Journal of the American Medical Informatics Association, 26(1), 2019, 9–18
doi: 10.1093/jamia/ocy140
Advance Access Publication Date: 13 December 2018
Research and Applications

Research and Applications

Electronic cigarette usage patterns: a case study combining survey and social media data

Yongcheng Zhan,[1] Jean-François Etter,[2] Scott Leischow,[3] and Daniel Zeng[1]

[1]Department of Management Information Systems, University of Arizona, Tucson, Arizona, USA, [2]Institute of Global Health, Faculty of Medicine, University of Geneva, Switzerland, and [3]College of Health Solutions, Arizona State University, Phoenix, Arizona, USA

Corresponding Author: Yongcheng Zhan, Department of Management Information Systems, Eller College of Management, The University of Arizona, McClelland Hall, Room 430, P.O. Box 210108, Tucson, AZ 85721–0108, USA (yongchengzhan@e-mail.arizona.edu)

Received 31 May 2018; Revised 13 August 2018; Editorial Decision 24 September 2018; Accepted 9 October 2018

ABSTRACT

Objective: To identify who were social media active e-cigarette users, to compare the use patterns from both survey and social media data for data triangulation, and to jointly use both datasets to conduct a comprehensive analysis on e-cigarette future use intentions.
Materials and Methods: We jointly used an e-cigarette use online survey ($n = 5132$) and a social media dataset. We conducted analysis from 3 different perspectives. We analyzed online forum participation patterns using survey data. We compared e-cigarette use patterns, including brand and flavor types, ratings, and purchase approaches, between the 2 datasets. We used logistic regression to study intentions to use e-cigarettes using both datasets.
Results: Male and younger e-cigarette users were the most likely to participate in e-cigarette-related discussion forums. Forum active survey participants were hardcore vapers. The e-cigarette use patterns were similar in the online survey data and the social media data. Intention to use e-cigarettes was positively related to e-liquid ratings and flavor ratings. Social media provided a valuable source of information on users' ratings of e-cigarette refill liquids.
Discussion: For hardcore vapers, social media data were consistent with online survey data, which suggests that social media may be useful to study e-cigarette use behaviors and can serve as a useful complement to online survey research. We proposed an innovative framework for social media data triangulation in public health studies.
Conclusion: We illustrated how social media data, combined with online survey data, can serve as a new and rich information source for public health research.

Key words: electronic cigarette; use pattern; survey; social media; health informatics; public health surveillance

BACKGROUND AND SIGNIFICANCE

Recent years have witnessed both the fast evolvement of e-cigarette devices and the fast growth of e-cigarette sales. The global e-cigarette market size grew from 2.76 billion U.S. dollars in 2014 to 7.1 billion U.S. dollars in 2016, according to Hexa Research.[1] The United States is the biggest market for e-cigarettes.[2] It is estimated that the global e-cigarette industry will have a total market value of 50 billion U.S. dollars by 2025.[3]

Many studies of e-cigarette use patterns relied on survey methods.[4–15] Social media, however, create a new channel of access to user-generated content and provide large datasets for e-cigarette research. For instance, Twitter text data was used in sentiment analysis[16] and marketing and use pattern recognition.[17] E-cigarettes are also discussed in online forums, where discussions are lengthier compared to tweets, thus containing much richer text information for fine-grained text mining analysis. For example, Reddit was used for analysis of e-cigarette refill liquids ("e-liquid") components and flavors,[18,19] and JuiceDB was used for e-liquids opinion analysis.[20] In addition, combining information from different social media sources may provide new insights. For example, Chu and colleagues examined the marketing strategies of leading e-cigarette brands on Twitter, Facebook, Google+, and Instagram, showing that different strategies were exploited on different websites to broadcast context-specific messages.[21] Another study utilized text data from Twitter, Reddit, and JuiceDB and found that different topic features were mentioned across different platforms.[22] However, very few studies combined survey and social media data in analyses. For example, some survey studies collected data on e-cigarette users' social media behavior,[23,24] which, however, focused on the behavior itself but did not utilize rich information from social media.

In this study, we assessed whether the user-generated content extracted from online communities could be usefully combined with online survey data to better understand the patterns of e-cigarette use. Both survey data and social media data have their limitations and biases, but observing a phenomenon from 2 different points of view provides a better grasp of reality, just as indicated from the triangulation method in social science research.[25–27] Data triangulation employs the idea of using different sources of data, including different times, places, people,[28] and collection processes,[29] to increase the validity of research results. It can also be viewed as "less a strategy for validating results and procedures than an alternative to validation which increases scope, depth and consistency in methodological proceedings."[30] Currently, much of the data triangulation in public health studies focuses on the traditional combination of qualitative data and quantitative data collected from surveys and interviews.[31–34] Very little, however, considers the possibility of integrating social media data. Therefore, in this study, we contribute to the literature by innovatively proposing a data triangulation method to include social media data in the domain of public health research.

MATERIALS AND METHODS

Data collection

Two survey datasets (survey A and survey B hereafter) were used in this study. The original surveys were designed to study e-cigarette users' profiles, utilization patterns, satisfaction, and perceived effects, as well as the relation between vaping and smoking behavior. Both of the survey data were collected from questionnaires posted on the smoking cessation website Stop-Tabac.ch from August 2011 to January 2013[35] and from October 2012 to December 2015, in English and French.[36] The survey links were published in different e-cigarette discussion forums and informing or selling e-cigarette websites. The surveys were independent of manufacturers and retailers of electronic cigarettes and e-liquids, and of the pharmaceutical and tobacco industries. The studies were originally conducted by researchers at the University of Geneva and funded by the Swiss Tobacco Prevention Fund (an agency of the Swiss Government). Since the questions from these 2 surveys were similar, we combined them in a single dataset for this study.

The main purpose of the survey was to better understand who used e-cigarettes and how they used these products. The questions covered demographic information, e-cigarette use behaviors, and smoking behaviors. Participation in the survey was voluntary. Participants who were at least 18 years old and were current e-cigarette users were eligible. Overall, there were 5132 participants from several nations including France, United States, Switzerland, United Kingdom, and Canada.

We collected social media data for the same time period (August 1, 2011, to December 31, 2015) from Reddit. As of 2017, Reddit had about 542 million monthly visitors and 234 million unique users, ranking as the fourth most visited website in the United States and the ninth in the world.[37] Reddit allows users with similar interests to form communities, which are called "subreddits." The largest subreddit for e-cigarettes is called "/r/electronic_cigarette," and has more than 150 000 subscribers. With the help of the SMILE platform (www.smileportal.org), we collected 332 906 posts from 42 major e-cigarette-related subreddits, all of which have more than 2000 subscribers. Note that all content extracted from Reddit was publicly available and thus consent was not required from the users.

