Published in final edited form as:

J Health Commun. 2019 ; 24(12): 889–899. doi:10.1080/10810730.2019.1682724.

Combining crowd-sourcing and automated content methods to improve estimates of overall media coverage: Theme mentions in e-cigarette and other tobacco coverage

Laura A. Gibson[1],
Annenberg School for Communication, University of Pennsylvania, Philadelphia, PA, US

Leeann Siegel,
Annenberg School for Communication, University of Pennsylvania, Philadelphia, PA, US

Elissa Kranzler[2],
Annenberg School for Communication, University of Pennsylvania, Philadelphia, PA, US

Allyson Volinsky,
Annenberg School for Communication, University of Pennsylvania, Philadelphia, PA, US

Matthew B. O'Donnell,
Annenberg School for Communication, University of Pennsylvania, Philadelphia, PA, US

Sharon Williams,
Annenberg School for Communication, University of Pennsylvania, Philadelphia, PA, US

Qinghua Yang[3],
Annenberg School for Communication, University of Pennsylvania, Philadelphia, PA, US

Yoonsang Kim,
NORC at the University of Chicago, Chicago, IL, US

Steven Binns,
NORC at the University of Chicago, Chicago, IL, US

Hy Tran,
NORC at the University of Chicago, Chicago, IL, US

Veronica Maidel Epstein[4],
NORC at the University of Chicago, Chicago, IL, US

Timothy Leffel,
NORC at the University of Chicago, Chicago, IL, US

Michelle Jeong[5],
Annenberg School for Communication, University of Pennsylvania, Philadelphia, PA, US

[1]Current affiliation: Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, US
[2]Current affiliation: The Wharton School, University of Pennsylvania, Philadelphia, PA, US
[3]Current affiliation: School of Communication, Texas Christian University, Fort Worth, TX, US
[4]Current affiliation: EarlySense, Ramat Gan, Israel
[5]Current affiliation: School of Public Health, Rutgers University, New Brunswick, NJ, US
[6]Current affiliation: Communication Studies, University of Georgia, Athens, GA, US
[7]Current affiliation: Harvard T.H. Chan School of Public Health, Harvard University, Brookline, MA, US

Jiaying Liu[6],
Annenberg School for Communication, University of Pennsylvania, Philadelphia, PA, US

Stella Lee[7],
Annenberg School for Communication, University of Pennsylvania, Philadelphia, PA, US

Sherry Emery,
NORC at the University of Chicago, Chicago, IL, US

Robert C. Hornik
Annenberg School for Communication, University of Pennsylvania, Philadelphia, PA, US

Abstract

Exposure to media content can shape public opinions about tobacco. Accurately describing content is a first step to showing such effects. Historically, content analyses have hand-coded tobacco-focused texts from a few media sources which ignored passing mention coverage and social media sources, and could not reliably capture over-time variation. By using a combination of crowd-sourced and automated coding, we labeled the population of all e-cigarette and other tobacco-related (including cigarettes, hookah, cigars, etc.) 'long-form texts' (focused and passing coverage, in mass media and website articles) and social media items (tweets and YouTube videos) collected May 2014-June 2017 for four tobacco control themes. Automated coding of theme coverage met thresholds for item-level precision and recall, event validation, and weekly-level reliability for most sources, except YouTube. Health, Policy, Addiction and Youth themes were frequent in e-cigarette long-form focused coverage (44%−68%), but not in long-form passing coverage (5%−22%). These themes were less frequent in other tobacco coverage (long-form focused (13–32%) and passing coverage (4–11%)). Themes were infrequent in both e-cigarette (1–3%) and other tobacco tweets (2–4%). Findings demonstrate that passing e-cigarette and other tobacco long-form coverage and social media sources paint different pictures of theme coverage than focused long-form coverage. Automated coding also allowed us to code the amount of data required to estimate reliable weekly theme coverage over three years. E-cigarette theme coverage showed much more week-to-week variation than did other tobacco coverage. Automated coding allows accurate descriptions of theme coverage in passing mentions, social media, and trends in weekly theme coverage.

Exposure to media content shapes public perceptions and opinions of public health issues (Hornik, 2002). Researchers have used content analyses to study tobacco media coverage's impact on tobacco use and beliefs (Dunlop, Cotter, Perez, & Chapman, 2011; Dunlop & Romer, 2010; Myers et al., 2018; Smith et al., 2008). Accurately describing content is a first step to showing such effects. Prior content analyses have described samples of tobacco-related media coverage (cf. Menashe & Siegel, 1998; D. Nelson et al., 2015; Smith et al., 2008; Smith, Wakefield, & Edsall, 2006; Wackowski, Lewis, Delnevo, & Ling, 2013; Wakefield, Brennan, Durkin, McLeod, & Smith, 2011). Invariably, these studies found that content favored tobacco control (U.S. Department of Health and Human Services, 2014). However, until recently, content analyses were conducted by hand-coding relatively small samples of texts from traditional news sources.

