Vaping at the VA: Developing an Annotated Corpus of Electronic Cigarette Mentions in Clinical Notes at the Department of Veterans Affairs

Mike Conway, MSc, PhD[1], Patrick R Alba, MS[2,3] ,
Shu-Hong Zhu, PhD[4], Olga V Patterson, PhD[2,3]

[1] Department of Biomedical Informatics, University of Utah, Salt Lake City, UT;
[2] Department of Internal Medicine, University of Utah, Salt Lake City, UT;
[3]VA Salt Lake City Health Care System, Salt Lake City, UT;
[4] Herbert Wertheim School of Public Health & Human Longevity Science, University of California San Diego, La Jolla, CA

Abstract

Use of Electronic Nicotine Delivery Systems (ENDS, colloquially known as "electronic cigarettes") has increased substantially in the United States in the decade since 2010. However, currently relatively little is known regarding the documentation of ENDS use in clinical notes. With this study, we describe the development of an annotation scheme (and associated annotated corpus) consisting of 4,351 ENDS mentions derived from Department of Veterans Affairs clinical notes during the period 2010-2020. Analysis of our corpus provides important insights into ENDS documentation practices at the VA, in addition to providing a resource for the future development and validation of Natural Language Processing algorithms capable of reliably identifying ENDS-use status.

Introduction & Motivation

Electronic cigarettes — *e-cigarettes*, *e-cigs*, *vapes*, or **E**lectronic **N**icotine **D**elivery **S**ystems (**ENDS**) — have exploded in popularity over the last ten years in the United States. It is estimated that by 2012, 75% of American adults had heard of ENDS, and 8.1% had tried them.[1] By 2014, 94% of Americans had some awareness of ENDS, and 12.6% had tried them.[2,3] Despite the growth in popularity of the product, little consensus currently exists regarding the safety of ENDS devices, with regulatory authorities, professional associations, and individual clinicians divided as to whether ENDS constitute a valuable smoking cessation tool,[4,5] or rather are a potentially harmful technology that risks eroding hard-won achievements in denormalizing smoking.[6,7] This uncertainty regarding the potential safety risk of ENDS use was further amplified by the 2019 outbreak of **E**-cigarette or **V**aping Product Use-**A**ssociated **L**ung **I**njury (**EVALI**) in the United States,[8] a development that sparked calls for enhanced regulation and surveillance of the products.[9]

Clinical practices regarding the documentation of ENDS use in the **E**lectronic **H**ealth **R**ecord (**EHR**) are currently poorly understood, and what evidence that does exist regarding documentation patterns suggests that ENDS use is massively under-documented.[10,11] This under-documentation is perhaps partially due to the absence of a standardized means of recording ENDS use in EHR systems,[12] a situation that — at least for some health systems — has recently seen changes with the introduction of specific fields for capturing ENDS use.[13] However, it holds true that, in the majority of cases, ENDS documentation, if it exists at all, is dispersed across clinical notes, thus necessitating the application of **N**atural **L**anguage **P**rocessing (**NLP**) methods or manual abstraction to extract relevant information. Further, the *type* of documentation commonly found in clinical notes (e.g. *type of ENDS device*, *duration of ENDS use*, *frequency of ENDS use*, *ENDS start date*, *ENDS end date*) is currently not well understood.

With this study, we describe the development of an annotation scheme (and associated annotated corpus) consisting of 4,351 ENDS-related mentions derived from Department of **V**eterans **A**ffairs (**VA**) clinical notes during the period 2010 to 2020, in order to better understand how clinicians document ENDS use, and to develop a validated resource for the future training and evaluation of NLP algorithms to execute the task of automatically determining ENDS use status.

Background

The VA EHR provides a unique national dataset with which to investigate ENDS documentation and usage patterns, given that it is the largest healthcare system in the United States with 1,255 healthcare facilities nationwide and is used by over nine million veterans per year.[14] The VA patient demographic is particularly interesting from a tobacco

control perspective given that smoking rates among veterans (29% as of 2015 [15]) are substantially greater than is found in the general population (14% as of 2019 [16]).

