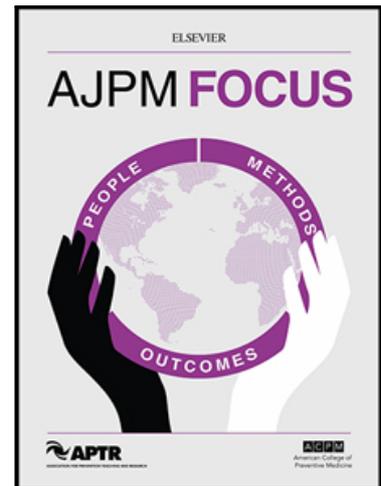


Journal Pre-proof

An Investigation of Age-Differentiated Conversations about Electronic Nicotine Delivery Systems on Reddit

Mario A. Navarro PhD , Andrea Malterud PhD ,
Zachary Cahn PhD , Laura Baum MA, MPH ,
Thomas Bukowski MPA, MA , Caroline Kery MS ,
Robert Chew MS , Annice Kim PhD

PII: S2773-0654(22)00043-8
DOI: <https://doi.org/10.1016/j.focus.2022.100045>
Reference: FOCUS 100045



To appear in: *AJPM Focus*

Received date: 21 April 2022
Revised date: 24 October 2022
Accepted date: 24 October 2022

Please cite this article as: Mario A. Navarro PhD , Andrea Malterud PhD , Zachary Cahn PhD , Laura Baum MA, MPH , Thomas Bukowski MPA, MA , Caroline Kery MS , Robert Chew MS , Annice Kim PhD , An Investigation of Age-Differentiated Conversations about Electronic Nicotine Delivery Systems on Reddit, *AJPM Focus* (2022), doi: <https://doi.org/10.1016/j.focus.2022.100045>

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Published by Elsevier Inc. on behalf of The American Journal of Preventive Medicine Board of Governors.

This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>)

An Investigation of Age-Differentiated Conversations about Electronic Nicotine Delivery

Systems on Reddit

Mario A. Navarro, PhD^a

Andrea Malterud, PhD^a

Zachary Cahn, PhD^a

Laura Baum, MA, MPH.^b

Thomas Bukowski, MPA, MA^b

Caroline Kery, MS^b

Robert Chew, MS^b

Annice Kim, PhD^b

^aOffice of Health Communication and Education

Center for Tobacco Products, Food and Drug Administration (FDA)

10903 New Hampshire Avenue, Silver Spring, MD, 20993, USA

^bRTI International, Research Triangle Park, NC

3040 East Cornwallis Road, Research Triangle Park, NC, 27709, USA

Author Note:

Correspondence concerning this article should be addressed to Mario A. Navarro, FDA Center for Tobacco Products, 10903 New Hampshire Avenue, Silver Spring, MD, 20993.

Contact: Mario.Navarro@fda.hhs.gov

Declaration of Interests: This publication represents the views of the author(s) and does not represent FDA Center for Tobacco Products position or policy. The authors have no conflicts of interest.

Funding: This work was funded by contract with Center for Tobacco Products, US Food and Drug Administration, US Department of Health and Human Services (No. HHSF223201510002B-Order #75F40119F19020).

Disclaimer: The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Food and Drug Administration.

Word Count: 1,943/2,000

Highlights

- Machine learning and qualitative coding provide context to social media analysis.
- Predicted Redditor's age allow nuanced comparisons on thematic topic by age group.
- Opposition to flavor restrictions was prominent for both age groups.
- Emergent themes by age group 13-20 years were opposition to minimum age laws and flavored ENDS discussions.
- Posts by age group 21-54 years commonly mentioned general vaping use behavior.

Abstract

Introduction: The current study analyzes age-differentiated Reddit conversations about Electronic Nicotine Delivery System [ENDS] .

Methods: The current study combines two methods to 1) predict Reddit users' age into two categories (13-20 [Underage], 21-54 [Of Legal Age]) using a machine learning algorithm and 2) qualitatively code ENDS related Reddit posts within the two groups. The 25 posts with the highest karma score (number of upvotes minus number of downvotes) for each keyword search (i.e., query) and each predicted age group were qualitatively coded.