DATA ANALYSIS

We do not assume that different e-cigarette users from different countries have the same use behaviors because of different local market conditions, local anti-vaping policies, and differences in smoking behaviors (eg, ratio of men/women who smoke). Therefore, to set up a more reliable comparison, we used survey data from participants living in the United States only ($n = 1057$, 21% of 5132). We cannot know where a Reddit user comes from, but a previous Reddit survey showed that 80% of users are from the United States.[38]

We tried to answer 3 research questions from the comparative analysis of online survey data and social media data.

1. (Identification) For those survey participants, who is active in social media? How is the pattern consistent with social media data?
2. (Validation) How consistent are the e-cigarette use patterns identified from social media and from surveys?
3. How can social media and survey data be used jointly to conduct a comprehensive analysis and reach a better interpretation of both types of data?

For the first research question, three survey questions were directly related to e-cigarette users' online behaviors:

- Have you ever visited a website or an online discussion forum dedicated to electronic cigarettes (labeled "Website" thereafter)?
  - Five response options: never, 1 time, 2–5 times, 6–10 times, 11 or more times.
- If you did, did these websites or these forums encourage you to use the electronic cigarette (labeled "Incite" thereafter)?
  - Four response options: not at all, not really, somewhat, a lot.
- Have you ever posted a message on a discussion forum devoted to electronic cigarettes (labeled "Forum" thereafter)?
  - Five response options: never, 1 time, 2–5 times, 6–10 times, 11 or more times.

We conducted Fisher's Exact test to test the relationship between these 3 online behaviors and age and gender (196 males and 140 females). Age was a continuous variable that had $\max = 75$, $\min = 18$, and $\text{mean} = 43$. We constructed 5 age intervals for analysis based on the histogram of the age distribution: $\text{age} < 30$ ($n = 59$), $30 \leq \text{age} < 40$ ($n = 64$), $40 \leq \text{age} < 50$ ($n = 94$), $50 \leq \text{age} < 60$ ($n = 80$), and $\text{age} \geq 60$ ($n = 25$).

Besides the demographic features, we can also infer that Reddit participants were more likely to be hardcore vapers instead of casual users. We evaluated this statement by comparing forum-active survey participants and non-forum-active survey participants on use duration, time to first puff every day, and number of puffs per day.

Thus, for these hardcore vapers, the use patterns described by the survey data and by the Reddit data were comparable. We answered the second research question by extracting relevant survey answers and corresponding Reddit text to compare use patterns including flavor types, flavor rating, nicotine level, cartridge type, purchase approach, and brand and model. Cosine similarity was normalized to the interval [0, 1] and could be used in a wide range of different data types. These advantages were utilized in previous literature to evaluate the similarity among different information sources,[39] or different medical cases.[40] To confirm the two datasets were similar, cosine similarity was measured in each feature to numerically evaluate the distance between the 2 samples.

In the survey data, based on answers to the open-ended question, "What flavor do you use the most?", we summarized flavor type distribution across 8 flavor types we identified from previous research.[19] Keywords from this previous research[19] were used to search the whole Reddit dataset and count the occurrence of flavors.

There was another flavor-related question in the survey: "Please rate the flavor you just mentioned above." The answer options were a 5-level Likert scale from "very bad" to "very good." We used a sentiment analysis tool called VADER to study flavor rating in Reddit text data. VADER is a lexicon and rule-based sentiment analysis tool and specifically attuned to sentiment expressed in social media.[41] For each of the posts containing a certain type of flavor, we fed the text into this tool and got a numerical score varied from -1 (the most negative) to $+1$ (the most positive). Then we divided the interval $[-1, 1]$ to 5 intervals $[-1, -0.8)$, $[-0.8, -0.3)$, $[-0.3, 0.3)$, $[0.3, 0.8)$, $[0.8, 1]$ in order to meet the 5-level Likert scale. We drew histograms to compare the 2 ratings. Note that these unevenly divided intervals reflected the fact that extreme ratings, for instance, "very good" or "very bad," were more difficult to obtain. We also tested evenly distributed intervals and the result did not change much.

We conducted a similar analysis concerning the nicotine level in refilled liquids. Nicotine plays a significant role in the taste and pharmacological effects of e-liquid.[18] We analyzed answers to the survey question, "What is the concentration of nicotine in the liquid or cartridge that you are currently using, on average?" and compared this information with nicotine levels self-reported in Reddit.

We focused on e-cigarette devices in the next 3 comparisons. The survey question, "Do you use prefill or refill e-cigarette?" inquired about the types of e-cigarette cartridges. E-cigarette devices have evolved in recent years.[42] At the time of the survey, prefilled e-cigarettes were more likely to be a "cigalike," which was one of the oldest versions of e-cigarettes, while refilled devices were emerging products and dominated the market later on.

We then analyzed the purchase approach of e-cigarettes based on the survey question, "Where do you usually buy your e-cigarettes?" Five response options: Internet, tobacco shop, vape shop, mall (kiosk), and pharmacy. It is noteworthy that vape shops are generally considered as adversaries of the traditional tobacco industry.[43] Most vape shops do not sell disposable or rechargeable brands that are owned by "Big Tobacco" companies, such as Blu (owned by Lorillard) and Vuse (owned by R. J. Reynolds).[43] Regular expression-based[44] keyword search was used to find corresponding materials from Reddit.

Finally, we studied e-cigarette brands and models. The survey required participants to list their most frequently used e-cigarette brands and models (open-ended responses). We used the brand and model names from the survey answers as keywords to search Reddit data for comparison. We used Levenshtein[45] distance to identify misspelling brand names (eg, "greensmoke" vs. "green smoke") in the data processing. Words with Levenshtein distance less than or equal to 2 were manually checked with the help of Google search. In the analysis, we included only keywords mentioned at least 3 times from the survey answers.

After we drew the conclusion that social media data and survey data were collected from similar samples, we could utilize the social media data in survey studies. Specifically, for our third research question, we wanted to assess whether social media data could be used to estimate opinions on e-cigarette products, so that these opinions could be utilized in further analyses. We inferred individual-level opinions on e-cigarette products from opinions collected on the social media.

A piece of e-liquid rating is related to both the overall e-liquid quality and the consumer's general satisfaction of using e-cigarettes. For example, if an e-cigarette user personally likes to use e-cigarettes, and an e-liquid brand is quite good from peer reviews, this user has a high probability of enjoying this brand of e-liquid. We constructed a Naïve Bayes classifier[46] for e-liquid rating estimation and achieved f1-score = 0.764. The details are described in Supplementary Appendix S1: Liquid Rating Estimation.

Next, we combined survey and social media data in a multivariate logistic regression analysis to study associations between e-cigarette product opinions and intention to use e-cigarettes in the future.

