older drugs over time. Nonetheless, adalimumab, despite being an older biologic medication, had the highest comment volume of any biologic medication for all years examined, indicating sustained interest in widely used established medications.

In recent years, the r/eczema subreddit has had a decrease in the percentage of comments containing terms related to antibacterial treatments, lifestyle, and diet approaches. These changes may reflect a true decrease in interest in these topics or may reflect a relative reduction due to increased interest in newer therapeutic options that are of greater interest to Reddit users.

The comment volume pattern for dupilumab within the r/eczema subreddit was slightly different from those of the newly FDA-approved psoriasis medications, possibly because it was the first biologic agent approved for this indication. There was a moderate comment volume in the years before the drug was FDA approved, with 1% of comments in the subreddit containing the term dupilumab in 2015 and 2016. By contrast, the newer psoriasis drugs were discussed in very few comments before FDA approval and were almost exclusively referred to by brand name, likely a reflection of the fact that other effective medications for psoriasis were already available at the time that newer psoriasis drugs were FDA approved.

Comment volume within the subreddits appears to reflect changes in disease management and recommendations that would be expected over time in the dermatologic community and suggests an understanding of current recommended management among Reddit users. Sentiment analysis was also used to explore trends in patients' perceptions of treatments. The most consistent trend across medications was that in the year a drug was FDA approved, sentiment tends to be high, followed by a decrease in subsequent years. Before the use of the drug, discussion may be focused on the anticipation of the drug and perceived benefits over previous medications, which may diminish after actual usage, even in cases where the drug is an improvement over previous options. Likewise, medications that have been FDA approved for many years, such as etanercept and adalimumab, have decreased in sentiment over time, possibly because they are increasingly compared against newer and more effective drugs, which may hamper enthusiasm. Sentiment analysis showed the highest sentiment for topical medications, intermediate sentiment for injected medications, and the least positive sentiment for oral medications.

Personal reports of potential ADRs were found using RE to search for comments containing the name of a drug and a known adverse reaction. In each case, the search for the medication name and the relevant ADR revealed numerous comments related to the ADR, including self-reported ADRs. A search of isotretinoin-related comments revealed hundreds of comments mentioning mood. An apremilast ADR search revealed many comments related to nausea and diarrhea, and a search for comments related to dupilumab and conjunctivitis revealed 68 comments (although this is likely an underestimate given different lay terms for the condition).

This study shows that many patients self-report symptoms related to their medication use, and therefore NLP tools may be a valuable tool to search through thousands of forum comments for potential ADRs. In cases where an ADR is already known, surveillance of comments with ADR descriptions may help to better characterize the ADR. In addition, tools focused more generally on ADRs may be used to identify patterns within big data to identify possible rare ADRs that have not been previously identified during the initial clinical studies. Exploring such data for the identification of aggregated patient-reported outcome measures offers an additional powerful use case.

The findings of this study are limited to comments reported within the subreddits r/psoriasis and r/eczema and therefore may not be generalizable to individuals with all dermatologic conditions. Individuals commenting within these subreddits may be more invested in their disease management, possibly representing a population with greater disease impact than the average dermatology patient with these conditions. In addition, Reddit users must also have technical knowledge to participate in this social media forum. Although NLP allows for the processing of large quantities of unstructured data, the unstandardized nature of social media has inherent limitations, such as uncaptured data due to alternate spellings, word choices, or transition of certain topics to alternate forums. The bag of words (BOW) methodology approach helped to mitigate this limitation by identifying the most commonly used terms.

Another distinct limitation is the lack of transparent data on the sales or prescription volume for each drug for specific indications during the years of the study, limiting the ability to directly compare comment volumes with a metric such as the number of prescriptions for each individual drug.

Any study that relies primarily on NLP for data analysis will tend to lose some of the finer subtleties and meanings that may be identified by manual examination. The BOW approach used in this study captures individual words but does not provide the full meaning of the sentence or comment, thereby limiting the interpretation of the intended meaning of the comment. In the ADR portion of this study, a manual review of a random sample was completed to validate the data as representing true ADRs. Future studies may combine the use of NLP tools to identify comments that are likely to represent ADRs with a manual review of each potential ADR comment to obtain more detailed information.