Results: The top three, out of nine, topics that emerged were “Flavor Restriction Policies”, “Tobacco 21 Policies”, and “Use”. Opposition to flavor restriction policies was a prominent sub-category for both groups, but more common in the 21-54 group. The 13-20 group was more likely to discuss opposition to minimum age laws as well as access to flavored ENDS products. The 21-54 group commonly mentioned general vaping use behavior.

Conclusions: Users predicted to be in the Underage group posted about different ENDS-related topics on Reddit than users predicted to be in the Of Legal Age group.

Keywords: Social Media, Machine Learning, E-cigarettes, Reddit, Young Adults

Introduction

Electronic Nicotine Delivery Systems [ENDS] are the most commonly used tobacco product among U.S. youth with an estimated 4.43 million high school students and 860,000 middle school students having ever used an ENDS product as of 2021.¹ Youth ENDS use increased rapidly after 2016, in part due to the appeal of flavored products.² By 2019, mean nicotine product days (NPDs) among high school students had nearly doubled in just 2 years,³ and NYTS estimated past-30 day use of any tobacco product reached its highest level since 2000.³ In addition to the addictive properties of nicotine, reviews have identified several other harms and potential harms associated with ENDS use, including inhalation of toxins and decreases in lung function.^{4,5,6} As the ENDS product landscape is rapidly changing,⁷ social media listening provides unique methodologies to obtain rapid insights and surveillance on product discussions.

Recent qualitative studies using social media for tobacco prevention and control research rely heavily on thematic coding and content analysis of posted material. For example, Wang and colleagues posts on the social media site, Reddit, to investigate ENDS flavor mentions⁸ and Brett and colleagues coded Reddit posts to find influences and barriers to use and perceptions of JUUL.⁹ However, the lack of publicly available demographic information on users is a limitation of working with social media and may prevent researchers from understanding at-risk audiences.¹⁰ To better understand conversations of tobacco public education target audiences on Reddit, Chew et al. developed an algorithm that examines users' posts and metadata to predict and categorize Reddit users' ages into one of two groups: 13-20 (i.e., Underage, UA) or 21-54 (i.e., Of Legal Age, OLA).¹¹ These two age groups were to separate users' legal use of tobacco products and to provide an appropriate model as there were very few age references for those

above the age of 54. The current exploratory study, using the Chew et al. algorithm, investigates ENDS conversations, with a focus on Flavor Restriction and Tobacco 21 Policy discussions for posts originating from predicted UA and OLA groups.

Methods

Figure 1 summarizes the three overarching steps of identification undertaken in the current study. First, Reddit posts about vaping in general, flavor restriction policies, and Tobacco 21 policies were identified and downloaded from Brandwatch.com, a social media listening platform. Multiple search keywords were used to identify relevant posts about general vaping (e.g., vape, vaping, e-cigarette), flavor restriction policies (e.g., flavor policy) and Tobacco 21 policies (e.g., “minimum age laws,” and tobacco related words such as “cigarettes,” “vapes,” and “cigars”). These keyword groups formed three separate queries to pull the data. Searches were also restricted to English language-only posts.

Second, a previously developed age prediction model was used to predict the age group for each author as either Underage (UA, 13-20), Of Legal Age (OLA, 21-54), or uncertain.¹¹ These categories were used to examine differences in conversations depending on if the user was of legal age to use tobacco. The lower bound was selected because Reddit users must be 13 or older and those above 54 could not be appropriately classified due to the small number of individuals who fell into this category during the development of the model. The age prediction model uses the gradient boosted trees algorithm¹² to predict the probability that each user belongs to either the UA or OLA age groups. Analogous to logistic regression, predicted probabilities are generated by multiplying the trained model weights by the input variable values for each new observation, summing them together, and applying an inverse logit transformation. There are 15 input variables required for the model to generate predictions, spanning literary

characteristics (e.g., “sentences per comment”) to subreddit posting frequencies (e.g., “proportion of user’s posts or comments in the *r/teenagers* subreddit”). A full list of variables used in the model, as well as further background on other variables considered, variable importance, and model performance, can be found in Chew et al. 2021.¹¹ Since the model does not produce perfect predictions (test set F1 score, ~0.79), we reduced the likelihood that the model returned false positives by only considering predictions with a predicted probability greater than 0.6 for either age group. This process of rejecting predictions for which the model is most uncertain is referred to as “Classification with a reject option”¹³ in the literature. After applying the age prediction model on the posts from each query, we selected the 25 posts in each predicted age group and query with the highest karma scores (number of upvotes – number of downvotes). This resulted in 150 total posts across both age groups and three queries.