Hand-coding can be quite accurate at the item-level, but it constrains the questions asked, and possibly study conclusions, because it limits the number and types of texts included. The amount of textual data that can be collected far outstrips what human coders can reasonably code. The quantity limitations of hand-coding have led to ignoring texts mentioning tobacco only in "passing" (e.g., movie stars smoking during interviews), constrained the analysis of social media (characterized by large quantities of items), and made hand-coding enough texts for stable weekly estimates over a long period impractical.

The exclusion of passing mentions and social media can result in inaccurate descriptions of content. Previous studies of tobacco coverage have not included passing mention content, and only one included social media content for all tobacco products (Myslín, Zhu, Chapman, & Conway, 2013). Those omitted content forms may have more conversational information that is less uniformly against tobacco use (Smith, Niederdeppe, Blake, & Cappella, 2013), and they may influence people's views. The extended elaboration likelihood model (Slater & Rouner, 2002) suggests that passing mentions may exert more influence on people's beliefs than content focused on tobacco because texts focused on different topics elicit less counter-arguing. Additionally, past studies of different tobacco product coverage in social media sources have generally characterized smaller samples of thousands of items (Cole-Lewis, Pugatch, et al., 2015; Luo et al., 2014; Myslín et al., 2013; Rose et al., 2017). Even if randomly selected, small samples from very large corpora may inaccurately estimate topic coverage, especially across long time periods. In this study, we use automated coding methods to examine the proportions of documents, including both passing mentions and social media in addition to focused coverage, that fall within a predetermined set of themes.

Automated coding methods allow many more items to be coded—in our case, the entire corpus from that time period: millions of tweets and hundreds of thousands from other sources—and enable the calculation of sensitive weekly estimates of tobacco coverage over a long timeframe. This methodological leap is critical for looking at changes in topic coverage over time, needed for analyses of effects of coverage, when time is the unit of analysis. We will show that high levels of item-level accuracy are not required for sensitive weekly estimates; one simply needs large numbers of texts each week. This is the key trade-off in using automated coding rather than hand-coding: classification errors may reduce the quality of coding for any single item, but the feasibility of coding large numbers of items every week increases, so one can reliably distinguish weeks with high and low coverage. When concerned with weekly accuracy, reliability is appropriately assessed at the weekly, rather than the item, level.

Automated coding

We searched for a broad topic from multiple sources including passing mention texts and social media sources, for three years (2014–2017). All tobacco product coverage, not just specific tobacco policies or products, was collected from long-form (broadcast news transcripts, Associated Press (AP) wire stories, newspapers, popular websites) and social media (Twitter and YouTube; see Appendix for details). Hand-coding was not feasible due to the ambitious amount of data (González-Bailón & Paltoglou, 2015; Grimmer & Stewart, 2013). There are two predominant automated approaches: unsupervised machine learning

(UML), and methods that rely on a small training/test sample of annotated documents to build an automated classifier (like both dictionary coding and supervised machine learning (SML)). UML describes underlying clusters of textual data from unlabeled texts (e.g., Jain, Zhu, & Conway, 2015; Zhan, Liu, Li, Leischow, & Zeng, 2017), but relies on researchers to name potentially theoretically uninteresting clusters *post hoc* (Grimmer & Stewart, 2013). In contrast, we wanted to use existing tobacco themes to directly compare the proportions of documents from new sources (i.e., passing mentions and social media) and traditional sources. For comparing across *a priori* specified themes, SML and dictionary methods are more appropriate (L. Nelson, Burk, Knudsen, & McCall, 2018).

SML and dictionary techniques rely on small samples of annotated texts to build coding algorithms that can be applied to new (unannotated) texts. Dictionary coding (e.g., Stone, Dunphy, & Smith, 1966) involves developing a list of keywords and iteratively testing how well they identify labeled training texts. We located relevant e-cigarette texts this way, but it was not ideal for analyzing themes with less predictable relevant terms. SML, in contrast, uses the co-occurrences of textual features in labeled training samples to develop algorithms predicting those labels. SML has been applied to tobacco tweets (Cole-Lewis, Varghese, et al., 2015; Kostygina, Tran, Shi, Kim, & Emery, 2016; Myslín et al., 2013), but not to tobacco coverage in other sources.