We assumed that consumers with positive attitudes would be more likely to continue to use e-cigarettes in the future. Based on the theory of planned behavior (TPB),[47] we tested the following hypothesis.

> *Hypothesis 1 (H1): Users with higher e-cigarette product evaluations [device (H1A) and e-juice (H1B) rating] will have higher levels of intention to use e-cigarettes in the future.*

In survey questions, the device was evaluated from 2 perspectives, model and cartridge. The e-juice was evaluated from 2 perspectives as well, flavor and liquid. Detailed survey question descriptions can be found in Supplementary Appendix S2: Regression Variables. We used a stepwise backward variable selection method and Akaike information criterion (AIC) to measure the logistic regression model quality. The result showed that only the "Liquid" and "Flavor" variables were left in the final model, which indicated that the evaluations to e-cigarette devices did not have a significant effect on the future use intention (H1A rejected). Instead, the evaluations of e-liquid and corresponding flavors were the key factors influencing the future use intention (H1B not rejected).

The regression model after the variable selection is:

$$\log\_odds(intention) = \beta_0 + \beta_1 Liquid + \beta_2 Flavor + \epsilon \quad (1)$$

Survey B had all the variables needed to conduct the regression analysis. However, survey A did not cover e-liquid ratings. Thus, we tested the validity of our approach by substituting the estimated e-liquid rating for actual e-liquid rating in the survey B dataset (only 240 out of 363 records listed e-liquid brands), using the regression:

$$\log\_odds(intention) = \beta_0 + \beta_1 LiquidEstimate + \beta_2 Flavor + \epsilon \quad (2)$$

Finally, we applied the trained classifier to the survey A dataset using regression model (2) to test the usefulness of our e-liquid rating estimator.

Note that our dataset was unbalanced regarding the e-cigarette future use intention. Almost all participants stated that they had future use intention (54 vs. 2 in survey A and 363 vs. 7 in survey B).

Since logistic regression can sharply underestimate the probability of rare events,[46] given the limitation of the dataset, we applied rare-event logistic regression instead, which conducted prior correction or data weighing compensation in the maximum likelihood estimation process. An R package "Zelig" was used to conduct the analysis.[48,49] For robustness check, using normal logistic regression would not change the statistical significance level of the analysis.

RESULTS

Demographic features and online activities

The test results are summarized in Figure 1 (Gender) and Figure 2 (Age).

As shown in Figure 1, males were more prone to visit e-cigarette online discussion forums and posted more messages than females ($P < .001$ in Fisher's Exact test). This finding is consistent with the social media data. Although we cannot obtain user gender information from Reddit directly, by using keywords husband/boyfriend and wife/girlfriend as a rough estimation, we found that there were 2432 male Reddit users and 479 female Reddit users. We admitted the fact that the LGBTQ population might bias the result. But given the large number gap we identified (2432 vs. 479), the actual gender ratio would not change much had we had the ability to fully consider the LGBTQ population. The approximated result served as a reference evidence of the consistency of survey data and social media data.

However, there was no significant difference between men and women, whether these websites encouraged them to use e-cigarettes ($P = 1.00$ in Fisher's Exact test).

Similarly, Figure 2 showed that different age groups had different forum visit and message posting behavior ($P = .023$, .007 in Fisher's Exact test). However, there was no significant differences among age groups for website e-cigarette use encouragement ($P = .60$ in Fisher's Exact test). We found that younger users participated more in online forums than older users. This is consistent with previous research findings that social media users are younger.[37]

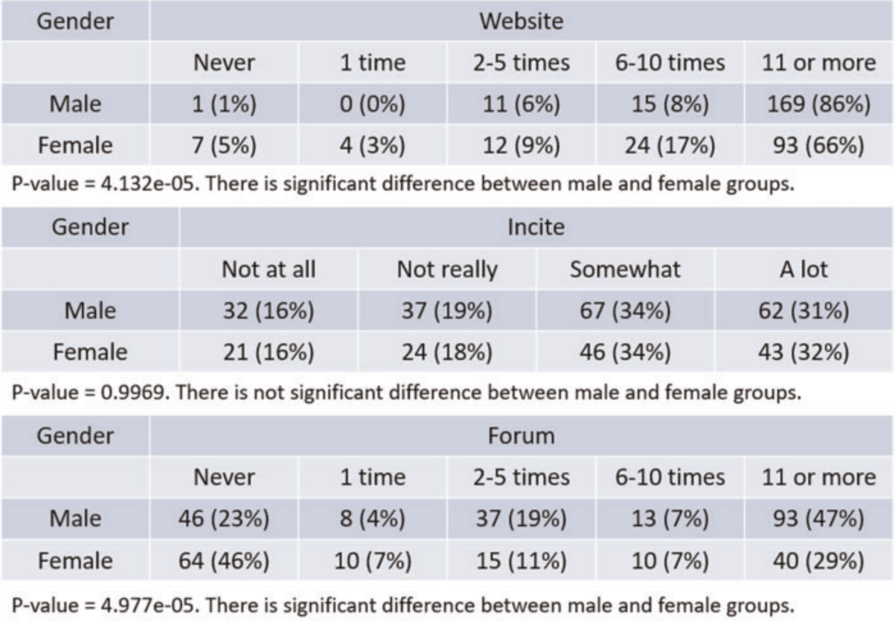
Next, we evaluated whether forum-active survey participants were hardcore vapers instead of casual users. The forum-active participants had used e-cigarettes for 209 days on average (sd = 274.48) at the time when they took the survey, while non-forum-active participants used for only 104 days (sd = 197.79). Forum-active users waited for 33 minutes on average (sd = 51.05) after waking up to vape the first puff while non-forum-active users waited for 56 minutes (sd = 109.52). Forum-active users usually drew 207 puffs per day on average (sd = 152.26) while non-forum-active users drew 110 puffs per day (sd = 144.23). We used an independent 2-sample t test to compare the results and found all $P < .001$, which indicated that e-cigarette users who were active in forums were statistically heavier users comparing to those who were not active in forums.

Thus, for these hardcore vapers, the use patterns described by the survey data and by the Reddit data were comparable. We extracted relevant survey answers and corresponding Reddit text to compare use patterns including flavor types, flavor rating, nicotine level, cartridge type, purchase approach, and brand and model.

USE PATTERN COMPARISON

The flavor type comparison results are summarized in Table 1. The cosine similarity of the two distributions is 0.991, which indicated a high consistency. Tobacco is the most welcomed flavor and counted for around 45% of the total flavors mentioned. Note that "RY4" is the most famous of them. If we did not use this keyword in Reddit text extraction, we could only obtain 676 (47% of 1430) posts mentioning tobacco flavor. This information cannot simply be obtained from the survey data, but using Reddit data gained more insights.