Third, two coders were trained using a standardized codebook and after achieving sufficient inter-rater reliability (percent agreement reached at least 70%), independently coded the study sample. All themes listed in the Results section were the themes in the codebook. Not all themes were present; more information is in the Results section. Posts were excluded if they mentioned marijuana/THC/CBD, were not in the English language, or not relevant to e-cigarettes.

Results

Descriptive Statistics

Eighteen posts were excluded from the predicted UA group and 24 posts were excluded from the predicted OLA group leaving 57 UA (General Vaping: 18, Flavor Restriction Policies: 18, Tobacco 21 Policies: 21) and 51 OLA (General Vaping: 13, Flavor Restriction Policies: 17, Tobacco 21 Policies: 21) posts. For each query, the range of karma scores for coded posts was

large, suggesting that the majority of highly engaged posts (i.e., high karma scores) were captured (General Vaping: Predicted UA Group: *Mean* = 3,715, *Min* = 1,212, *Max* = 8,327; Predicted OLA Group: *Mean* = 1,071, *Min* = 553, *Max* = 2,352; Flavor Restriction Policy: Predicted UA Group: *Mean* = 476, *Min* = 42, *Max* = 5,837; Predicted OLA Group: *Mean* = 438, *Min* = 248, *Max* = 1,188; Tobacco 21 Policy: Predicted UA Group: *Mean* = 62, *Min* = 17, *Max* = 376; Predicted OLA Group: *Mean* = 185, *Min* = 20, *Max* = 1,259).

Post Categories

Table 1 reports the frequency and percentages of each post code category and sub-category. Coding categories included: Flavor Restriction Policies, Tobacco 21 Policies, Use, Harm Perceptions, Products, Memes/Jokes, COVID-19, Motivations for Vaping, Barriers to Vaping, Access, Campaigns by The Center for Tobacco Products, and Other. Barriers to Vaping and Campaigns by the Center for Tobacco Products did not emerge as code categories, even though they were originally in the codebook.

For both UA and OLA groups, the categories of Flavor Restriction Policies and Tobacco 21 Policies were the most prominent (> 40%). Between the two groups, the Products and Memes/Jokes categories were more prominent for UA than the OLA. The categories of Use and Harm Perceptions were more prominent for OLA.

Demonstrating nuances between the groups, sub-category differences continued between predicted age groups. For Flavor Restriction policies, opposition was a primary sub-category for both predicted age groups, but many flavor restriction posts fell into the Other sub-category for the UA group and Skepticism for the OLA group. To clarify, Opposition was defined as “voicing clear opposition or encouraging work against an ordinance” and skepticism as “doubt about the motives behind or effectiveness of an ordinance.” Posts coded as Other were dominated by news

stories in both groups. The OLA group had nearly twice as many opposition codes as the UA group and the second most-common code for UA were links to news stories. A clear distinction between the groups is that the OLA group showed greater opposition and skepticism to flavor restriction policies.

For the Tobacco 21 Policy category, a similar pattern emerged for the UA and OLA groups with Opposition, Skepticism, and the Other sub-categories dominating the conversation, although for this topic the UA group showed greater opposition and skepticism while the OLA group posted mostly Other category news links. For the UA group, a sub-category code emerged that detailed the desire to allow 18-20 years-old ENDS users, who were previously able to use ENDS products, to continue having the ability to purchase ENDS products (Legacy Clause, sometimes referred to by posters as “grandfather clause”). Within the sub-categories of Use, the Vape Terms sub-category was the most prominent for both groups. These terms consisted of “Vapes,” “Vaping,” “Vape master,” “Ripping,” and “Juuling” for the predicted UA posts and “Fire up your rig,” “e-liquid,” “ejuice,” “nic juice,” “coils,” and “pod system” for OLA posts.

