AMERICAN MEDICAL WRITERS ASSOCIATION AMWA JOURNAL V38 N4 / 2023

ORIGINAL RESEARCH

Data Mining FDA Docket 2019-N-1482: Content, Sentiment, and Metadata

Michael J. Madson, PhD[1] **and Andrew Madson, MA**[2] / [1]Arizona State University, Mesa, AZ; [2]Western Governors University, Salt Lake City, UT

ABSTRACT

The legal status of cannabis continues to evolve, raising challenges for medical writers who work in population health and drug safety. To guide messaging, research has investigated how the public perceives cannabis, often relying on surveys or "big data" analyses of social media. However, these methods can be costly. As a supplement, we explored comments posted to a United States Food and Drug Administration docket on cannabis science and risk, which may offer an accessible, purposive, cost-effective source of data. We applied a multipronged methodology that involved content analysis, sentiment analysis, and metadata analysis. The findings suggest that broad messaging on cannabis may have limited effectiveness. Instead, medical writers should design messages that emphasize the risks of particular products as well as express empathy for consumers suffering from specific conditions. Moreover, among other things, the findings suggest that medical writers should use the terms "cannabis" and "marijuana" intentionally, considering the implications of each. In the future, research should develop methods to further segment drug consumers demographically and psychographically, building on the methodology that we present here. This research may inform not just messaging but regulatory writing practices and state drug policies.

The legal status of cannabis has been debated in numerous countries, including the United States (US), where the legal cannabis industry may exceed $43 billion in sales by mid-decade.[1] There have also been changes in public attitudes. A recent survey by the Pew Research Center found that over the past decade, the number of US adults who oppose cannabis legalization has fallen 20 percentage points, from 52% to 32%.[2] Moreover, 9 out of 10 US adults now support the legalization of cannabis for medical or recreational use, raising numerous questions for public health.[3]

The US Food and Drug Administration (FDA) subsequently convened a hearing on May 31, 2019, to "obtain scientific data and information about the safety, manufacturing, product quality, marketing, labeling, and sale of products containing cannabis or cannabis-derived compounds."[4] Although the in-person proceedings concluded at 6:00 PM that day, the discussion has continued through the comments posted to the hearing's docket. The docket comments are broadly accessible, excepting proprietary and other sensitive information.

Comments posted to federal dockets have received little attention from medical writers and researchers in adjacent fields. Yet, there are several reasons why these comments are potentially valuable. First, the commenters are invested in the legal status of cannabis, and thus their comments provide a form of purposeful sampling (see Palinkas et al.[5]). In aggregate, their comments, similar to social media posts, may texturize our understanding of how the public perceives cannabis, offering a quick and low-cost alternative to surveys.[6] Second, Regulations.gov, where FDA dockets are hosted, informs commenters that what they submit may be displayed there. The site relatedly informs commenters that, in addition to official agency uses, third parties may access or collect comments for their own purposes.[7-9] Third, the FDA has stated that comments "can, and do, influence agency decisions,"[10] potentially impacting the work of medical writers in regulatory settings.

Using a multipronged methodology, this study explored who the commenters are on FDA docket 2019-N-1482 and what they are commenting about. Our specific research questions were

1. What are common themes and concepts in the comments?
2. What sentiment is expressed in the comments?
3. How did the commenters self-identify, based on the demographic categories that the FDA provides?
4. What geolocations are the comments attached to?

The answers to these questions provided helpful insights into docket comments, suggesting ways that medical writers can gauge public perceptions of cannabis.

ISSN 2163-5315

METHODS

Our methods involved 3 general steps: scraping the data, clearing the data, and visualizing the data. We briefly explain each below.

Scraping the Data

Using a custom script in Python, we scraped all of the comments posted to the docket by January 2021 (n = 4,300). We also scraped commenter geolocation and demographic category (eg, individual consumers, industry representatives, health care professionals, members of government, etc.). Commenters can choose whether to include these metadata or not.

Cleaning the Data

This consisted of several sub-steps that are common in data analytics. We removed leading and trailing whitespace, standardized spellings (drug and chemical names, in particular), and filtered out stopwords. Our stopwords were honorifics "thanks," "thank you," and "sincerely" because these words convey phatic rather than substantive meaning in the data set. They also included prepositions (eg, "of," "to," "at") and coordinating conjunctions (eg, "so," "and," "but"), which tend to carry little semantic meaning.