This paper shows NLP tools successfully applied to eczema and psoriasis subreddits as a model that could be applied to other medical forums to better understand the patient experience with their disease, treatments, and ADRs. Examination of social media is uniquely valuable because it allows the physician and researcher to better understand the patient experience in the patient's own words without the restrictions of structured surveys or the bias of patients wishing to please a clinician or clinical researcher. Analysis of thousands of comments on social media forums can efficiently identify trends in sentiment and interest levels related to diseases and treatment interventions. Tools can also be used to extract highly valuable data, such as potential ADRs. Moreover, understanding patient use of language and trending topics may help clinicians to communicate better with patients, paving the way for enhanced clinical care in the future.

MATERIALS AND METHODS

Reddit is an open social media platform with over 430 million monthly active users. It is divided into a wide range of subreddit topics, including politics, sports, health, and others. All subreddit comments from r/eczema and r/psoriasis from 2013 to 2020 were obtained using the Pushshift API (Baumgartner et al., 2020). Each subreddit was analyzed individually using Python 3 (Van Rossum and Drake, 2009). All subreddit datasets were filtered by a process that included the removal of empty or whitespace comments (comments containing no characters other than whitespace); comments containing the string I am a bot (automated messages), deleted comments (comments removed by Reddit's moderators or the posting user and no longer visible to the public), and comments with only non-Unicode characters (such as some emojis and other images) were removed. All non-Unicode characters were removed from the comments.

All postfiltering comments from each subreddit were preprocessed using NLP techniques. All texts were made in lowercase, and punctuation was removed. One-letter words such as I and a as well as other isolated letters that were not part of a larger word were removed. Next, each comment was tokenized, converting each comment to a list of words. Stop words were removed from the lists of words using a union of Genism's and the Natural Language Toolkit's stop words lists, along with the words use and like (Bird et al., 2009; Řehůřek and Sojka, 2010). Stop words are common words that do not have much meaning and create noise in NLP data. In the subreddits r/eczema and r/psoriasis, the subreddit title words eczema and psoriasis, respectively, were treated as stop words on the basis of the assumption that almost all the comments in that subreddit would be related to that title word regardless of whether the word was written in the particular comment. After stop words were removed, the remaining words were lemmatized using Natural Language Toolkit's WordNet Lemmatizer (Bird et al., 2009). Lemmatization is a process that removes words' inflectional endings and reduces the words to their infinitive or dictionary form.

The processed comments were used to create a BOW model using a genism dictionary (Řehůřek and Sojka, 2010). A BOW model counts the number of occurrences of each word in a document. This model counts the words in a way that is easy to understand and implement, but it does not capture information related to the order of the words. For example, the comments "The president is not good and will fail" and "The president is good and will not fail" would both be represented as the same comment in a BOW model. After the BOW model was created, words that appeared fewer than 15 times or in more than 50% of the documents were removed because words with extremely high and low frequencies can create noise in NLP data. After comment preprocessing and filtering, the 1,000 words with the highest comment frequency (number of comments that contain the word at least once) were retrieved from the BOW model.

The list of 1,000 words for each subreddit was reviewed manually and reduced to fewer than 100 words of interest per subreddit on the basis of dermatologic relevance. Samples of comments from each subreddit were also manually reviewed to identify authorship (e.g., first person vs. family member). REs were used to identify and count comments that contained each of the selected words of interest within the corresponding subreddit. The RE searches were set as case insensitive. In addition to processing the RE analysis for the words individually, some words were processed as RE groups. Words that had the same meaning, such as generic and brand name descriptions of the same medication, were grouped together for RE analysis. For example, the words DUPIXENT and dupilumab were combined in an RE group such that the total number of comments containing DUPIXENT and/or dupilumab were counted. In addition, RE analysis was also completed for each of these words individually. Matplotlib was used to create all plots for each selected word within the subreddit, showing the percentage of comments containing the specified word each year during the specified time (Hunter, 2007). The percentage of comments was calculated by dividing the total number of comments within the subreddit containing the RE pattern by the total number of comments in the subreddit over the specified time.

Sentiment analysis with the Valence Aware Dictionary for sEntiment Reasoning was used to extract the sentiment of comments containing the name of a drug (Hutto and Gilbert, 2014). Valence Aware Dictionary for sEntiment Reasoning is a validated sentiment analyzer that determines the sentiment of a comment on a −1 to +1 scale, with −1 being the most negative sentiment possible, +1 being the most positive sentiment possible, and 0 being a completely neutral sentiment. Valence Aware Dictionary for sEntiment Reasoning determines the sentiment of text on the basis of lexicon and rules such as negation. The sentiment scores of comments containing the name of a specific drug were tracked over time, and the median sentiment of all drugs to treat psoriasis was compared. In this study, sentiment scores for the comments ranged from −0.9981 to 0.9997. RE was also used to identify ADRs by searching for comments containing the name of a drug and at least one term reflecting a known ADR for the drug.