The VA collects data regarding patient smoking history and current smoking status using several different approaches at the time of patient encounter. Chief among these is the utilization of *Health Factors* (i.e. a semi-structured data field that describes smoking status and history [17]). However, health factors cannot easily be used to reliably identify the *type* of tobacco product used, and — most salient to our research question — whether a patient is an ENDS user. In the VA context, this ENDS-use information is typically found embedded in clinical text.

Automatically identifying smoking status using the EHR is a well developed research area that has utilized a number of different methods, including the use of structured data, [18] semi-structured data, [19] and NLP approaches. [20,21,22] However, there has been relatively little work focused on how ENDS use — as opposed to combustible tobacco use — is currently documented in the EHR. Notable exceptions include ENDS annotation scheme and corpus development work at Fairview Health System in Minneapolis [23] and analyses of ENDS documentation using keyword matching methods conducted at the VA, [10,24] and on a larger scale, at Kaiser Permanente. [25] All these health systems exhibited an increase in ENDS use documentation (i.e. the prevalence of ENDS-related keywords) over time, but notable, this increase is below what would be expected given the prevalence of ENDS use in the general population, indicating that ENDS use is systematically under-documented in these systems. [10]

Our overarching aim with this study is to analyze ENDS documentation patterns in VA clinical notes, including how ENDS documentation practices have changed over time. More specifically, we report on the development of an annotation scheme (and a manually annotated corpus based on that annotation scheme) suitable for the annotation of ENDS mentions in VA clinical notes. Our resulting corpus consisted of 4,351 annotations across sixteen annotation categories related to ENDS usage (e.g. *potential-user*, *passive-exposure*, *attempting-to-quit-ENDS*) that can be utilized as a resource for the future development and validation of NLP algorithms. The contribution of this work consists in (a) the development of validated annotation guidelines for annotating ENDS mentions in VA clinical notes*; and (b) based on this annotation scheme, the development of a high quality manually annotated corpus that provides important insights into the range of current and historical ENDS use documentation practices, in addition to providing a resource for the future development and validation of NLP algorithms. The development of an annotated corpus is a necessary condition for the creation and validation of an NLP algorithm capable of reliably determining ENDS use status, which in turn is required to interrogate EHR systems regarding key open epidemiological questions concerning ENDS use (e.g. *what proportion of ENDS users quit tobacco? are clinicians recommending ENDS use as a cessation method?*)

Materials & Methods

Our initial cohort consisted of VA patients with a history of smoking (as determined by VA Health Factors) who had utilized clinical services between 2010 and 2020. To better understand how ENDS terminology has changed over time, we used a set of ENDS-related keywords derived from various sources, including previous work, [24] additional iterative corpus data analysis, and a word-embedding model trained on the totality of VA notes. This process resulted in 77 keywords† that were categorized into five broad groups (*electronic cigarette*, *e-cigarette*, *e-cig*, *vape*) or other keywords (*other variants*). Using these keywords, we performed a search across our corpus of 4,604,856 million patients in order to allow us to identify the prevalence of ENDS keywords and observe changes in ENDS terminology over time. [26] This process resulted in the identification of 1,638,884 ENDS mentions for 418,170 patients across 1,191,133 clinical notes. From our cohort of VA patients, we randomly sampled notes containing ENDS mentions. We then iteratively developed an annotation scheme for VA ENDS use documentation, using prior work reported in Winden et al. [23] as our starting point. Four individuals participated in the annotation scheme construction process (authors MC & PA, in addition to two annotators with a background in nursing). The resulting annotation scheme consisted of five top level annotation categories (*Active-User*, *Usage-Unknown*, *Irrelevant*, *Former-User*, and *Non-User*) and sixteen lower level categories (see **Figure 1** for a a graphical representation of the annotation scheme workflow, with additional examples provided in **Table 1**). In addition to annotation type, each ENDS mention is associated with demographic characteristics derived from structured data in the VA EHR (including *gender*, *urban/rural location*, *outpatient/inpatient status*, and *age range*). Further, each annotation is associated with its source note type (e.g. *primary care*, *mental health*,