The flavor rating comparison histogram is shown in Figure 3A. Overall, the 2 datasets shared a similar distribution across flavors. We obtained a 0.986 cosine similarity. Almost 60% of the flavor ratings were "very good," which indicated an overall satisfaction to this emerging product. We also analyzed the rating across different flavor categories using Reddit data. Figure 3B shows a consistent flavor rating pattern across these categories.



Gender	Website				
	Never	1 time	2-5 times	6-10 times	11 or more
Male	1 (1%)	0 (0%)	11 (6%)	15 (8%)	169 (86%)
Female	7 (5%)	4 (3%)	12 (9%)	24 (17%)	93 (66%)

P-value = 4.132e-05. There is significant difference between male and female groups.

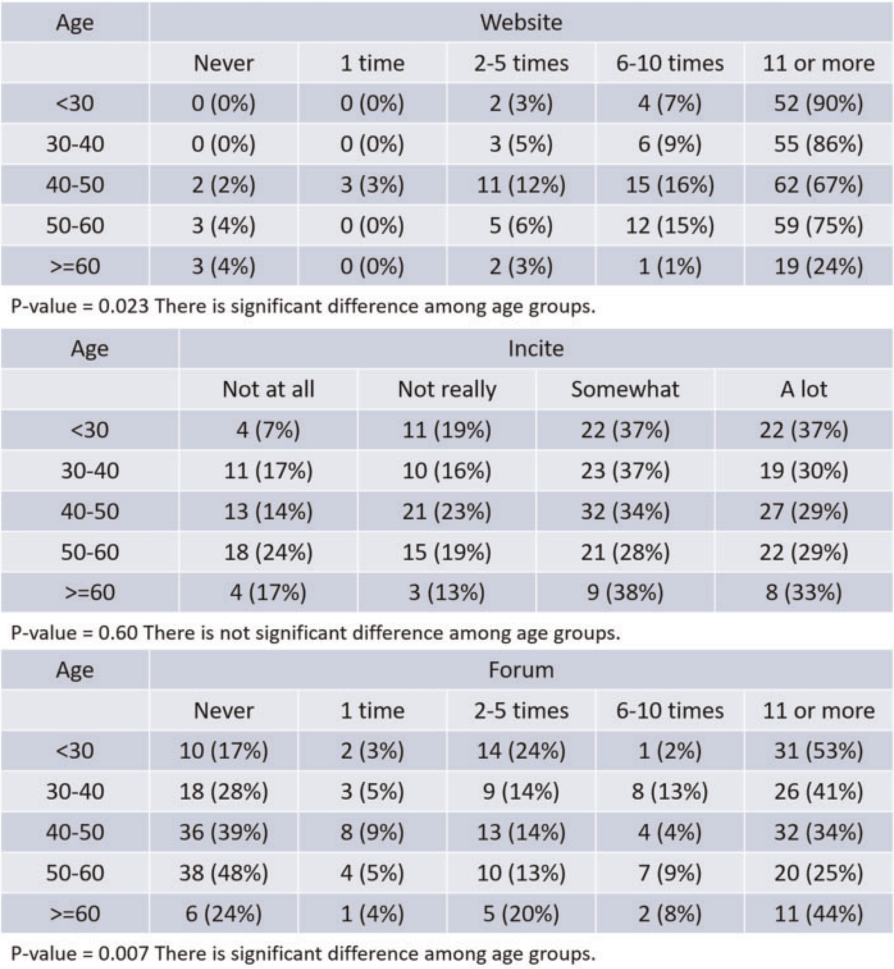
Gender	Incite			
	Not at all	Not really	Somewhat	A lot
Male	32 (16%)	37 (19%)	67 (34%)	62 (31%)
Female	21 (16%)	24 (18%)	46 (34%)	43 (32%)

P-value = 0.9969. There is not significant difference between male and female groups.

Gender	Forum				
	Never	1 time	2-5 times	6-10 times	11 or more
Male	46 (23%)	8 (4%)	37 (19%)	13 (7%)	93 (47%)
Female	64 (46%)	10 (7%)	15 (11%)	10 (7%)	40 (29%)

P-value = 4.977e-05. There is significant difference between male and female groups.

Figure 1. Fisher's Exact test for gender and online behaviors.



Age	Website				
	Never	1 time	2-5 times	6-10 times	11 or more
<30	0 (0%)	0 (0%)	2 (3%)	4 (7%)	52 (90%)
30-40	0 (0%)	0 (0%)	3 (5%)	6 (9%)	55 (86%)
40-50	2 (2%)	3 (3%)	11 (12%)	15 (16%)	62 (67%)
50-60	3 (4%)	0 (0%)	5 (6%)	12 (15%)	59 (75%)
>=60	3 (4%)	0 (0%)	2 (3%)	1 (1%)	19 (24%)

P-value = 0.023 There is significant difference among age groups.

Age	Incite			
	Not at all	Not really	Somewhat	A lot
<30	4 (7%)	11 (19%)	22 (37%)	22 (37%)
30-40	11 (17%)	10 (16%)	23 (37%)	19 (30%)
40-50	13 (14%)	21 (23%)	32 (34%)	27 (29%)
50-60	18 (24%)	15 (19%)	21 (28%)	22 (29%)
>=60	4 (17%)	3 (13%)	9 (38%)	8 (33%)

P-value = 0.60 There is not significant difference among age groups.

Age	Forum				
	Never	1 time	2-5 times	6-10 times	11 or more
<30	10 (17%)	2 (3%)	14 (24%)	1 (2%)	31 (53%)
30-40	18 (28%)	3 (5%)	9 (14%)	8 (13%)	26 (41%)
40-50	36 (39%)	8 (9%)	13 (14%)	4 (4%)	32 (34%)
50-60	38 (48%)	4 (5%)	10 (13%)	7 (9%)	20 (25%)
>=60	6 (24%)	1 (4%)	5 (20%)	2 (8%)	11 (44%)

P-value = 0.007 There is significant difference among age groups.

Figure 2. Fisher's Exact test for age and online behaviors.

Table 1. Flavor categories

Flavor	Survey data	Reddit data
Tobacco	198 (42%)	1430 (45%)
Fruit	112 (24%)	676 (21%)
Menthol	63 (13%)	263 (8%)
Beverages	26 (5%)	230 (7%)
Cream	35 (7%)	285 (9%)
Sweet	31 (7%)	172 (5%)
Seasonings	7 (1%)	89 (3%)
Nuts	2 (0%)	7 (0%)

A nicotine level comparison histogram is shown in Figure 4. We observed a similar pattern of nicotine level between survey and Reddit data. The average nicotine level for survey data was 16 mg/mL (sd = 9.62) while the average nicotine level for Reddit data was 16 mg/mL as well (sd = 8.34). The 2-sample independent t test cannot distinguish the mean difference of these 2 datasets ($P = 0.95$). We also obtained a cosine similarity 0.965.