Themes

Tobacco themes in the literature have traditionally emerged from hand-coding hundreds to thousands of tobacco-relevant newspaper articles. The most prevalent themes from this research were tobacco control-related: negative health effects; secondhand smoke; policy/regulation; addiction; youth access, purchase, possession and use (content collected 2001–2010; D. Nelson et al., 2015; Smith et al., 2008, 2006; Wakefield et al., 2011). After the advent of e-cigarettes, Yates and colleagues (2015) and Wackowski and colleagues (2017) found the most prevalent e-cigarette themes in traditional sources still included policy/regulation, health, and youth (content collected 1997–2015). Following this history, we chose four themes *a priori*: Health, Policy, Addiction, and Youth.

Combining automated and crowd-sourced content analysis

As noted above, SML and dictionary automated content analysis begins with hand-coding samples of texts. We leveraged crowd-sourced hand-coding to annotate these smaller training/test samples. Crowd-sourced labeling uses multiple coders, who require less extensive training than expert coders, to determine a text's label based on the label the majority of coders chose. It efficiently allows hand-coding of many texts (Lind, Gruber, & Boomgaarden, 2017; Morris, 1994; Weber, 1990; Wolfe, Gephart, & Johnson, 1993).

Themes in passing mention coverage

No prior content analysis of tobacco coverage in long-form sources has included passing mention coverage because texts were limited to those predominantly about tobacco. Previous hand-coded content analyses of tobacco themes found texts by requiring a certain amount

of topic coverage (e.g., topical text in 50% of paragraphs: Wakefield et al., 2011, topical text in the headline: Wackowski et al., 2017, or texts hand-coded as about tobacco issues: D. Nelson et al., 2015; Smith et al., 2008, 2006), effectively restricting texts to tobacco-focused coverage. Therefore, the proportion of these four typical themes in passing coverage is unknown. By searching all long-form sources for tobacco keywords, our corpus included both focused and passing coverage of tobacco. We developed a method to distinguish between long-form texts that were substantially about tobacco (i.e., similar to prior hand-coded tobacco texts) versus only mentioning it in passing (see Methods). In contrast, we assume that all tobacco-relevant tweets and YouTube videos, given their brevity and selection, were substantially about that topic.

Themes in social media sources

A few studies have examined these tobacco themes in social media sources. A systematic review of 27 articles that coded e-cigarette or other tobacco tweets for valence or themes (Lienemann, Unger, Cruz, & Chu, 2017) found mostly smaller scale studies providing a few results relevant to our questions. One study comparing specific themes across e-cigarette and other tobacco tweets found that 40% of tweets collected 2011–2012 focused on first-person experiences, and the most frequent theme was cessation (14%; Myslín et al., 2013). Another study that sampled only e-cigarette-related tweets found that 26% of tweets collected 2013–2014 were advertisements, and the most frequent theme was policy/government (20%; Cole-Lewis, Pugatch, et al., 2015). Luo and colleagues' (2014) study of e-cigarettes found most YouTube videos in 2013 included weblinks for purchase (81%), but still often discussed health (71%). These few results overlap with our chosen traditional tobacco control themes, but offer limited samples and timeframes, and do not make comparisons with traditional media sources.

Weekly estimates of themes

In the absence of sufficient texts for calculating reliable weekly estimates, no prior content analyses of tobacco theme coverage have estimated trends at the weekly level. Theme trends have been examined at the yearly level (Eversman, 2015; D. Nelson et al., 2015; Rooke & Amos, 2014; Wakefield et al., 2011; Yates et al., 2015), and have shown that newspapers and broadcast TV covered each of the targeted themes proportionately less over time (e.g., TV health coverage decreased from 71% in 2004 to 23% in 2010; D. Nelson et al., 2015). More granular estimates of trends in theme coverage (e.g., at the weekly level) would allow accurately connecting tobacco theme coverage in the media to shorter-term public events, as well as to individuals' transient cognitions and behaviors.

The current study

The objective of the current study is to highlight the advantages of the proposed approach of combining automated and crowd-sourced content analysis. First, the approach allows us to assess coverage of well-described themes in previously unmeasured texts (i.e., passing mention tobacco coverage). Second, the approach allows a more accurate and comprehensive description of theme coverage in social media. Finally, the approach supports

the creation of reliable weekly estimates of theme coverage. All of these analyses are made possible because we can code large numbers of texts with considerably less effort relative to hand-coding by experts.

Methods

In the Appendix, we provide a detailed description of the coding processes and evidence for text-level coding validity. Here we present our methods in broad strokes.