*Annotation guidelines are available at: `https://tinyurl.com/fhbdbnmt`

†A complete list of keywords is available at: `https://tinyurl.com/7pfambrt`

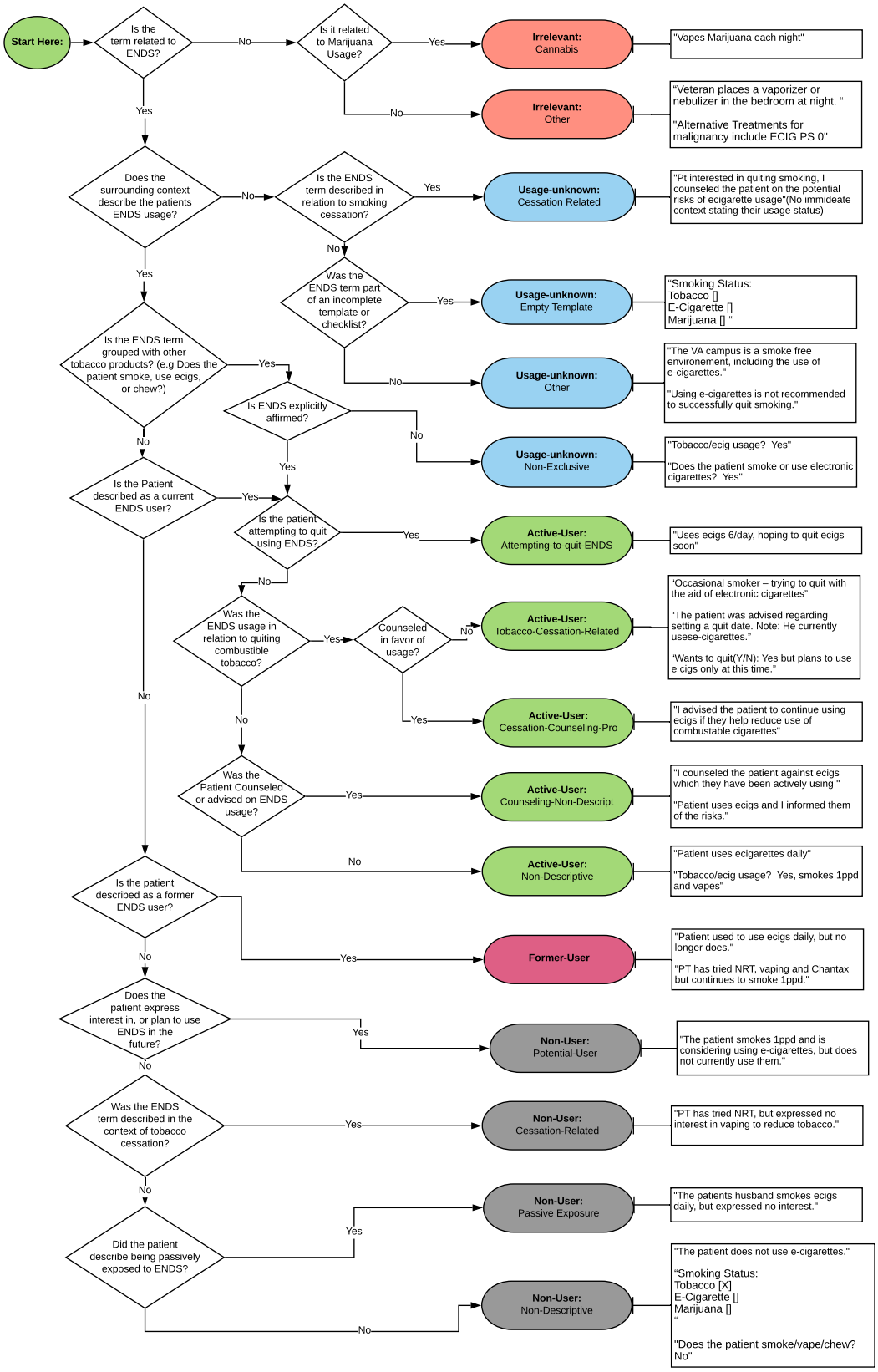

Figure 1: Annotation scheme flowchart

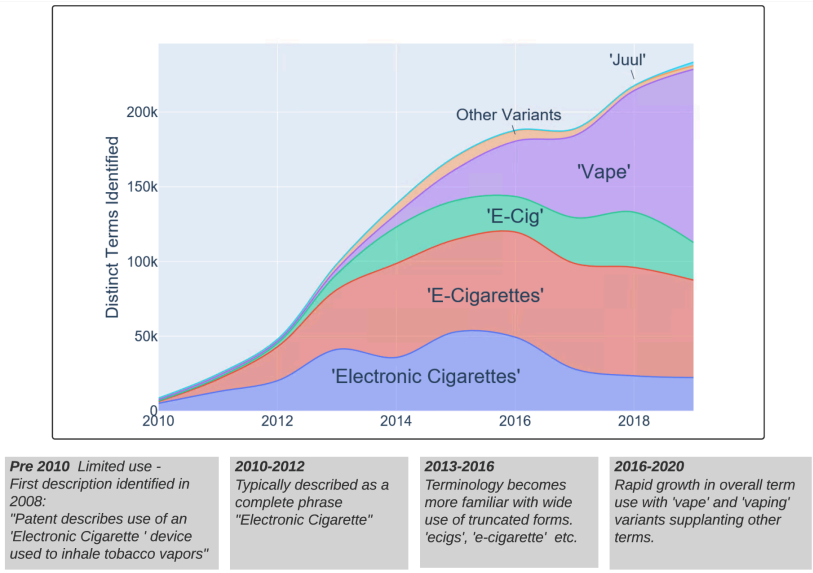

Figure 2: Changes in the prevalence of ENDS-related keywords found in VA clinical notes (2010-2020)

social work).

As can be seen from **Figure 2** there have been striking changes in ENDS-related terminology since 2010, with "vape"-related keywords superseding previously dominant ENDS keywords like "e-cigarette".[‡] Given this observation — i.e. that keywords used to refer to ENDS in clinical documentation shifted substantially over time — we elected to construct a stratified sample of ENDS mentions for the development of our annotated corpus. The initial stratified set of documents was created by selecting 200 random ENDS mentions every year from 2010-2020 (n = 2,200). In order to account for the increased variation and changes in usage in recent years, a second set of instances was selected using 250 randomly selected mentions every year from 2016-2020 (n = 1,250). The third and final set of annotations was selected to be held out for final testing of any NLP algorithms. This third set was selected using a random set of ENDS mentions proportionate to the total number of mentions identified each year (n = 901). In total, our annotated corpus consists of 4,351 ENDS mentions. Annotations were completed at the mention-level using *Chex*,[27,28] a web-based, VA-developed application. Chex provides a simple user interface that allows annotators to view a small snippet of text surrounding an anchored keyword with the ability to expand to a complete document for further context when needed.

In order to ensure annotation quality, we double annotated sections of the corpus (10% of the training set and 100% of the test set), achieving comparable agreement scores to those gained during the annotation scheme development process. We computed agreement using Cohen's kappa.[29] Our annotation effort resulted in 4,351 ENDS mentions, derived from 4,345 clinical notes and 4,213 individual patients between 2010 and 2020.[§]

Results

Using our cohort of 4,604,856 patients we identified 1,638,884 ENDS mentions for 418,170 patients across 1,191,133 clinical notes in the VA EHR and observed rapid growth in both the volume and variety of ENDS-related terms over time (see **Figure 2**). The total volume of clinical notes containing ENDS-related keywords exhibited an average annual growth rate of 52.8% over the 10 year period, with 8,234 clinical notes identified in 2010 and 202,842 in 2019; this is

[‡]Note that a simple string matching approach was used to determine prevalence of a predetermined list of ENDS-related keywords that occurred in notes over the period 2010-2020. See previous footnote for a link to the complete set of keywords used.