For content analysis, we used the stemming algorithm in Leximancer, a data analytics program that is commonly used in health-related research.[11-14] For sentiment analysis, we lemmatized the data to optimize output from Valence Aware Dictionary and sEntiment Reasoner (VADER), as Symeonidis et al.[15] recommend.

Visualizing the Data

We visualized the data both demographically and psychographically. To do so, we applied content analysis, sentiment analysis, and what we called "metadata analysis." For content analysis, we uploaded the data set to Leximancer, as mentioned above. Leximancer calculates the presence and frequency of key concepts as well as their co-occurrence.[16(p8)] Concepts are clusters of terms that tend to "travel together" in a data set and, when grouped together as themes, maximize the relevancy of all the other words in a data set.[16(p11)] Based on the concepts it detects, Leximancer produces a heat map showing the relationships between themes and their underlying concepts as well as frequency. The former is indicated by the location of a theme or concept on the map and the latter by its color: the "hotter" the color (with red being the hottest, purple the coldest), the greater the frequency.

For sentiment analysis, we used VADER, which takes a "bag of words" approach. That is, it analyzes lexical features that, based on their meanings, are typically perceived as positive, negative, or neutral.[17] In our study, we used VADER to calculate a compound sentiment score for each docket comment and then average a final score for the whole data set. For both subjectivity and polarity, sentiment scores are normalized between –1.0 (negative sentiment) and 1.0 (positive sentiment).[17]

For metadata analysis, we focused on how commenters self-identified as well as where the comments were geolocated. Specifically, we quantified the frequency of each FDA demographic category, each country attached to the comments, and each US state attached to the comments. We charted these findings through Microsoft Excel and Tableau.

RESULTS

What Are Common Themes and Concepts in the Docket Comments?

Our content analysis with Leximancer identified 10 common themes in the data set, which are displayed in Figure 1.

The most common theme was CBD, referring to cannabidiol (10,954 occurrences). Its primary concept, CBD, tended to co-occur with oil (2,093 co-occurrences), use (1,902), take (1,260), helped (1,110), relief (419), milligrams (364), daily (340), doctor (255), and dose (243).

The next most common theme was pain (8,138 occurrences). Its primary concept, pain, tended to co-occur with chronic (537 co-occurrences), anxiety (519), life (400), sleep (369), work (309), arthritis (280), able (275), tried (251), started (230), better (214), year (182), depression (175), old (121), days (118), down (101), and symptoms (99).

The third most common theme was medical (6,883 occurrences). Its primary concept, medical, tended to co-occur with effects (305 co-occurrences), prescription (175), need (167), people (154), issues (112), conditions (106), patients (97), cause (74), treatment (73), active (38), and disease (34).

The fourth most common theme was products (6,188 occurrences). Its primary concept, products, tended to co-occur with hemp (633 co-occurrences), consumer (395), testing (288), benefits (279), supplement (273), extract (266), pharmaceutical (136), food (266), companies (245), potential (190), and form (133).

The fifth most common theme was health (5,591 occurrences). Its primary concept, health, tended to co-occur with believe (69 co-occurrences), levels (67), access (58), children (57), consider (56), natural (49), available (45), allow (43), provide (42), medicine (42), THC or tetrahydrocannabinol (39), and quality (35).

The sixth most common theme was cannabis (5,043 occurrences). Its primary concept, cannabis, tended to co-occur with regulations (288 co-occurrences), plant (269), FDA (231), support (196), compounds (140), safety (95),

public (91), industry (85), market (81), cannabinoids (74), information (53), data (51), and based (35).

The seventh most common theme was drug (4,197 occurrences). Its primary concept, drug, tended to co-occur with legal (97 co-occurrences), control (70), alcohol (49), research (43), studies (40), states (33), and government (33).

The eighth most common theme was time (2,824 occurrences). Its primary concept, time, tended to co-occur with seizures (53 co-occurrences), family (36), body (29), and cancer (20).

The ninth most common theme was marijuana (1,833). Its primary concept, marijuana, tended to co-occur with substance (55), law (49), DEA or Drug Enforcement Agency (43), respondent (31), and money (20).

The tenth most common theme was months (878 occurrences). Its primary concept, months, tended to co-occur with night (25).