Sources

The data consisted of long-form and social media texts published 2014–2017. We took three steps to identify content for analysis: (1) searching for potentially relevant content with keywords, (2) refining this corpus to texts clearly relevant to tobacco products using automated methods (counts listed refer to this stage), and (3) coding these texts for themes using automated methods. Long-form texts (n=135,764) were collected from Lexis-Nexis (broadcast TV and radio news transcripts, n=4,275; AP newswire, n=8,522; popular U.S. newspapers, n=52,561) and the MIT MediaCloud database (limited to websites popular among young people, n=70,406) using 12 search terms (e.g., “smoking”, “e-cigarette”, “vaping”, “nicotine”, “hookah”, “cigar”). Using many more search terms, tweets were collected from the Gnip Twitter Historical Powertrack for full historic access to all public tweets (n=75,322,911) and YouTube videos were collected using YouTube search APIs (n=12,262). See the Appendix for details on all search terms.

Locating product relevant items

We define *e-cigarettes* as electronic nicotine delivery systems used with or without nicotine. Given the novelty of e-cigarettes and disagreement in the public health community about their potential harm-reduction benefits (Royal College of Physicians, 2016; Truth Initiative, 2015; World Health Organization, 2014) versus their unknown long-term risks (Glasser et al., 2016; U.S. Department of Health and Human Services, 2016; Walley & Jenssen, 2015), we decided to separate e-cigarette coverage from coverage of all *other tobacco products* including both combustible and smokeless tobacco. Texts mentioning both e-cigarettes and other tobacco products were labeled as e-cigarette. In Step 1, expert-developed broad keyword searches located these two classes of texts (see Appendix for details on keyword selection). In Step 2, automated classifiers then refined the returned texts for relevancy to the product (Stryker, Wray, Hornik, & Yanovitzky, 2006) through iterative training and testing on fresh samples until adequate *precision* (proportion of cases hand-coded as relevant out of all items classified as a product) and *recall* (proportion of cases classified as a product out of all relevant items) was established. Precision was high for both long-form and social media sources (>.87). For almost all sources, recall was comparably high (>.86); only the classifier for YouTube other tobacco products located fewer of the relevant cases (recall=.72), perhaps because analysis was limited to available text rather than image features.

Identifying ‘more than passing mention’ (MPM) texts

For the long-form sources, we defined more than passing mention (MPM) coverage as texts having at least three of the long-form tobacco search terms (i.e., keywords) within 100

words as a proxy for a paragraph, as that is the low-end of paragraph length (Hacker & Sommers, 2010). Keywords also had to appear in more than one sentence operationalized as having more than 20 words between the first and last keyword, the upper-end of sentence length (Cutts, 2013). Finally, having a keyword in the title also categorized texts as MPM, as titles summarize content. Coding a sample of 771 texts showed this definition of MPM had precision=.95 and recall=.71. All other long-form texts were labeled passing mentions (PM). Twitter and YouTube content were all considered MPM, as noted above.

Assigning themes to items, item-level reliability

In Step 3, we coded this relevant corpus of texts for four themes using automated methods (see Figure 1). The themes were defined as: (1) Health: effects of use on the *user's* health excluding effects on non-users such as secondhand smoke, (2) Policy: mandatory policy/law/regulation by a government, company, or institution, (3) Addiction: explicit addiction references, and (4) Youth: use, access, or purchase by anyone ≤21 years old. Our process of coding themes was similar across products (e-cigarette and other tobacco) and sources (long-form, Twitter, and YouTube) allowing comparisons not possible previously. First, samples of texts were hand-coded for the theme definitions described above requiring reliability >.75. Then, algorithms were developed using automated methods to replicate the hand-coding based on features of the texts. Algorithms were deemed of adequate quality if precision and recall for a held-out test set were >.70. Finally, the theme algorithms were applied to the population of relevant source items (i.e., long-form texts (*N*=135,764), tweets (*N*=75,322,941), and YouTube videos (*N*=12,262)). Components of this process differed for long-form and social media sources as described below and in more detail in the Appendix.

In the long-form process, coders were recruited from Amazon Mechanical Turk (MTurkers), trained to detect theme presence or absence, assessed, and retained if they showed adequate skill. Second, retained coders (N=163 across themes) hand-coded a sample of texts for the theme and their confidence in that judgment on a 1 (*not at all*) to 5 (*very confident*) scale. Each of 2,400 texts was hand-coded by 7 to 9 MTurkers. Texts were retained for the training and test samples if 75% of MTurkers agreed on the code (alpha>.75).[8] Third, 80% of the retained texts (the training sample) were used to iteratively develop an optimum logistic regression SML algorithm for theme classification using the Python *scikit learn* package (Pedregosa, Varoquaux, Gramfort, Michel, & Thirion, 2011). Each algorithm produced predicted probabilities that texts were theme-relevant. Fourth, the final algorithm quality was judged by the strength of the theme probabilities' correlation with the MTurkers' assessments of the held-out test set texts (i.e., the mean confidence rating per text), and precision and recall estimates. We assessed reliability for all long-form sources combined assuming that the quality of coding would not differ by source.