[§]In line with our approved University of Utah Institutional Review Board protocol, and VA policy, the annotated corpus reported on in this paper cannot be publicly shared as it contains sensitive medical data.

compared to a 2.3% annual growth rate across all clinical notes, with or without ENDS documentation.

Our annotated corpus of ENDS mentions consisted of 4,351 annotations from the period 2010-2020. **Table 1** shows a breakdown of the annotation types, with five top-level annotation categories and sixteen lower-level annotation categories. It can be seen that *Active-User* is the most frequent annotation type (43%) with *Former-User* the least frequent annotation type (3%). Of the *Active-User* group, the majority of mentions were related to tobacco cessation (either *tobacco-cessation-related*, *cessation-counseling-pro* or *counseling-non-descriptive*), with only a small proportion relating to attempts to quit ENDS. Four percent of mentions were found to be *Irrelevant* by annotators, with the majority of these *Irrelevant* posts referring to vaping cannabis. The initial annotation scheme development process yielded an agreement (kappa) of 0.7. The subsequent annotation effort yielded an overall agreement of 0.65 (0.85 for the top-level annotation categories *Irrelevant*, *Usage-Unknown*, *Active-User*, *Non-User*, and *Former-User*) on those portions of the corpus that were double annotated.

As discussed above, each ENDS mention is associated with patient demographic information derived from structured data in the VA EHR. **Table 2** shows that — consistent with the wider VA patient population — 91% of annotated ENDS mentions are derived from the notes of male patients, and 9% of annotated ENDS mentions are derived from the notes of female patients. The difference between urban and rural locations was less marked than that between male and female, with 65% of annotated ENDS mentions derived from the notes of urban patients — or rather, patients who attended a VA facility in an urban area — and 35% derived from the notes of rural patients. Regarding age range, the majority of ENDS mentions were derived from individuals between the ages of 50 and 69, with the lowest number of mentions derived from the youngest age group (≤ 29). Documentation occurred in a range of clinical note types, with *Other* (a "catch-all" category that covers several internal medicine specialties) the most frequent note type (32%). Additionally, *primary care* notes and *mental health* notes were both well represented (31% and 12%, respectively).

Annot. Type	Annot. Subtype	#Annotations	Examples
Irrelevant	Total Annotations	155 (3.56%)	
	Cannabis	86 (1.98%)	*vaping marijuana to help him with sleep.*
	Other	69 (1.59%)	*used vaporizer when sick with flu*
Usage-unknown	Total Annotations	1651 (37.95%)	
	Cessation-counseling	382 (8.78%)	*e-cigarettes may be bad for your health*
	Empty-template	13 ($< 1\%$)	*electronic cigarette: Y/N*
	Other	871 (20.02%)	*current tobacco user (excluding ecigs & hookah*
	Non-exclusive	3858 (8.84%)	*tobacco/ecig: yes*
Active-user	Total Annotations	1862 (42.79%)	
	Attempting-to-quit-ENDS	112 (2.57%)	*tapering off ecig*
	Tobacco-cessation-related	929 (21.35%)	*smoked tobacco but recently switched to vaping*
	Cessation-counseling-pro	12 ($< 1\%$)	*patient advised to continue ecig use*
	Counseling-non-descriptive	126 (2.90%)	*smoking ecigs: education provided*
	Non-descriptive	683 (15.70%)	*smokes ecigs*
Non-user	Total Annotations	540 (12.41%)	
	Potential-user	135 (3.10%)	*pt is planning on using ecig for cessation*
	Cessation-related	29 ($< 1\%$)	*pt not interested in using ecigs to quit*
	Passive-exposure	8 ($< 1\%$)	*girlfriend uses ecig*
	Non-descriptive	368 (8.46%)	*tobacco/ecig usage: No*
Former-user	Total Annotations	143 (3.29%)	*no longer vapes*

Table 1: Distribution of annotation categories (4,351 annotations in total)