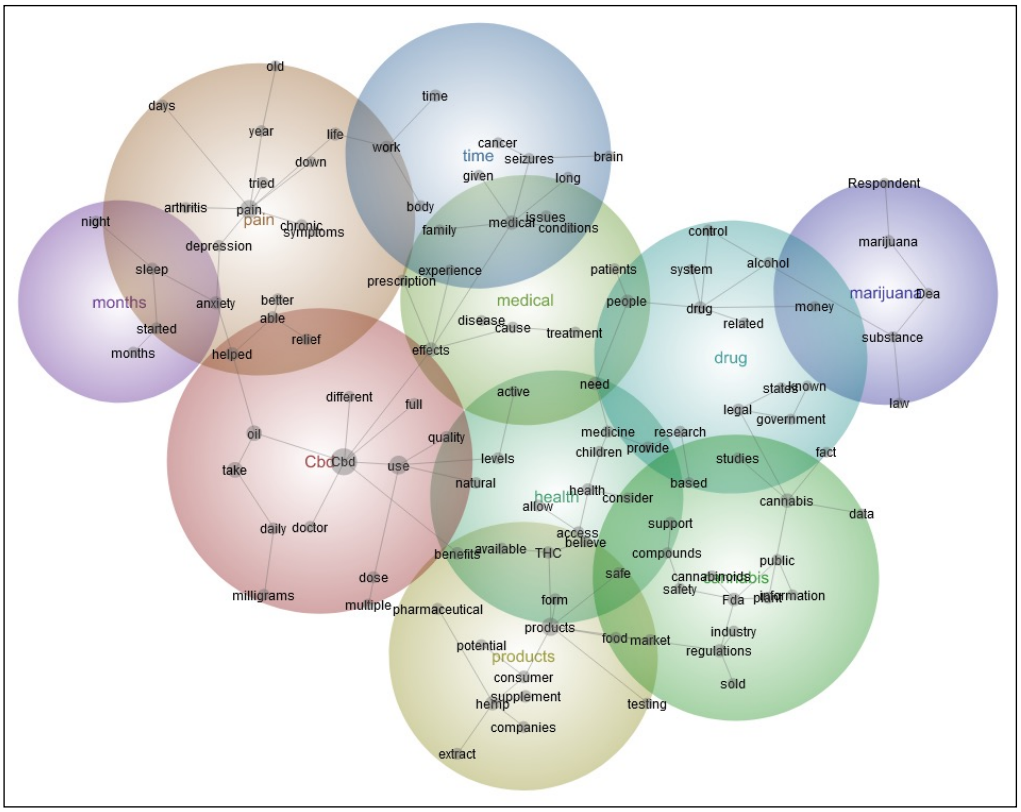

Figure 1. A heat map generated by Leximancer showing the relationships between themes and their underlying concepts as well as frequency. Relationships are indicated by the location of a theme or concept, and frequency is indicated by hue.

What Is the Sentiment in the Comments?

We found that the data set had a mean subjectivity score of 0.418, with a standard deviation of 0.190. The data set had a mean polarity score of 0.121, with a standard deviation of 0.167.

How Did the Commenters Self-Identify?

The FDA docket did not require commenters to self-identify by selecting a demographic category. In our sample, only 467 commenters did choose to self-identify: as individual consumers (81 commenters), health professionals (8), international public citizens (1), or representatives of various organizations.

Most commenters affiliated with an organization chose the most general demographic categories, such as other organizations (157 commenters), association (103), or private industry (32). Some were more specific, self-identifying as representatives of the drug industry (23), a consumer group (16), the food industry (8), a health care association (5), or international industry (4). Some commenters self-identified as representatives of local (1), state (2), federal (3), or other government (15) as well as academia (5) or the media (2). See Figure 2.

What Geolocations Are the Comments Attached to?

The majority of comments were not geolocated, but slightly more than two-fifths (1,821 comments) were. A few comments were reportedly from a geolocation outside of the US: the United Kingdom (3), Canada (3), Australia (2), Norway (1), South Korea (1), or Germany (1).

Most were from a geolocation in the US, as displayed in Figure 3. All 50 states were represented, and so was the District of Columbia. The states with the most comments were California (183 comments), Texas (152), and Florida (128), Missouri (82), New York (73), North Carolina (66), Illinois (62), Colorado (57), Kansas (57), and Wisconsin (54). Several other states had at least 50 comments: Georgia (52 comments), Oklahoma (51), and Virginia (50).