In contrast, for social media, *experts* first coded samples for all themes (5,427 tweets and 2,088 videos). Second, a subset of these items was double-coded and reliability assessed with prevalence-adjusted bias-adjusted kappa (PABAK, >.75 for all themes). Third, for Twitter, experts developed keywords and exclusion terms for each theme, and iteratively

[8]All texts were coded for themes using the finalized algorithms. However, only clear texts were used to train the algorithm (74–91% of texts were retained depending on theme).

refined the search terms using Python on the hand-coded training set. For YouTube, the text features used to develop four variations of SML classifiers included the title, descriptions, and transcripts of the training set. An automated ensemble method was used to assign theme classification to training set videos based on the majority judgment from five classifiers: the Twitter theme classifier and four SML classifiers developed with different parameters. Though it would have been preferable to use exactly the same methods for all sources, this was not possible, as discussed below.

Event validation

We assessed the external validity of theme coding by asking whether the measured quantity of theme coverage was responsive to relevant high salience public events. On days when five or more broadcast news national shows were coded for a theme, we qualitatively reviewed the transcript texts to confirm the event. Three e-cigarette events and four other tobacco events met this criterion. For the remaining long-form sources, Twitter, and YouTube, we summed theme coverage during these events. Our validity test compared those estimates to overall average daily coverage using t-tests. Significantly higher theme coverage on event days (compared to average) would indicate that the classifiers were effective at locating days with appropriately large theme coverage.

Weekly-level reliability

The third aim of this study is to show that this coding method generates reliable weekly estimates. Therefore, we also assess algorithm quality at the weekly level. Even with substantial item-level error, weekly-level estimates of theme coverage may be accurate, as long as there are enough items each week to create stable estimates and there is true variation across weeks. To assess weekly reliability, we first calculated the consistency of weekly theme coverage estimates from randomly split halves. We re-estimated the split-half correlation 100 times and averaged those. Second, we used the Spearman-Brown prophecy formula to estimate the reliability of a weekly measure summing two halves (Eisinga, te Grotenhuis, & Pelzer, 2013). We required reliability of >.70 for weekly estimates (Lacy, Watson, Riffe & Lovejoy, 2015).

Data analysis plan

We assessed the reliability and validity of our theme coding using the item-level precision and recall, event validation, and weekly-level reliability described above. We used this coding to compare coverage in PM versus MPM texts. We used t-tests to test differences in numbers of themes per text by type of coverage (PM vs MPM) within product, and to compare coverage in social media versus MPM long-form texts. Finally, we compared the weekly variation across sources and themes using the coefficient of variation (CV=standard deviation/mean), modified signed-likelihood ratio test (SLRT) for equality of CVs (Krishnamoorthy & Lee, 2014), and visual inspection of over time graphs.

Results

At the item level, precision and recall of coding in the held-aside test sets met our threshold of >.70 in all sources except YouTube (see Table 1; correlations with mean confidence

rating per long-form text all >.87). Testing the validity of this coding at the event-level, YouTube was again the only source that did not consistently cover these themes more on event days than on average (see Table 2; all t-tests comparing average theme coverage to event days were significant). Testing the reliability of weekly estimates, e-cigarette and other tobacco coverage were consistently >.70 across sources except for YouTube and e-cigarette-PM long-form texts (see Table 3). Against all three of our criteria, the automated theme coding among long-form sources and Twitter is valid. In contrast, weekly e-cigarette PM coding, and most YouTube coding lacked adequate validity evidence. While we continue to present YouTube results, we do so with less confidence than for other sources. (See https://repository.upenn.edu/asc_papers/689 and https://repository.upenn.edu/asc_papers/687 for theme-product examples from each source.)

Themes in PM versus MPM texts

Comparing MPM and PM long-form texts for e-cigarette and other tobacco presents several striking results. First, the two product categories show different patterns. Other tobacco coverage included many texts not substantially about tobacco (84% PM), while e-cigarette coverage had many fewer texts not substantially about e-cigarettes (33% PM). PM texts were less likely to include one or more themes than were MPM texts (0.4 versus 2.1 themes for e-cigarettes ($t(10,597)=-80.0$, $p<.001$); 0.3 versus 1.0 for other tobacco ($t(125,163)=-182.9$, $p<.001$); Table 4), indicating that including PM texts yields different theme prevalence information than MPM texts alone. Policy was the theme most commonly discussed for all long-form e-cigarette texts (68% of MPM and 22% of PM); Youth was relatively infrequently the theme of other tobacco texts (13% of MPM and 4% of PM). Individual themes were not otherwise consistently ordered.