Discussion & Conclusion

Our research has shown a considerable increase in ENDS documentation in the VA EHR over the period 2010-2020 (see **Figure 2**), with an average year-on-year growth rate of 53%, broadly reflecting ENDS use prevalence in the general population over the same period. In addition to the increase in the *volume* of ENDS mentions during the study period, our results show that the language of ENDS documentation has evolved over time, with terms like "vaping"

Annotation Subtype	Number of Annotations
PATIENT GENDER	
Male	3947 (91%)
Female	404 (9%)
URBAN/RURAL	
Urban	2821 (65%)
Rural	1529 (35%)
OUTPATIENT/INPATIENT	
Outpatient	3394 (78%)
Inpatient	949 (22%)
AGE RANGE	
≤ 29	177 (4%)
30-49	977 (22%)
50-69	2,392 (55%)
≥ 70	805 (19%)
NOTE TYPE	
Other	1383 (32%)
Primary care	1368 (31%)
Mental health	511 (12%)
Surgery	287 (7%)
Pharmacy	187 (4%)
Psychology	138 (3%)
Pulmonary disease	119 (3%)
Smoking cessation	77 (2%)
Emergency medicine	72 (2%)
Social work	69 (2%)
Cardiology	63 (1%)
Optometry	41 (1%)
Anesthesiology	36 (1%)

Table 2: Corpus characteristics

and "juul" (an ENDS brand name) supplanting previously popular terms like "electronic cigarette" and "e-cigarette" towards the end of the decade. This finding suggests that NLP systems designed to automatically identify ENDS use status must be flexible and adapt to change in language use over time. Unlike smoking status (i.e. combustible tobacco smoking status), where terminology is, if not standardized, then at least relatively stable over time (e.g. "2ppd", "15 pack years"), ENDS documentation patterns are likely to continue to change quickly as technology and usage trends develop. As such, NLP systems designed to identify ENDS use must also evolve over time.

Winden et al.'s work on ENDS-related annotation scheme construction and corpus development provided a starting point for our VA-specific annotation scheme development process.[23] However, there were some key differences in both the motivation of our work and in the specific characteristics of VA clinical note data that led us to substantially modify Winden et al.'s model for our specific needs. The work reported in this paper was primarily motivated by the goal of developing an annotation scheme and annotated corpus as a means of training and validating an NLP algorithm for identifying ENDS use status in the VA EHR. Given this motivation, it was necessary to include an *Irrelevant* annotation category for examples that include "false positive" ENDS-related keywords. Most of the examples of irrelevant mentions relate to cannabis use, particularly cannabis vaporization (e.g. "pt vapes marijuana"). A further distinctive feature of our annotation scheme is that it explicitly accounts for numerous types of semi-structured templates embedded in VA notes (e.g. "e-cigarette: Yes/No") and use of facility-specific boilerplate language characteristic of VA notes (e.g. "the VA campus is a smoke free environment, including the use of e-cigarettes"). Excluding cannabis-related mentions, unfilled templates, and boilerplate language that does not directly refer to a patient's specific ENDS use circumstances is a core requirement for an NLP algorithm capable of identifying ENDS use status at the VA.

Research presented in this paper is not without limitations. First, our data collection spanned the period from January 1st 2010 to March 1st 2020 (i.e. immediately before the first major impacts of the COVID-19 pandemic in the United States were felt). Given systematic and drastic changes in ENDS documentation patterns at the VA since the onset of the pandemic and the increased scrutiny of potential ENDS-related COVID-19 transmission at the VA, it will likely be necessary to annotate additional contemporary ENDS mentions in order to ensure that a resulting NLP algorithm can be successfully applied to both contemporary and post-pandemic clinical notes. Second, our annotation scheme was designed primarily for the annotation of VA clinical notes, and may not be suitable without modification for other EHR environments. Third, given the relatively small size of the corpus compared to the universe of VA clinical notes and the stratified sampling method adopted in the corpus construction process, the corpus cannot in itself form a basis for forming epidemiological conclusions regarding, for example, ENDS use prevalence in VA patients.

In conclusion, the research described in this paper provides useful insights into changes in VA ENDS-related documentation practices during the period 2010-2020, as well as providing a resource for the future development and validation of NLP algorithms designed to identify ENDS use status.