Eight states had fewer than 50 comments but at least 30: Washington (46 comments), Arizona (43), Michigan (41), Ohio (40), Massachusetts (35), Pennsylvania (35), Oregon (32), and Tennessee (30).

Nineteen states and the District of Columbia had fewer than 30 comments but at least 10: New Jersey (29 comments), Maryland (27), South Carolina (26), Indiana (24),

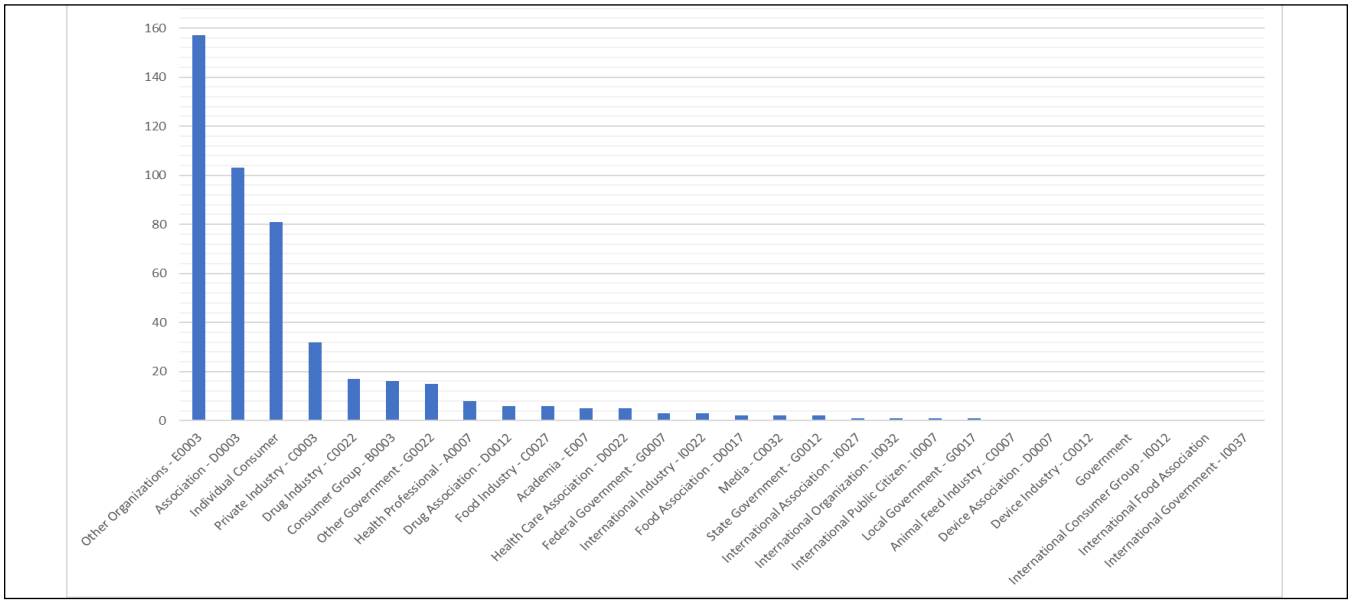

Figure 2. How the commenters self-identified, based on the demographic categories provided by the FDA (n = 467).

Kentucky (24), Arkansas (23), the District of Columbia (23), Alabama (22), Minnesota (22), Nebraska (20), Connecticut (19), Nevada (19), Utah (19), Iowa (17), New Mexico (15), Louisiana (11), Montana (11), Vermont (11), Idaho (10), and New Hampshire (10).

The states with the fewest comments were Mississippi (9 comments), West Virginia (9), Hawaii (8), Rhode Island (8), Alaska (6), Wyoming (6), Maine (5), North Dakota (3), South Dakota (3), and Delaware (2). The average number of comments per state was 35.5 with a standard deviation of 36.4.