For Motivations for Vaping, the primary motivation mentioned for both UA and OLA posts was the desire to avoid cigarettes. Harm Perception posts were primarily identified for the OLA group and ranged in topic from vaping-related illnesses to feeling better after quitting. The Product category was primarily made up of brand names. For the UA group, this brand was exclusively JUUL, but OLA included others such as Lava Pods. Memes/Jokes emerged predominantly among UA posts and included visual jokes for various sorts of media and contained jokes mocking vaping. COVID-19 information, in the form of news articles, was discussed mostly by OLA posts. The Other category posts were more prominent for OLA posts

and consisted of an individual's personal relationship with vaping, usually with a form of judgment.

Discussion

This mixed method analysis of Reddit posts provided insight into ENDS online conversations by differentiating conversations by two age group (13-21, 21-54). Differences between predicted age groups emerged for both frequency of code categories and more specific content within categories. Posts were coded into the categories of Flavor Restriction Policies, Tobacco 21 Policies, Use, Motivations for Vaping, Harm Perceptions, Products, Memes/Jokes, COVID-19, and Other. Looking at the sub-categories, a more nuanced story emerged such that most posts for the UA group fell into the Other category, and Skepticism posts were most prevalent for the OLA group. A similar pattern emerged for the Tobacco 21 Policy category. One differentiating sub-category for the Tobacco 21 Policy category was the Legacy Clause code for UA posts. This study aligns with previous research, which found age restriction opposition by UA Reddit users, using e-cigarettes to avoid cigarettes,⁹ and significant discussion about JUUL¹⁴ and flavor access.^{8,15} Findings are in line with recent ENDS studies that demonstrate Twitter users' positive sentiment toward flavors.¹⁵

Future Directions

This study has implications for future research and for public health surveillance. The mixed methodologies (i.e., data science models and qualitative coding) used in the current study can be applied to a vast amount of public health topics. In addition, age algorithms have been applied to other platforms in the past and there can be an expansion of the platforms that are analyzed.¹⁶ Finally, automated data science methodologies (e.g., topic modeling) could provide a

way auto-categorize posts, making it easier to provide thematic analyses for a large amounts of data and provide a more rapid form of surveillance.

Limitations

The current study had several limitations. The keywords used in the query did not reflect all the relevant keywords that could differentiate between posts written by UA and OLA users of Reddit. Relevant posts could have been missed. Although a sample of the top 25 most engaged posts was used, the sample size is still fairly small. This may limit the generalizability of the results. Since the sample was small, it was inappropriate to conduct statistical analyses.

Conclusion

Reddit provides a robust public access data source that can be used by researchers.^{8,10-}

¹¹The current study uses a combination of methodologies to paint a picture of the current ENDS landscape. Differences were found across all three queries (General Vaping, Flavor Restriction Policies, and Tobacco 21 Policies). These differences highlight the importance of using a combination of classification tools and qualitative coding which allows researchers and public health professionals to better understand perceptions and knowledge, attitudes, and beliefs about a product, to develop more targeted messaging.

CRedit Statement

Mario Navarro: Conceptualization, Methodology, Writing – Original Draft, Writing – Review & Editing, Supervision **Andrea Malterud:** Methodology, Writing – Original Draft, Writing – Review & Editing **Zachary Cahn:** Writing – Original Draft, Writing – Review & Editing **Laura Baum:** Project Administration, Methodology, Formal Analysis, Writing – Original Draft, Writing – Review & Editing **Thomas Bukowski:** Software, Formal Analysis, Data Curation, Writing – Original Draft, Writing – Review & Editing **Caroline Kery:** Formal Analysis, Data Curation, Writing – Review & Editing **Robert Chew:** Software, Formal Analysis, Data Curation, Writing – Original Draft, Writing – Review & Editing **Annice Kim:** Conceptualization, Supervision, Writing – Review & Editing