Acknowledgments

We would like to take this opportunity to thank Mr Gregory Stoddard, MPH for the provision of valuable statistical advice that served to guide our sampling and stratification strategy.

Funding Statement

Research reported in this publication was partially supported by the National Institute on Drug Abuse of the National Institutes of Health under award number R03DA047577 and received further support in terms of resources and facilities from the Department of Veterans Affairs (VA) Informatics and Computing Infrastructure under award number VA HSR RES 13-457. The views expressed in this article are those of the authors and do not necessarily reflect the position or policy of the National Institutes of Health, the Department of Veterans Affairs, or the United States government.

Ethics Statement

The research reported on in this paper was approved by the University of Utah Institutional Review Board (IRB_00088382).

References

1. S. Zhu, A. Gamst, M. Lee, S. Cummins, L. Yin, and L. Zoref. The use and perception of electronic cigarettes and snus among the US population. *PLoS One*, 8(10):e79332, 2013.
2. T. Huerta, D. Walker, D. Mullen, T. Johnson, and E. Ford. Trends in e-cigarette awareness and perceived harmfulness in the US. *Am J Prev Med*, 52(3):339–346, Mar 2017.
3. C. Schoenborn and R. Gindi. Electronic cigarette use among adults: United States, 2014. Technical Report 217, National Center for Health Statistics, 2015.
4. A. McNeill, L. Brose, R. Calder, L. Bauld, and D. Robson. Evidence review of e-cigarettes and heated tobacco products 2018: A report commissioned by Public Health England. Technical report, Public Health England, 2018.
5. I Grabovac, M Oberndorfer, J Fischer, W Wiesinger, S Haider, and T E Dorner. Effectiveness of electronic cigarettes in smoking cessation: a systematic review and meta-analysis. *Nicotine Tob Res*, Sep 2020.
6. J. Gornall. Public Health England's troubled trail. *BMJ*, 351:h5826, Nov 2015.
7. American Medical Association. AMA calls for total ban on all vaping products not approved by FDA, 2019. https://www.ama-assn.org/press-center/press-releases/ama-calls-total-ban-all-vaping-products-not-approved-fda.
8. V. Krishnasamy, B. Hallowell, J. Ko, A. Board, K. Hartnett, P. Salvatore, M. Danielson, A. Kite-Powell, E. Twentyman, L. Kim, A. Cyrus, M. Wallace, P. Melstrom, B. Haag, B. A King, P. Briss, C. Jones, L. Pollack, S. Ellington, and Lung Injury Response Epidemiology/Surveillance Task Force. Update: characteristics of a nationwide outbreak of e-cigarette, or vaping, product use-associated lung injury - United States, August 2019-January 2020.

MMWR Morb Mortal Wkly Rep, 69(3):90–94, Jan 2020.