Themes in social media versus MPM long-form texts

Overall, there was much more coverage of other tobacco products than of e-cigarettes; only YouTube had more e-cigarette than other tobacco items (3.0 times more, see Table 4). The novelty of e-cigarettes did not result in greater e-cigarette coverage relative to other tobacco products. As we move to inclusion of themes and away from overall coverage, the patterns are different. For MPM long-form sources, e-cigarette texts were much more likely than other tobacco texts to address each theme (a median of 2.0 times more likely across themes).

Twitter and YouTube contained fewer themes per item than MPM long-form texts (e-cigarettes: 0.1 tweets and 0.2 videos versus 2.1 MPM texts; other tobacco: 0.1 tweets and 0.3 videos versus 1.0 MPM texts, all $p<.001$; Table 4). MPM long-form texts have, on average, many more words than tweets, and therefore more space to cover multiple themes.

Weekly estimates

Our dataset contained enough items per week to assess weekly variation in coverage of these themes by product and source using automated methods. As described above, for MPM long-form texts, PM long-form texts (other tobacco only), and Twitter, these estimates reliably distinguished between weeks of high and low theme coverage. Comparing coefficients of variation by product among MPM long-form texts and Twitter, weekly e-cigarette theme coverage had significantly more variability than weekly other tobacco theme

coverage (7 of 8 tests, $p<.001$; Table 5). Averaging across the four themes, Twitter had more weekly variation in theme coverage (M_{ecig}=86%, M_{other}=53%) than MPM long-form texts (M_{ecig}=72%, M_{other}=40%), which in turn had more variation than PM long-form texts (for other tobacco products, M_{other}=20%). Figure 2 shows weekly MPM long-form e-cigarette health coverage over-time with key events labeled (e.g., the deeming rule). See Appendix for all over time graphs.

Discussion

This paper outlines and provides validity evidence for a coding approach combining automated and crowd-sourced content analysis of tobacco-relevant items. By relying on automated coding, we assigned codes for product and theme to a large quantity of items: 135,764 long-form texts, 75,322,941 tweets, and 12,262 YouTube videos, capturing e-cigarette and other tobacco media coverage over 162 weeks. There is good evidence that the automated coding was valid for most sources and both product categories at the item-, event-, and week-level, although YouTube was an exception.[9] The coding allowed comparisons of theme coverage in MPM long-form texts versus PM and social media items, and estimates of weekly variation in theme coverage over time.

Given strong evidence for valid coding, we are confident in our conclusions that, for long-form texts, other tobacco products received more attention than e-cigarettes overall, but e-cigarette texts were more substantive, with more themes discussed and fewer PM texts. One explanation for the greater presence of e-cigarette MPM texts and theme coverage in long-form sources is that the novelty of the product and corresponding media coverage of its health effects, related policies, addictive properties, and relevance to youth is more “newsworthy”. E-cigarette policies were widely covered across all sources.

Our theme coverage estimates among MPM e-cigarette long-form texts are larger than prior estimates (e.g., Wackowski et al., 2017: their e-cigarette Policy=45%, ours=68%; their e-cigarette Health=22%, ours=44%). The rising popularity of e-cigarettes and their increasing prominence in public discussion since prior research was conducted may explain this difference. Additionally, our estimates of Twitter e-cigarette theme coverage are lower than previous ones (policy: 3% vs. 20%, health: 2% vs. 13%; Cole-Lewis, Pugatch, et al., 2015). Of course, differences in methods, particularly search comprehensiveness, as well as time periods may also produce different results. Our estimates add to the prior literature by including theme coverage in PM texts, which is especially important for other tobacco long-form texts that are mostly PM (84%). In line with prior research, our data suggest less coverage of these themes related to e-cigarettes on social media compared to MPM long-form texts.

Strengths

We have collected close to a census of e-cigarette and tobacco relevant texts from diverse sources, both within source and across media type, over 162 weeks. Though we did not

[9]YouTube classification was disadvantaged because analysis was limited to available text features and ignored the image features of the videos.

cover the entire range of media sources contributing to the tobacco and e-cigarette public communication environment, this text corpus represents a large proportion. One important benefit of the breadth of coverage is that we can make reliable over-time estimates for use in longitudinal analyses of coverage effects.

Our validation of theme coding in e-cigarette and other tobacco coverage is quite strong relative to prior work, as the only previous study to use SML for theme coding did not calculate validity statistics (Cole-Lewis, Pugatch, et al., 2015). Our use of crowd-sourced coding for a small sample of long-form data was a cost-effective method for labeling many texts reliably without investing resources in coders on site (Lind et al., 2017). Crowdsourcing decreases the influence of individual bias, minimizes negative effects due to rater fatigue, and produces better time- and cost-efficiency compared to other approaches to hand-coding (Morris, 1994; Weber, 1990; Wolfe et al., 1993). This is the first communication study to take advantage of crowdsourcing to facilitate automated content analysis.