References

1. Gentzke AS, Wang TW, Cornelius M, et al. Tobacco product use and associated factors among middle and high school students – National Youth Tobacco Survey, United States, 2021. *MMWR* 2022;71:1-29.
2. Groom AL, Thanh-Huyen, TV, Kesh A, et al. Correlates of youth vaping flavor preferences. *Prev Med Reports* 2020;18:1-6. doi: 10.1016/j.pmedr.2020.101094
3. Sun R, Mendez D, Warner KE. Trends in nicotine product use among US adolescents, 1999-2020. *JAMA Network Open* 2021;4:1-11. doi: 10.1001/jamanetworkopen.2021.18788
4. Domenico L, DeRemer CE, Nichols KL, et al. Combatting the epidemic of e-cigarette use and vaping among students and transitional-age youth. *Curr Psychopharmacol* 2021;10:5-16. doi: 10.2174/2211556009999200613224100.
5. Singh S, Windle SB, Filion KB, et al. E-cigarettes and youth: Patterns of use, potential harms, and recommendations. *Prev Med* 2020;133:1-8. doi:10.1016/j.ypmed.2020.106009.
6. Cahn Z, Drope J, Douglass CE, et al. Applying the population health standard to the regulation of electronic nicotine delivery systems. *Nicotine Tob Res* 2021;23:780-789. doi:10.1093/ntr/ntaa190.
7. Owotomo O, Walley S. The youth e-cigarette epidemic: updates and review of devices, epidemiology and regulation. *Curr Prob Pediatr Ad* 2022; in press.
8. Wang L, Zhan Y, Li Q, Zeng DD, Leischow SJ, Okamoto J. An examination of electronic cigarette content on social media: Analysis of e-cigarette flavor content on Reddit. *Int J Environ Res Pu* 2015;12:14916-14935. doi:10.3390/ijerph121114916.
9. Brett EI, Stevens EM, Wagener TL, Leavens EL, Morgan TL, Cotton WD, Hébert ETA. content analysis of JUUL discussions on social media: Using Reddit to understand patterns and perceptions of JUUL use. *Drug Alcohol Depend* 2019;194:358-362. doi: 10.1016/j.drugalcdep.2018.10.014.
10. Sharma R, Wigginton B, Meurk C, Ford P, Gartner CE. Motivations and limitations associated with vaping among people with mental illness: A qualitative analysis of Reddit discussions. *Int J Environ Res Pu* 2017;14: 7-21. doi:10.3390/ijerph1401007.
11. Chew R, Kery C, Baum L, Bukowski T, Kim A, Navarro M. Predicting Age Groups of Reddit Users based on Posting Behavior and Metadata: Comparative Study of Classification Models. *JMIR Public Health and Surveillance*, 2021;7,1-13. doi:10.2196/25807.
12. Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Statist* 2001; 29:1189-1232.

13. Herbei R, Wegkamp MH. Classification with reject option. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique* 2006;34:709-721.
14. Zhan Y, Zhang Z, Okamoto JM, et al. Underage JUUL use patterns: content analysis of Reddit messages. *J Med Internet Res* 2019;21:e13038. doi:10.2196/13038.
15. Lu X, Chen L, Yuan J, Luo J, Luo J, Xie Z, Li D. User perceptions of different electronic cigarette flavors on social media: Observational study. *J Med Internet Res* 2020;22:1-12. doi:10.2196/17280
16. Kim AE, Chew R, Wegner M. Estimated ages of JUUL Twitter followers. *JAMA Pediatrics* 2019;173:690-692. doi: 10.1001/jamapediatrics.2019.0922