9. D. Cao, K. Aldy, S. Hsu, M. McGetrick, G. Verbeck, I. De Silva, and S. Feng. Review of health consequences of electronic cigarettes and the outbreak of electronic cigarette, or vaping, product use-associated lung injury. *J Med Toxicol*, 16(3):295–310, 07 2020.
10. M. Conway, D. Mowery, B. South, G. Stoddard, W. Chapman, O. Patterson, and S-H. Zhu. Documentation of ENDS use in the Veterans Affairs electronic health record. *American Journal of Preventive Medicine*, 56(3):474–475, 2019. [PMID: 30777165].
11. K. Young-Wolff, D. Klebaner, B. Folck, A. L Tan, R. Fogelberg, V. Sarovar, and J. Prochaska. Documentation of e-cigarette use and associations with smoking from 2012 to 2015 in an integrated healthcare delivery system. *Prev Med*, 109:113–118, Apr 2018.
12. S. Hurst and M. Conway. Exploring physician attitudes regarding electronic documentation of e-cigarette use: A qualitative study. *Tob Use Insights*, 11:1179173X18782879, 2018.
13. T. Jose, J. Hays, and D. Warner. Improved documentation of electronic cigarette use in an electronic health record. *Int J Environ Res Public Health*, 17(16), 08 2020.
14. Veterans Health Administration, 2021. https://www.va.gov/health/aboutvha.asp.
15. S. Odani, I. Agaku, C. Graffunder, M. Tynan, and B. Armour. Tobacco product use among military veterans - United States, 2010-2015. *MMWR Morb Mortal Wkly Rep*, 67(1):7–12, Jan 2018.
16. M. Cornelius, T. Wang, A. Jamal, C. Loretan, and L. Neff. Tobacco product use among adults - United States, 2019. *MMWR Morb Mortal Wkly Rep*, 69(46):1736–1742, Nov 2020.
17. P. Barnett, A. Chow, and N. Flores. Using health factors data for VA health service research. Technical Report 28, Health Economic Resource Center, 2014.
18. L. Wiley, A. Shah, H. Xu, and W. Bush. ICD-9 tobacco use codes are effective identifiers of smoking status. *J Am Med Inform Assoc*, 20(4):652–8, 2013.
19. K. McGinnis, C. Brandt, M. Skanderson, A. Justice, S. Shahrir, A. Butt, S. Brown, M. Freiberg, C. Gibert, M. Goetz, J. Kim, M. Pisani, D. Rimland, M. Rodriguez-Barradas, J. Sico, H. Tindle, and K. Crothers. Validating smoking data from the Veteran's Affairs health factors dataset, an electronic data source. *Nicotine Tob Res*, 13(12):1233–9, Dec 2011.
20. L. Da Silva, T. Ginter, T. Forbus, N. Nokes, B. Fay, T. Mikuls, G. Cannon, and S. DuVall. Extraction and quantification of pack-years and classification of smoker information in semi-structured medical records. In *Proceedings of the 28th International Conference on Machine Learning*. Bellevue, WA, 2011.
21. C. Clark, K. Good, L. Jezierny, M. Macpherson, B. Wilson, and U. Chajewska. Identifying smokers with a medical extraction system. *J Am Med Inform Assoc*, 15(1):36–9, 2008.
22. G. Savova, P. Ogren, P. Duffy, J. Buntrock, and C. Chute. Mayo clinic NLP system for patient smoking status identification. *J Am Med Inform Assoc*, 15(1):25–8, 2008.
23. T. Winden, E. Chen, Y. Wang, I. Neil Sarkar, E. Carter, and G. Melton. Towards the standardized documentation of e-cigarette use in the electronic health record for population health surveillance and research. *AMIA Jt Summits Transl Sci Proc*, 2015:199–203, 2015.
24. D. Mowery, B. South, O. Patterson, S-H Zhu, and M. Conway. Investigating the documentation of electronic cigarette use in the Veteran Affairs electronic health record: a pilot study. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing, Vancouver (BioNLP 2017)*, page 282–286. 2017.
25. K. Young-Wolff, D. Klebaner, B. Folck, L. Carter-Harris, R. Salloum, J. Prochaska, R. Fogelberg, and Andy S. Tan. Do you vape? leveraging electronic health records to assess clinician documentation of electronic nicotine delivery system use among adolescents and adults. *Prev Med*, 105:32–36, Aug 2017.
26. P. Alba, O. Patterson, and M. Conway. A decade of vaping: identifying shifts in clinical documentation and terminology in the Department of Veterans Affairs (2010-2020). In *AMIA Annu Symp Proc*, page 1684, 2020.

27. S. Duvall, R Cornia, T. Forbrush, C. Halls, and O Patterson. Check it with Chex: A validation tool for iterative NLP development. In *AMIA Annu Symp Proc*, page 204, 2014.

28. S. Duvall, O. Patterson, T. Forbrush, A. Karmauu, C. Reyes, Y. Yim, and G. Bowen. Using advanced healthcare data analytics to identify patients with advanced basal cell carcinoma in a large nationwide healthcare institution. In *45th Annual Meeting of the American College of Mohs Surgery (ACMS)*, 2014.

29. J. Carletta. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254, 1996.