For the cartridge type comparison, we observed a similar ratio of prefilled and refilled cartridge use for the survey (28 vs. 265) and corresponding Reddit dataset (61 vs. 374) ($P = .083$ in Fisher's Exact test). There were approximately 10 times more users of refillable cartridges than prefilled "cigalike" e-cigarettes. The cosine similarity for this attribute between the 2 datasets is 0.998.

The purchase approach results in Table 2 show that the Internet was the most popular channel to purchase e-cigarettes. The result showed consistency between the 2 datasets with a cosine similarity 0.999. Almost 90% of the purchase happened on the Internet.

The results of the brand comparison are listed in Table 3. The patterns indicate that survey data and Reddit data had a similar distribution of popular e-cigarette brands and models. "Ego," "joye," "provape," "volcano," and "kanger" were identified as the most used products in both datasets. The cosine similarity of the 2 brand lists is 0.852.

We compared use patterns in survey and social media data from 6 different angles: flavor types, flavor rating, nicotine level, cartridge type, purchase approach, and brands. We summarized cosine similarity of these 6 different features and got 0.965 on average with sd = 0.057. Based on the Chebyshev's inequality and the 3-sigma-rule, 0.794 was considered as the lower bound for 89% (99% if normally distributed, eg, approximated by the central limit theorem) of the observed cosine similarity values, which was a reasonably high value to state the 2 datasets were similar.

Regression analysis on intentions to use e-cigarettes

First, we used the regression model (1) on the survey B dataset. The effect of each variable to the odds ratio and corresponding 95%

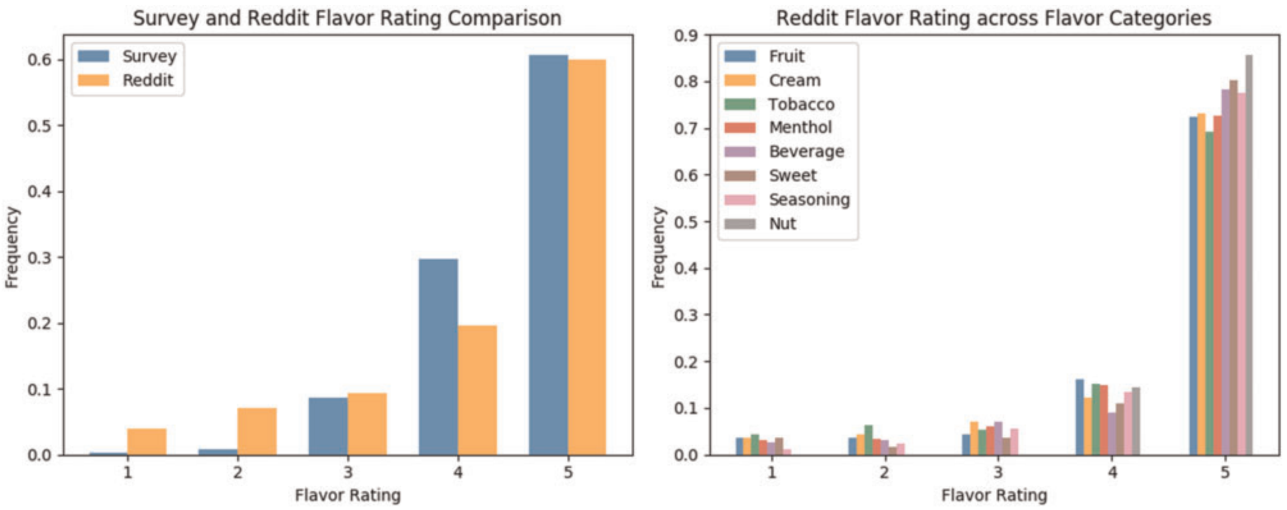


Figure 3. Flavor rating. 3A. Survey and Reddit comparison. 3B. Reddit flavor rating across flavor categories.

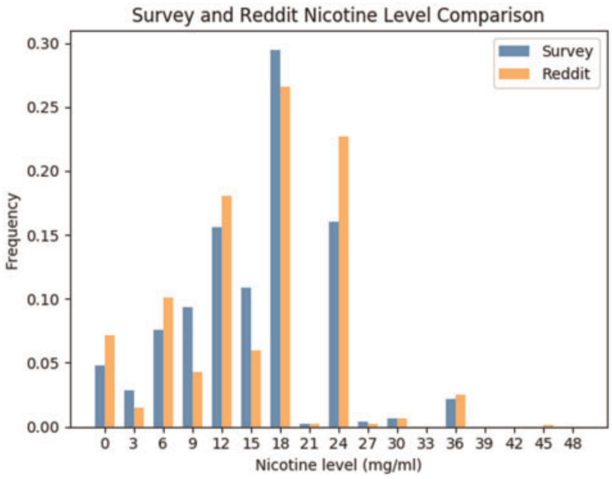


Figure 4. Nicotine level.

Table 2. Purchase approach

Approach	Survey	Reddit
Internet	205 (93%)	101 (89%)
Tobacco shop	6 (3%)	3 (3%)
Vape shop	8 (4%)	5 (4%)
Mall, kiosk	2 (1%)	3 (3%)
Pharmacy	0 (0%)	1 (1%)

confidence interval is shown in Table 4, column (1). One point more of e-liquid rating is associated with 6.69 increase of e-cigarette future use intention odds ratio ($P = .001$). One point more of flavor rating is associated with 2.61 increase of e-cigarette future use intention odds ratio ($P = .034$).

By using a regression model (2) on the survey B dataset, we got results in Table 4, column (2). The similar results between column (1) and column (2) showed that the substitution does not change the regression significance level or even the estimated coefficients. It means that the individual e-cigarette product evaluation can be estimated by jointly using the product quality from social media and the individual's general satisfaction from survey questions. Sentiment analysis from online community discussion reveals the overall quality of products.