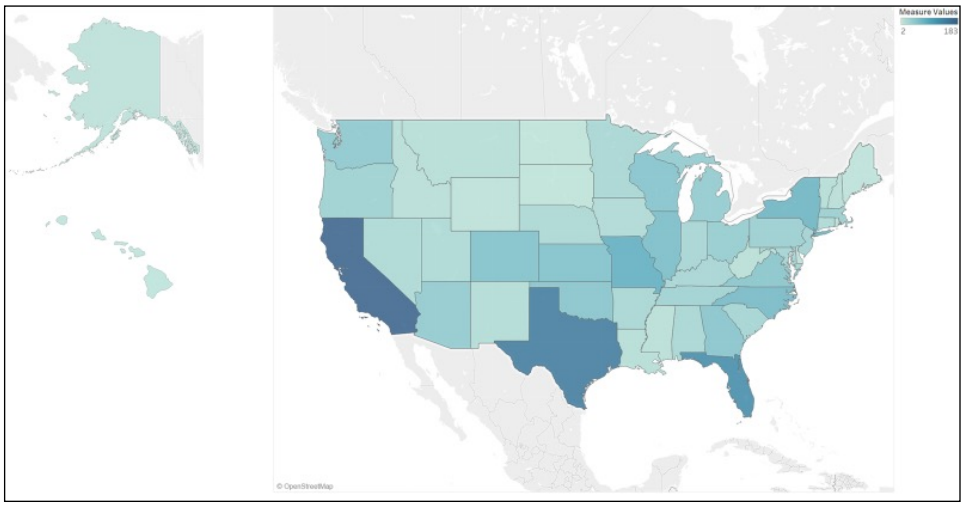

Figure 3. The geolocation of comments in the data set, specific to the US (n = 1,810). Darker shades indicate a greater number of comments.

DISCUSSION

Prior "big data" research that explores public perceptions of cannabis has generally focused on social media posts.[18-28] Expanding on this research, our study investigated public comments to FDA docket 2019-N-1482, applying content analysis, sentiment analysis, and metadata analysis in ways that may be relevant for medical writers.

The content analysis suggests that the commenters were less concerned with cannabis in the abstract and more concerned with specific products and symptoms. Of particular concern were CBD and hemp and the treatment of pain, anxiety, and sleep issues. This finding may have relevance for public health messaging: rather than targeting cannabis in general, messages might be more effective if they discuss the risks associated with particular products or if they express empathy for consumers suffering from particular symptoms or conditions.

The concepts "use" and "take" appeared frequently in the data. This makes sense, given that a large share of commenters who chose a demographic category self-identified as individual consumers. Future messaging should strategically employ different verbs, such as "use" and "take," so that medical writers can evaluate the effects. On first glance, "take" may have a stronger association with health and medical discourses. "Use" may have a stronger association with illicit or recreational activity. Such associations may have a significant influence on a message's overall effectiveness.

The content analysis also suggests that cannabis and marijuana have different semantic orientations in the data

set. "Cannabis" was associated with concepts that seem regulatory and scientific, such as safety, public, regulations, compounds, industry, cannabinoids, and data. "Marijuana" may have a more legalistic or punitive orientation, considering its co-occurrences with concepts like substance, law, and money. Future studies could test how participants respond to messages about "cannabis" compared with messages about "marijuana." In the meantime, medical writers should use the 2 terms intentionally, considering the possible implications of each. Although common in everyday speech, "marijuana" may carry more stigma.

The sentiment scores indicated positive polarity and subjectivity. The polarity score suggests that commenters generally had neutral or favorable views of cannabis, which should be confirmed through additional research. The subjectivity score suggests that commenters tended to express personal feelings, opinions, and preferences. It is unknown whether FDA officials will consider these subjectivities to be "sound grounds" for decision-making.10 Because the number of comments per state was so variable (the average being 35.5 with a standard deviation of 36.4), we did not calculate sentiment scores by state. A richer level of granularity that allows comparisons across states would improve on the methodology that we reported here. That granularity could also support inter- and intra-state policy evaluations, suggesting how cannabis policies may have "moved the needle."

The metadata analysis was small scale, as only 10.9% of the comments indicated the commenters' demographic category. More than half of these comments were from other organizations or associations, and slightly less than a fifth were from individual consumers. Because these demographic categories are self-reported, they cannot be fully verified. Future studies might develop techniques of categorizing demographic information in the comments themselves, beyond the limited categories provided by the FDA. It would be interesting, for instance, to examine how sentiment may vary by occupation, education level, income, age, and gender. The findings could support more targeted medical and regulatory communication regarding cannabis as well as policy development.