Moving from binary coding to continuous measures of theme likelihood yields greater measurement sensitivity. Using average predicted probabilities from the long-form SML classifiers allowed us to capture more variation and nuance at the weekly level, especially for sources with few relevant texts. The development of an SML algorithm is fairly computationally intensive; however, for experienced programmers, it is faster than hand-coding, and is the only solution when working with complex concepts in very large datasets.

Finally, assessing PM texts in long-form coverage expands our understanding of the communication environment. PM texts are particularly important to measure because the average person may be more likely to read articles about celebrities with passing mentions of tobacco, for example, than entire articles devoted to tobacco control. Including PM texts in future analyses of media effects may better capture actual exposure and thus the impact of the public media environment on tobacco beliefs and intentions.

Limitations

One limitation of this study is that we did not use identical coding methods for all social media and long-form sources. Although the definitions were the same and the quality of the classifiers met the same threshold, there were differences in search terms, coders of the small hand-coded samples, and specific SML algorithms. Search terms had to be different because social media sources require a larger set of search terms in order to pull in the comprehensive corpus of tobacco and e-cigarette texts. This same exhaustive search would pull in too many irrelevant long-form texts. Algorithms routinely differ to reach optimum quality for different training sets. As for coders, it would have been better to use the same methods for all sources, but when working with such a large quantity of data from very different sources in parallel, this was not possible. Still, we trust that our comparison across sources is a step forward from comparing sources analyzed by different researchers with different definitions and thresholds for quality across different papers.

A second limitation of this study is that theme coding was applied to all texts, but conclusions were drawn from looking separately at themes within e-cigarette and other

tobacco product coverage. Texts mentioning both e-cigarettes and other tobacco products were labeled as e-cigarette. Therefore, the proportion of e-cigarette texts mentioning each theme (e.g., Addiction) may be over-estimated, since addiction in the text might actually be about addiction to cigarettes, not addiction to e-cigarettes. Additionally, texts were only coded for themes, not for the valence of these themes. Thus, items coded for health effects might be supportive of the use of e-cigarettes, but opposed to the use of other tobacco products. Future research should test whether texts independently coded as containing e-cigarettes and specific themes are indeed combinations of e-cigarette specific themes, and code for the valence of these themes.

Another limitation is that we pre-defined our four themes for simplicity. Although UML approaches may have enabled us to uncover unexpected themes (Jain et al., 2015; Lazard, Wilcox, Tuttle, Glowacki, & Pikowski, 2017; Prier, Smith, Giraud-Carrier, & Hanson, 2011; Zhan et al., 2017), those themes may not have been of practical importance. By choosing our themes *a priori*, we ensured our ability to assess and compare themes of interest. Future work might use UML to uncover other conceptually-important themes.

Finally, this is a first step on the way to effect studies. We recognize that we have only measured opportunities for exposure, and not captured the specific media diet of a population. To use this content information in effects studies, we need to assume that these measures reflect what people are exposed to, i.e., that they are indicators of the public communication environment that affects people. We can establish the legitimacy of that assumption only as we move from content analysis of media content to evidence for effects of that content, whether on policy views, or on individual beliefs and behavior.

Conclusion

Automated coding methods are increasingly important as social scientists struggle to uncover meaning from large amounts of textual data. This study shows that automated methods, though not perfect, can be applied to any tobacco or e-cigarette reference (no matter how brief), including from Twitter, and can be used to estimate weekly coverage. Furthermore, results from tests at the item, daily, and weekly levels instill confidence in the validity of our coding. In the future, we plan to examine the impact of these coded themes at the weekly level on young people’s tobacco and e-cigarette beliefs, intentions, and use.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

Cole-Lewis H, Pugatch J, Sanders A, Varghese A, Posada S, Yun C, … Augustson E (2015). Social listening: A content analysis of e-cigarette discussions on Twitter. Journal of Medical Internet Research, 17(10), e243. 10.2196/jmir.4969 [PubMed: 26508089]

Cole-Lewis H, Varghese A, Sanders A, Schwarz M, Pugatch J, & Augustson E (2015). Assessing electronic cigarette-related tweets for sentiment and content using supervised machine learning. Journal of Medical Internet Research, 17(8), e208. 10.2196/jmir.4392 [PubMed: 26307512]

Cutts M (2013). Oxford Guide to Plain English (4 edition). Oxford: Oxford University Press.