Table 3. E-cigarette Brands

Brand/Model	Survey	Reddit
Ego	112 (30.1%)	1516 (39.9%)
Joye	98 (26.3%)	286 (7.5%)
Provape (provari)	22 (5.9%)	485 (12.8%)
Volcano	14 (3.8%)	147 (3.9%)
Kanger	13 (3.5%)	363 (9.6%)
Smokeless Image	10 (2.7%)	99 (2.6%)
Blu	9 (2.4%)	158 (4.2%)
Lavatube	8 (2.2%)	216 (5.7%)
Apollo	8 (2.2%)	48 (1.3%)
Kgo	7 (1.9%)	54 (1.4%)
Madvapes	7 (1.9%)	92 (2.4%)
Silver Bullet	7 (1.9%)	97 (2.6%)
Puresmoker	6 (1.6%)	14 (0.4%)
Altsmoke	6 (1.6%)	53 (1.4%)
Vapor 9	6 (1.6%)	21 (0.6%)
V4l	5 (1.3%)	21 (0.6%)
Slb	5 (1.3%)	12 (0.3%)
Vapage	5 (1.3%)	9 (0.2%)
E Power	5 (1.3%)	8 (0.2%)
Indulgence	4 (1.1%)	3 (0.1%)
Vmod	3 (0.8%)	31 (0.8%)
Super-t manufacturer	3 (0.8%)	4 (0.1%)
Njoy	3 (0.8%)	40 (1.1%)
Prodigy	3 (0.8%)	9 (0.2%)
Smoktek	3 (0.8%)	15 (0.4%)

Finally, we used the regression model (2) on the survey A dataset. We obtained the result as shown in Table 4, column (3). For survey A, 1 point of e-liquid rating increase was associated with 10.40 increase of e-cigarette future use intention odds ratio ($P = .434$). One point flavor rating increase was related to 4.03 increase of e-cigarette future use intention odds ratio ($P = .427$). Though we did not obtain a statistical significant estimation, the coefficients were still consistent with the results obtained from survey B. The unbalanced

Table 4. Logistic regression on future use intention

	(1) Intention (survey B)	(2) Intention (survey B)	(3) Intention (survey A)
Liquid	6.69 (2.11, 21.26) **		
Estimated liquid		6.73 (2.03, 22.27) **	10.40 (0.03, 3699.72)
Flavor	2.61 (1.07, 6.36) *	2.86 (1.31, 6.27) **	4.03 (0.13, 125.80)
AIC	52.97	55.79	6.69
McFadden's R^2	0.32	0.28	0.96

P < 0.001 ***, P < 0.01 **, P < 0.05 *.

dataset (2 no clear intention vs. 54 having intention) might be the reason for this situation. Nonetheless, the proposed method has shown its potential to be explored by future studies.

DISCUSSION

Contributions

Previous studies utilized either survey data or social media data but seldom both to analyze e-cigarette use patterns. The combination of online survey data and social media data sheds light on new approaches to study this field.

From the methodological perspective, we contributed to the literature by innovatively proposing a new method in integrating social media data in data triangulation. We first identified survey participants who were active on social media and used data from these users to compare the use patterns between the two datasets. The cosine measure was employed to display the similarity of the datasets and confirmed the consistency of the samples. Then we used information from the online community to infer survey participants' opinions and further used the opinions in intention prediction. The consistent logistic regression results showed potential to estimate individual's attributes by integrating social media data.

From a practical sense, first, we found that males/youngers were more likely to visit e-cigarette online discussion forums and tended to post more messages than females/elders. Furthermore, these forum-active users were identified as hardcore vapers who had vaped for a longer time, took less time to have the first puff after waking up, and puffed more times per day. This conclusion was supported by the social media dataset. Second, the 2 datasets shared similar patterns across flavor type, flavor rating, nicotine level, cartridge type, purchase approach, and brand and model. We obtained a 0.965 cosine similarity on average. Third, by collecting opinions on e-liquid brand from social media, we built a Naïve Bayes classifier and achieved 0.764 f1-score, which could be used to infer survey participant's e-liquid rating. The intentions to use e-cigarettes in the future were positively related to e-liquid ratings and flavor ratings.

By adopting more accurate and advanced methods, as well as applying features extracted from surveys, we expect social media data to be used more frequently and interactively in e-cigarette health surveillance. We propose a research framework that utilizes survey and social media data interactively, which is shown in Figure 5. In this framework, as we have identified, hardcore vapers are more likely to be involved in both the survey datasets and social media datasets. They are the target population to be studied in this mixed-method approach. Survey questions are designed to collect basic demographic features and e-cigarette use patterns. Then the corresponding social media dataset is collected to gain further individual opinions and online community information. Finally, by applying the features to analytical models, we could conduct health surveillance, such as pattern recognition, trend detection, and intention inference.

LIMITATIONS

The first limitation is the different population composition of the 2 datasets. Our research framework is suitable for research to hardcore vapers. Casual e-cigarette users, however, were much less likely to visit specific e-cigarette forums, and thus information extracted from forums was less relevant. This limits our study approach to be applied to the whole e-cigarette user population. Though understanding the use patterns of heavy users can help depict the e-cigarette use health consequences, studying casual users and the reasons that they started using e-cigarettes might be another important research domain to investigate to stop improper e-cigarette initiation.

Another limitation is the limited sample size for regression. Although we had 1057 survey participants, not all of them reported e-liquid brand and rating, flavor rating, or future e-cigarette use intention. The unbalanced dataset might cause biases in the regression results.

FUTURE RESEARCH

We propose several possible approaches for future studies. First, a better-designed survey based on theory (eg, the TPB) could be employed to explore the relationship between e-cigarette use intentions and attitudes, subjective norms, and perceived behavior control, with the help of social media data.

Second, more social media platforms could be incorporated. Data from Twitter, Facebook, YouTube, Instagram, and many other social media platforms can generate new perspectives and shed light on innovative approaches to combine the survey data and social media data. For example, Twitter has location information of e-cigarette users. This location information could be utilized in a temporal–spatial analysis. Note that the geospatial data can be obtained only if the user actively chooses to tag the tweets. The selection bias needs to be thoroughly considered as a future work direction.

Third, more fine-grained qualitative and quantitative methods could be applied to investigate the current dataset. For example, the reason that some of the brands were popular could be studied from both the open-ended survey questions and social media text.

Finally, using a unique identifier, survey data and social media data could be directly linked at individual level in volunteers. This can help relieve the limitation that our approach can be applied only to hardcore vapers.