Geographically, the metadata analysis indicated that the docket has attracted comments from all 50 US states and the District of Columbia as well as 6 countries besides the US. More than half of the comments were not geolocated. Of those that were, about a third came from just 5 states: California, Texas, Florida, Missouri, and New York. Because most comments were not geolocated, it is not possible to determine the representativeness of the docket comments. Indeed, the comments may not be representative of public opinions toward cannabis writ large. Yet, the median number of comments per state and standard deviation suggests considerable geographic variation in the docket's "public participation" and "open exchange of ideas."[29]

CONCLUSION

At minimum, federal docket comments seem well suited to hypothesis generation based on themes/concepts, sentiment, and metadata. Future studies should explore ways to further segment drug consumers demographically and psychographically, building on the multipronged methodology we described here. These studies may inform not just messaging but regulatory communication and state drug policy, supporting the work of medical writers.

Author declaration and disclosures: *The authors note no commercial associations that may pose a conflict of interest in relation to this article.*

Author contact: *Michael J. Madson, michael.madson@asu.edu*

References

1. Dorbian I. Legal cannabis market projected to rack up $43 billion by 2025. *Forbes*. Published June 18, 2021. Accessed October 4, 2022. https://www.forbes.com/sites/irisdorbian/2021/06/18/legal-cannabis-market-projected-to-rack-up-43-billion-by-2025-says-new-study/?sh=35f070d336b4
2. Daniller A. Two-thirds of Americans support marijuana legalization. Pew Research Center. Published November 14, 2019. Accessed October 4, 2022. https://www.pewresearch.org/fact-tank/2019/11/14/americans-support-marijuana-legalization/
3. Van Green T. Americans overwhelmingly say marijuana should be legal for recreational or medical use. Pew Research Center. Published April 16, 2021. Accessed October 4, 2022. https://www.pewresearch.org/fact-tank/2021/04/16/americans-overwhelmingly-say-marijuana-should-be-legal-for-recreational-or-medical-use/
4. Food and Drug Administration. Scientific data and information about products containing cannabis or cannabis-derived compounds; public hearing; request for comments. 84 FR 12969. Published April 3, 2019. Accessed October 4, 2022. https://www.federalregister.gov/documents/2019/04/03/2019-06436/scientific-data-and-information-about-products-containing-cannabis-or-cannabis-derived-compounds
5. Palinkas LA, Horwitz SM, Green CA, Wisdom JP, Duan N, Hoagwood K. Purposeful sampling for qualitative data collection and analysis in mixed method implementation research. *Adm Policy Ment Health*. 2015;42(5):533-544. doi:10.1007/s10488-013-0528-y
6. Tran T, Kavuluru R. Social media surveillance for perceived therapeutic effects of cannabidiol (CBD) products. *Int J Drug Policy*. 2020;77:102688. doi:10.1016/j.drugpo.2020.102688
7. Privacy notice. Regulations.gov. Accessed October 4, 2022. https://www.regulations.gov/privacy-notice
8. User notice. Regulations.gov. Accessed October 4, 2022. https://www.regulations.gov/user-notice
9. Food and Drug Administration. Consumer comments—public posting and availability of comments submitted to Food and Drug Administration dockets. 80 FR 56469. Published September 18, 2015. Accessed October 4, 2022. https://www.federalregister.gov/documents/2015/09/18/2015-23389/consumer-comments-public-posting-and-availability-of-comments-submitted-to-food-and-drug#furinf
10. Food and Drug Administration. The importance of public comment to the FDA. Published September 14, 2018. Accessed