Sentiment scores over time or between treatments were compared using boxplots and Kruskal−Wallis rank-sum tests. Confidence intervals for percentages are based on the Wilson score intervals. Two-sided P-values < 0.05 were considered statistically significant. All analyses are done using Python (version 3.7) and R (version 4.1.2).

Data availability statement

The datasets generated during and/or analyzed in this study are available from the corresponding author upon reasonable request. Primary data for this study may be accessed at https://github.com/JackCummins493/eczema_psoriasis.

CONFLICT OF INTEREST

The authors state no conflict of interest.

ACKNOWLEDGMENTS

Institutional Review Board approval was waived by the Mass General Brigham Institutional Review Board on the basis of a determination of not human subjects research (identification number 514).

AUTHOR CONTRIBUTIONS

Conceptualization: JC, VN; Data Curation: JC; Formal Analysis: JC, GZ, VN; Investigation: JC, GZ, VN; Methodology: JC, GZ, VN; Project Administration: VN; Resources: JC, GZ; Software: JC, GZ; Supervision: VN; Validation: JC, GZ; Writing – Original Draft Preparation: JC; Writing – Review and Editing: GZ, VN

REFERENCES

Baumgartner J, Zannettou S, Keegan B, Squire M, Blackburn J. The Pushshift reddit dataset. Proceedings of the int AAAI conf. web soc Media 2020;14:830–9.

Bird S, Klein E, Loper E. Natural language processing with Python. O'Reilly; 2009.

Buntinx-Krieg T, Caravaglio J, Domozych R, Dellavalle RP. Dermatology on Reddit: elucidating trends in dermatologic communications on the world wide web. Dermatol Online J 2017;23:13030.

Couto FM. Text processing. Adv Exp Med Biol 2019;1137:45–60.

Falissard B, Simpson EL, Guttman-Yassky E, Papp KA, Barbarot S, Gadkari A, et al. Qualitative assessment of adult patients' perception of atopic dermatitis using natural language processing analysis in a cross-sectional study [published correction appears in Dermatol Ther (Heidelb) 2020;10:307–10] Dermatol Ther (Heidelb) 2020;10:297–305.

Heaton E, Levender MM, Feldman SR. Timing of office visits can be a powerful tool to improve adherence in the treatment of dermatologic conditions. J Dermatolog Treat 2013;24:82–8.

Hunter JD. Matplotlib: a 2D graphics environment. Comput Sci Eng 2007;9:90–5.

Hutto C, Gilbert E. VADER: a parsimonious rule-based model for sentiment analysis of social media text. Proceedings of the int AAAI conf web soc Media 2014;8:1.

Moberg C, Alderling M, Meding B. Hand eczema and quality of life: a population-based study. Br J Dermatol 2009;161:397–403.

Murage MJ, Tongbram V, Feldman SR, Malatestinic WN, Larmore CJ, Muram TM, et al. Medication adherence and persistence in patients with rheumatoid arthritis, psoriasis, and psoriatic arthritis: a systematic literature review. Patient Prefer Adherence 2018;12:1483–503.

Okon E, Rachakonda V, Hong HJ, Callison-Burch C, Lipoff JB. Natural language processing of Reddit data to evaluate dermatology patient experiences and therapeutics. J Am Acad Dermatol 2020;83:803–8.

Patel N, Feldman SR. Adherence in atopic dermatitis. Adv Exp Med Biol 2017;1027:139–59.

Patel NU, D'Ambra V, Feldman SR. Increasing adherence with topical agents for atopic dermatitis. Am J Clin Dermatol 2017;18:323–32.

Řehůřek R, Sojka P. Software framework for topic modelling with large corpora. In: Proceedings of the LREC 2010 workshop new chall NLP Framew. Valletta, Malta: ELRA; 2010. p. 45–50.

Sunkureddi P, Doogan S, Heid J, Benosman S, Ogdie A, Martin L, et al. Evaluation of self-reported patient experiences: insights from digital patient communities in psoriatic arthritis. J Rheumatol 2018;45:638–47.

Van Rossum G, Drake FL. Python 3 reference manual. Scotts Valley, CA: CreateSpace; 2009.

Wu Q, Xu Z, Dan YL, Zhao CN, Mao YM, Liu LN, et al. Seasonality and global public interest in psoriasis: an infodemiology study. Postgrad Med J 2020;96:139–43.