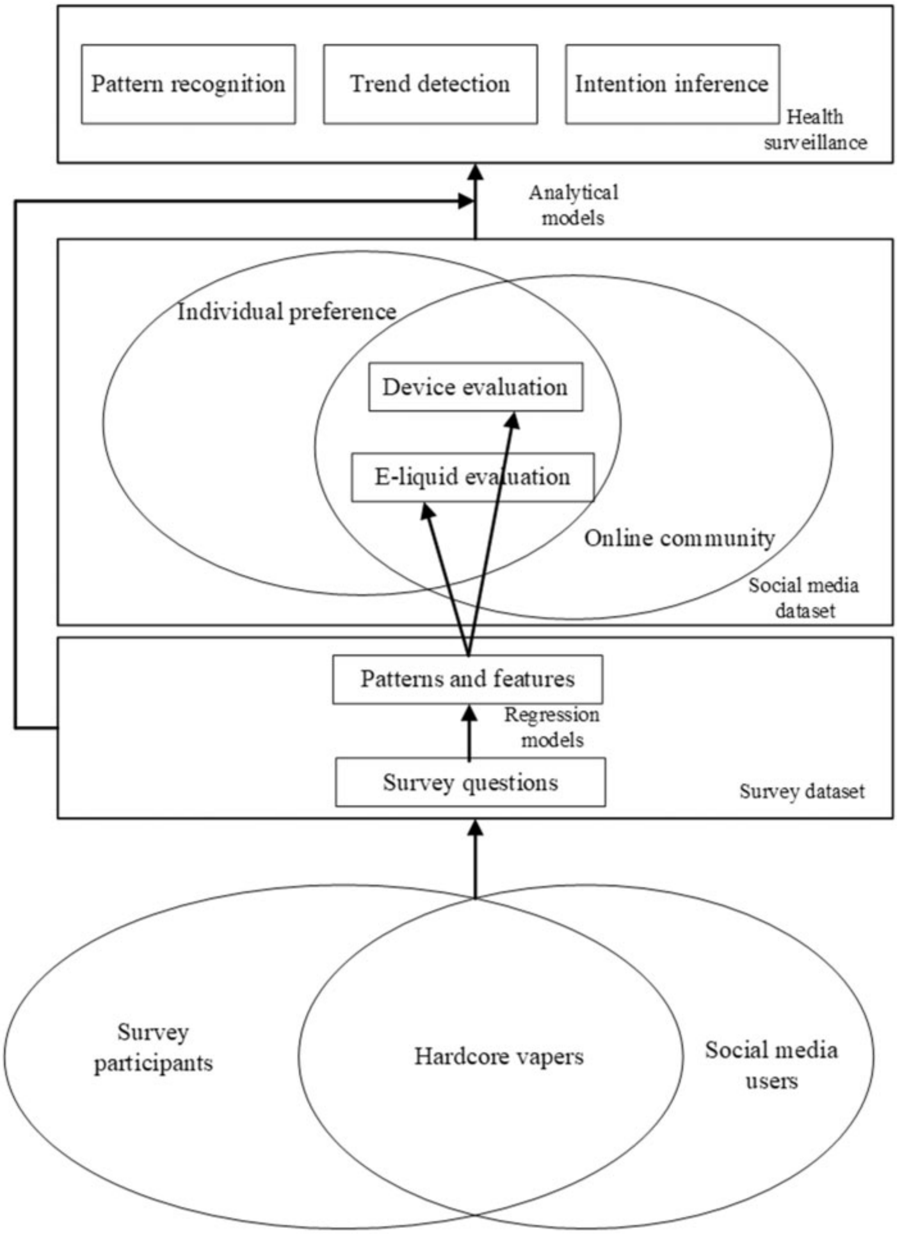


Figure 5. Framework for combining survey dataset and social media dataset in e-cigarette health surveillance studies.

CONCLUSION

We studied e-cigarette use patterns by combining survey data and corresponding social media data. We found e-cigarette users who were active online were more likely to be hardcore vapers. The use patterns identified from the corresponding survey dataset and social media dataset shared commonalities. The future use intention was studied by combining the 2 datasets and found to be related to e-liquid ratings and flavor ratings. We hope this study can serve as an example of social media data triangulation study in the public health surveillance research community and be utilized by other researchers and policymakers.

FUNDING STATEMENT

This work was supported by the NIH Grant 5R01DA037378-05, NNSFC Innovative Team Grant 71621002, and CAS Grant ZDRW-XH-2017-3.

CONTRIBUTORS

Yongcheng Zhan (YZ), Jean-François Etter (JFE), Scott Leischow (SL), and Daniel Zeng (DZ) conceived the idea for this study. YZ and JFE designed the study, conducted the data analysis, and drafted the manuscript. JFE, SL, and DZ provided critical feedback, helped interpret the analysis of results, and revised the manuscript. All authors read and approved the final manuscript.

Conflict of interest statement. None declared.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