October 4, 2022. https://www.fda.gov/drugs/information-consumers-and-patients-drugs/importance-public-comment-fda
11. Dadich A, Wyer M. Patient involvement in healthcare-associated infection research: a lexical review. *Infect Cont Hosp Ep*. 2018;39(6):710-717. doi:10.1017/ice.2018.62
12. Dadich A, Moore L, Eapen V. What does it mean to conduct participatory research with Indigenous peoples? A lexical review. *BMC Public Health*. 2019;19(1):1388. doi:10.1186/s12889-019-7494-6
13. Singleton JA, Lau ET, Nissen LM. Waiter, there is a drug in my soup–using Leximancer® to explore antecedents to pro-environmental behaviours in the hospital pharmacy workplace. *Int J Pharma Pract*. 2018;26(4):341-350. doi:10.1111/ijpp.12395
14. Viana JN, Edney S, Gondalia S, Mauch C, Sellak H, O'Callaghan N, Ryan JC. Trends and gaps in precision health research: a scoping review. *BMJ Open*. 2021;11(10):e056938. doi:10.1136/bmjopen-2021-056938
15. Symeonidis S, Effrosynidis D, Arampatzis A. A comparative evaluation of pre-processing techniques and their interactions for Twitter sentiment analysis. *Expert Syst Appl*. 2018;110:298-310. doi:10.1016/j.eswa.2018.06.022
16. Leximancer user guide: release 4.5. Leximancer. Published March 10, 2021. Accessed October 4, 2022. https://static1.squarespace.com/static/5e26633cfcf7d67bbd350a7f/t/60682893c386f915f4b05e43/1617438916753/Leximancer+User+Guide+4.5.pdf
17. Hutto CJ. VADER sentiment analysis. GitHub. Accessed October 4, 2022. https://github.com/cjhutto/vaderSentiment
18. Cavazos-Rehg PA, Krauss M, Fisher SL, Salyer P, Grucza RA, Bierut LJ. Twitter chatter about marijuana. *J Adolesc Health*. 2015;56(2):139-145. doi:10.1016/j.jadohealth.2014.10.270
19. Cavazos-Rehg PA, Sowles SJ, Krauss MJ, Agbonavbare V, Grucza R, Bierut L. A content analysis of tweets about high-potency marijuana. *Drug Alcohol Depend*. 2016;166:100-108. doi:10.1016/j.drugalcdep.2016.06.034
20. Cavazos-Rehg PA, Krauss MJ, Cahn E, Lee KE, Ferguson E, Rajbhandari B, Sowles SJ, Floyd GM, Berg C, Bierut LJ. Marijuana promotion online: an investigation of dispensary practices. *Prev Sci*. 2019;20(2):280-290. doi:10.1007/s11121-018-0889-2
21. Lamy FR, Daniulaityte R, Sheth A, Nahhas RW, Martins SS, Boyer EW, Carlson RG. "Those edibles hit hard": exploration of Twitter data on cannabis edibles in the US. *Drug Alcohol Depend*. 2016;164:64-70. doi:10.1016/j.drugalcdep.2016.04.029
22. Månsson J. A dawning demand for a new cannabis policy: a study of Swedish online drug discussions. *Int J Drug Policy*. 2014;25(4):673-681. doi:10.1016/j.drugpo.2014.04.001
23. Meacham MC, Paul MJ, Ramo DE. Understanding emerging forms of cannabis use through an online cannabis community: an analysis of relative post volume and subjective highness ratings. *Drug Alcohol Depend*. 2018;188:364-369. doi:10.1016/j.drugalcdep.2018.03.041
24. Meacham MC, Roh S, Chang JS, Ramo DE. Frequently asked questions about dabbing concentrates in online cannabis community discussion forums. *Int J Drug Policy*. 2019;74:11-17. doi:10.1016/j.drugpo.2019.07.036
25. Moreno MA, Gower AD, Jenkins MC, et al. Social media posts by recreational marijuana companies and administrative code regulations in Washington State. *JAMA Netw Open*. 2018;1(7):e182242.
26. Pang RD, Dormanesh A, Hoang Y, Chu M, Allem JP. Twitter posts about cannabis use during pregnancy and postpartum: a content analysis. *Subst Use Misuse*. 2021;56(7):1074-1077.
27. Sheikhan NY, Pinto AM, Nowak DA, et al. Compliance with Cannabis Act regulations regarding online promotion among Canadian commercial cannabis-licensed firms. *JAMA Netw Open*. 2021;4(7):e2116551. doi:10.1001/jamanetworkopen.2021.16551
28. Van Draanen J, Krishna T, Tsang C, Liu S. Keeping up with the times: how national public health and governmental organizations communicate about cannabis on Twitter. *Subst Abuse Treat Prev Policy*. 2019;14(1):38. doi:10.1186/s13011-019-0224-3
29. Executive Order 13563—improving regulation and regulatory review. Published January 18, 2011. Accessed October 4, 2022. https://obamawhitehouse.archives.gov/the-press-office/2011/01/18/executive-order-13563-improving-regulation-and-regulatory-review