Journal Pre-proof

LIST OF FIGURES

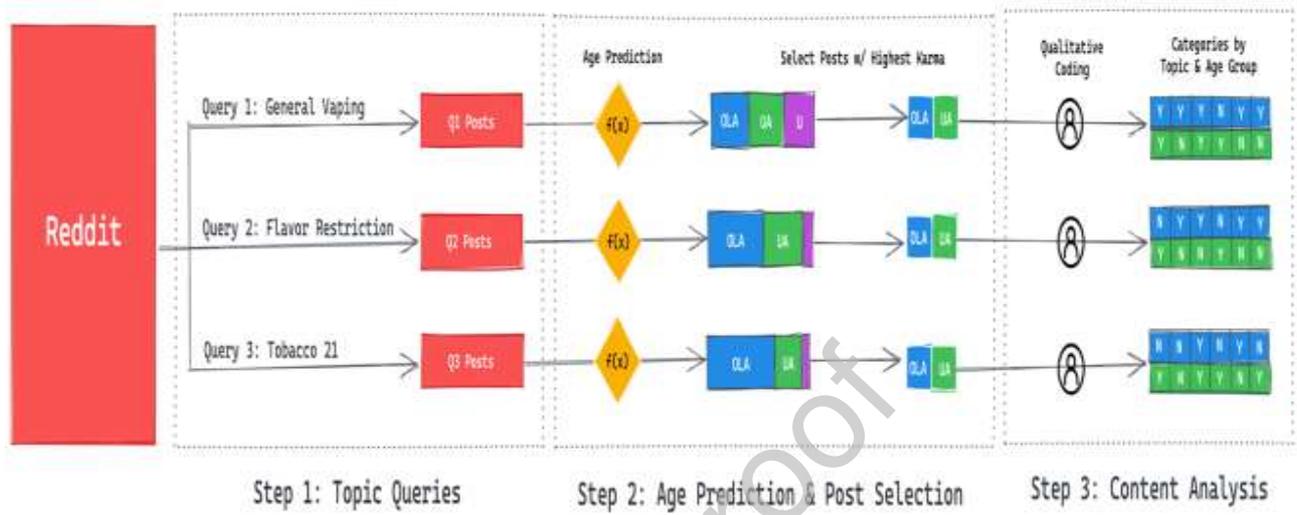
Figure 1: Study Design Flow Chart

Table 1: Post Category Prevalence in Both Predicted Underage and Of Legal Age Post Authors		
Post Category or Sub-Category	Underage <i>n</i> (%)	Of Legal Age <i>n</i> (%)
Flavor Restriction Policies ^a	26 (45.61)	37 (72.54)
Support	0 (0)	1 (2.70)
Oppose	9 (34.62)	17 (45.95)
Skepticism	1 (3.85)	8 (21.62)
Access	4 (15.38)	2 (5.41)
Switching	3 (11.53)	5 (13.51)
Quitting	1 (3.85)	0 (0)
Other	8 (30.77)	4 (10.81)
Tobacco 21 Policies ^b	27 (47.37)	21 (41.18)
Support	1 (3.71)	0 (0)
Oppose	8 (29.63)	4 (19.04)
Skepticism	4 (14.81)	1 (4.76)
Legacy Clause	4 (14.81)	0 (0)
Access	2 (7.41)	2 (9.52)
Switching	1 (3.70)	1 (4.77)
Quitting	0 (0)	0 (0)
Other	7 (25.93)	13 (61.91)
Use ^c	11 (19.30)	22 (43.14)
Dual	1 (9.10)	0 (0)
Switching	0 (0)	2 (9.09)
Quitting	2 (18.18)	1 (4.55)
Vape Terms	5 (45.45)	11 (50.00)
Other	3 (27.27)	8 (36.36)
Motivations for Vaping ^d	9 (15.79)	9 (17.65)
Harm Perceptions ^e	2 (3.51)	13 (25.49)
Products ^f	18 (31.58)	8 (15.69)
Memes/Jokes ^g	17 (29.82)	5 (9.80)
COVID-19 ^h	1 (1.75)	6 (11.76)
Other ⁱ	1 (1.75)	3 (5.88)

* Sub-category percentages are derived from the percent total of the parent code.

^aAny post referencing flavor restriction policies.

^bAny post referencing general mention of Tobacco 21 policies.

^cAny mention of other vaping use behaviors, including mentions of using vaping to quit cigarettes or other tobacco products.

^dAny mention or discussion of why someone vapes (e.g., makes them feel relaxed; to escape; for fun). Includes noting the motives of other users.

^eAny noted harms or perceived harms associated with vaping.

^fDescriptions, reviews, or questions about a vaping product.

^gVaping related Photo/GIF memes or jokes

^hAny general mention of COVID-19 related to vaping.

ⁱAny other conversations related to vaping tobacco not covered in the other codes.

Journal Pre-proof