REFERENCES

1. Hexa Research. E-cigarette Market Analysis By Product (Disposable, Rechargeable, Modular), And Segment Forecasts, 2014–2024. 2017. https://www.hexaresearch.com/research-report/e-cigarette-market. Accessed November 6, 2017.
2. Kaplan J. The U.S. E-Cigarette Market Is the Biggest in the World: Chart. 2016. https://www.bloomberg.com/news/articles/2016–06–20/the-u-s-e-cigarette-market-is-the-biggest-in-the-world-chart. Accessed November 6, 2017.
3. BIS Research. Electronic Cigarette & E Vapor (Vaporizer) Market Research Reports. 2016. https://bisresearch.com/industry-report/electronic-cigarette-market-size-forecast.html. Accessed November 6, 2017.
4. McCabe SE, Veliz P, McCabe VV, Boyd CJ. Smoking behaviors and intentions among current e-cigarette users, cigarette smokers, and dual users: a national survey of U.S. high school seniors. *Prev Med (Baltim)* 2017; 99: 228–35.
5. Volesky KD, Maki A, Scherf C, Watson LM, Cassol E, Villeneuve PJ. Characteristics of e-cigarette users and their perceptions of the benefits, harms and risks of e-cigarette use: survey results from a convenience sample in Ottawa, Canada. *Health Promot Chronic Dis Prev Can* 2016; 36 (7): 130–8.
6. Regan AK, Promoff G, Dube SR, Arrazola R. Electronic nicotine delivery systems: adult use and awareness of the “e-cigarette” in the USA. *Tob Control* 2013; 22 (1): 19–23.
7. Kralikova E, Novak J, West O, Kmetova A, Hajek P. Do e-cigarettes have the potential to compete with conventional cigarettes? A survey of conventional cigarette smokers’ experiences with e-cigarettes. *Chest J* 2013; 144 (5): 1609–14.
8. McCabe SE, West BT, Veliz P, Boyd CJ. E-cigarette use, cigarette smoking, dual use, and problem behaviors among U.S. adolescents: results from a National Survey. *J Adolesc Heal* 2017; 61 (2): 155–62.
9. Farsalinos KE, Romagna G, Tsiapras D, Kyrzopoulos S, Voudris V. Characteristics, perceived side effects and benefits of electronic cigarette use: a worldwide survey of more than 19,000 consumers. *Int J Environ Res Public Health* 2014; 11 (4): 4356–73.
10. Brown J, West R, Beard E, Michie S, Shahab L, McNeill A. Prevalence and characteristics of e-cigarette users in Great Britain: Findings from a general population survey of smokers. *Addict Behav* 2014; 39 (6): 1120–5.
11. Dawkins L, Turner J, Roberts A, Soar K. Vaping” profiles and preferences: an online survey of electronic cigarette users. *Addiction* 2013; 108 (6): 1115–25.
12. Etter J-F. Electronic cigarettes: a survey of users. *BMC Public Health* 2010; 10: 231.
13. Pearson JL, Richardson A, Niaura RS, Vallone DM, Abrams DB. E-cigarette awareness, use, and harm perceptions in US adults. *Am J Public Health* 2012; 102 (9): 1758–66.
14. Giovenco DP, Lewis MJ, Delnevo CD. Factors associated with e-cigarette use: a national population survey of current and former smokers. *Am J Prev Med* 2014; 47 (4): 476–80.
15. Adkison SE, O’Connor RJ, Bansal-Travers M, *et al*. Electronic nicotine delivery systems: international tobacco control four-country survey. *Am J Prev Med* 2013; 44 (3): 207–15.
16. Cole-Lewis H, Pugatch J, Sanders A, *et al*. Social listening: a content analysis of E-cigarette discussions on Twitter. Eysenbach G, ed. *J Med Internet Res* 2015; 17 (10): e243.
17. Kim AE, Hopper T, Simpson S, *et al*. Using Twitter data to gain insights into E-cigarette marketing and locations of use: an infoveillance study. *J Med Internet Res* 2015; 17 (11): e251.
18. Li Q, Zhan Y, Wang L, Leischow SJ, Zeng DD. Analysis of symptoms and their potential associations with e-liquids’ components: a social media study. *BMC Public Health* 2016; 16 (1): 674.
19. 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 Public Health* 2015; 12 (11): 14916–35.
20. Chen Z, Zeng DD. Mining online e-liquid reviews for opinion polarities about e-liquid features. *BMC Public Health* 2017; 17 (1): 633.
21. Chu K-H, Sidhu AK, Valente TW. Electronic cigarette marketing online: a multi-site, multi-product comparison. *JMIR Public Health Surveill* 2015; 1 (2): e11.
22. Zhan Y, Liu R, Li Q, Leischow SJ, Zeng DD. Identifying topics for E-cigarette user-generated contents: a case study from multiple social media platforms. Eysenbach G, ed. *J Med Internet Res* 2017; 19 (1): e24.
23. Sawdey, M. D., Hancock, L., Messner, M., & Prom-Wormley, E. C. (2017). Assessing the Association Between E-Cigarette Use and Exposure to Social Media in College Students: A Cross-Sectional Study. Substance use & misuse, 52(14), 1910–1917.
24. Link AR, Cawkwell PB, Shelley DR, Sherman SE. An exploration of online behaviors and social media use among hookah and electronic-cigarette users. *Addict Behav Reports* 2015; 2: 37–40.
25. Hussein A. The use of triangulation in social sciences research: can qualitative and quantitative methods be combined? *J Comp Soc Work* 2009; 1: 1–12.
26. Denzin NK. Triangulation 2.0. *J Mix Methods Res* 2012; 6 (2): 80–8.
27. Carter N, Bryant-Lukosius D, DiCenso A, Blythe J, Neville AJ. The use of triangulation in qualitative research. *Oncol Nurs Forum* 2014; 41: 545–7.
28. Denzin NK. *The research act: a theoretical orientation to sociological methods*. New York, NY: McGraw-Hill. 1978.
29. Begley CM. Using triangulation in nursing research. *J Adv Nurs* 1996; 24 (1): 122–8.
30. Flick U. *An Introduction to Qualitative Research*. Thousand Oaks, California: Sage. 2014.
31. Bossuyt N, Van Casteren V, Goderis G, *et al*. Public health triangulation to inform decision-making in Belgium. *Stud Health Technol Inform*. 2015; 210: 855–859.
32. O’Cathain A, Knowles E, Bishop-Edwards L, *et al*. Understanding variation in ambulance service non-conveyance rates: a mixed methods study. *Health Serv Deliv Res* 2018; 6 (19): 1–192.
33. Rutherford GW, McFarland W, Spindler H, *et al*. Public health triangulation: approach and application to synthesizing data to understand national and local HIV epidemics. *BMC Public Health* 2010; 10 (1): 447.
34. Johnson M, O’Hara R, Hirst E, *et al*. Multiple triangulation and collaborative research using qualitative methods to explore decision making in pre-hospital emergency care. *BMC Med Res Methodol* 2017; 17 (1): 11.
35. Etter J-F, Bullen C. Electronic cigarette: users profile, utilization, satisfaction and perceived efficacy. *Addiction* 2011; 106 (11): 2017–28.
36. Etter J-F. Electronic cigarette: a longitudinal study of regular vapers. *Nicotine Tob Res* 2018; 20 (8): 912–22.
37. Alexa. Reddit.com Traffic Statistics. 2017. https://www.alexa.com/siteinfo/reddit.com. Accessed November 8, 2017.
38. slackerChuck. [SURVEY] The Results Are In! 2014. https://www.reddit.com/r/electronic_cigarette/comments/2jigna/survey_the_results_are_in/. Accessed March 6, 2018.
39. Chaudhuri S, Le T, White C, Thompson H, Demiris G. Examining health information–seeking behaviors of older adults. *Comput Inform Nurs* 2013; 31 (11): 547.
40. Begum S, Ahmed MU, Funk P, Xiong N, Schéele B. Similarity of medical cases in health care using cosine similarity and ontology. In. David C. Wilson and Deepak Khemani. eds. 5th Workshop on CBR in the Health Sciences, ICCBR-07, Northern Ireland: Springer LNCS; 2007: 263–272.
41. Hutto CJ, Gilbert E. Vader: A parsimonious rule-based model for sentiment analysis of social media text. Eighth Int AAAI Conf Weblogs…. 2014: 216–225. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8109%5Cnhttp://comp.social.gatech.edu/papers/icwsm14.vader.hutto.pdf. Accessed March 6, 2018.
42. Farsalinos KE, Spyrou A, Tsimopoulou K, Stefopoulos C, Romagna G, Voudris V. Nicotine absorption from electronic cigarette use: comparison between first and new-generation devices. *Sci Rep* 2014; 4: 4133.
43. Kamerow D. The battle between big tobacco and vape shops. *BMJ* 2014; 349: g5810.
44. Thompson K. Programming techniques: regular expression search algorithm. *Commun ACM* 1968; 11 (6): 419–22.

45. Levenshtein VI. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady* 1966; 10: 707–10.
46. King G, Zeng L. Logistic regression in rare events data. *Polit Anal* 2001; 9 (02): 137–63.
47. Ajzen I. The theory of planned behavior. *Organ Behav Hum Decis Process* 1991; 50 (2): 179–211.
48. Imai K, King G, Lau O. Toward a common framework for statistical analysis and development. *J Comput Graph Stat* 2008; 17 (4): 892–913.
49. Choirat C, Honaker J, Imai K, King G, Lau O. *Zelig: Everyone's Statistical Software*. Version 5.0–15, 2017. 2017. http://zeligproject.org. Accessed March 6, 2018.