Dunlop SM, Cotter T, Perez D, & Chapman S (2011). Tobacco in the news: associations between news coverage, news recall and smoking-related outcomes in a sample of Australian smokers and recent quitters. Health Education Research, 27(1), 160–171. 10.1093/her/cyr105 [PubMed: 22156232]

Dunlop SM, & Romer D (2010). Relation between newspaper coverage of “light” cigarette litigation and beliefs about “lights” among American adolescents and young adults: the impact on risk perceptions and quitting intentions. Tobacco Control, 19(4), 267–273. 10.1136/tc.2009.032029 [PubMed: 20530139]

Eisinga R, te Grotenhuis M, & Pelzer B (2013). The reliability of a two-item scale: Pearson, Cronbach, or Spearman-Brown? International Journal of Public Health, 58(4), 637–642. [PubMed: 23089674]

Eversman MH (2015). Harm reduction in U.S. tobacco control: Constructions in textual news media. International Journal of Drug Policy, 26(6), 575–582. 10.1016/j.drugpo.2015.01.018 [PubMed: 25727451]

Glasser AM, Collins L, Pearson JL, Abudayyeh H, Niaura RS, Abrams DB, & Villanti AC (2016). Overview of electronic nicotine delivery systems: A systematic review. American Journal of Preventive Medicine 10.1016/j.amepre.2016.10.036

González-Bailón S, & Paltoglou G (2015). Signals of public opinion in online communication: A comparison of methods and data sources. The ANNALS of the American Academy of Political and Social Science, 659(1), 95–107. 10.1177/0002716215569192

Grimmer J, & Stewart BM (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. Political Analysis, 21(3), 267–297.

Hacker D, & Sommers N (2010). A Writer’s Reference (7th edition). Boston ; New York: Bedford/St. Martin’s.

Hornik RC (2002). Public Health Communication: Evidence for Behavior Change Mahwah, N.J.: L. Erlbaum Associates.

Jain S, Zhu S-H, & Conway M (2015). Exploring consumer attitudes towards hookah, cigarettes, and cigars using Twitter. Tobacco Regulatory Science, 1(3), 198–203. 10.18001/TRS.1.3.1

Kostygina G, Tran H, Shi Y, Kim Y, & Emery S (2016). ‘Sweeter Than a Swisher’: amount and themes of little cigar and cigarillo content on Twitter. Tobacco Control, 25, i75–i82. 10.1136/tobaccocontrol-2016-053094 [PubMed: 27697951]

Krishnamoorthy K & Lee M (2014). Improved tests for the equality of normal coefficients of variation. Journal Computational Statistics, 29(1–2), 215–232.

Lacy S, Watson BR, Riffe D, & Lovejoy J (2015). Issues and best practices in content analysis. Journalism & Mass Communication Quarterly, 92(4), 791–811.

Lazard AJ, Wilcox GB, Tuttle HM, Glowacki EM, & Pikowski J (2017). Public reactions to e-cigarette regulations on Twitter: a text mining analysis. Tobacco Control, 26, e112–e116. 10.1136/tobaccocontrol-2016-053295 [PubMed: 28341768]

Lienemann BA, Unger JB, Cruz TB, & Chu K-H (2017). Methods for coding tobacco-related Twitter data: A systematic review. Journal of Medical Internet Research, 19(3), e91. 10.2196/jmir.7022 [PubMed: 28363883]

Lind F, Gruber M, & Boomgaarden HG (2017). Content analysis by the crowd: Assessing the usability of crowdsourcing for coding latent constructs. Communication Methods and Measures, 11(3), 191–209. 10.1080/19312458.2017.1317338 [PubMed: 29118893]

Luo C, Zheng X, Zeng DD, & Leischow S (2014). Portrayal of electronic cigarettes on YouTube. BMC Public Health, 14(1). 10.1186/1471-2458-14-1028

Menashe CL, & Siegel M (1998). The power of frame: an analysis of newspaper coverage of tobacco issues - United States, 1985–1996. Journal of Health Communication, 3, 307–325. [PubMed: 10977260]

Morris R (1994). Computerized content analysis in management research: A demonstration of advantages & limitations. Journal of Management, 20(4), 903–931. 10.1177/014920639402000410

Myers AE, Southwell BG, Ribisl KM, Moreland-Russell S, Bowling JM, & Lytle LA (2018). State-level Point-of-Sale tobacco news coverage and policy progression over a 2-year period. Health Promotion Practice, 1524839917752108. 10.1177/1524839917752108

Myslín M, Zhu S-H, Chapman W, & Conway M (2013). Using Twitter to examine smoking behavior and perceptions of emerging tobacco products. Journal of Medical Internet Research, 15(8), e174. 10.2196/jmir.2534 [PubMed: 23989137]

Nelson D, Pederson LL, Mowery P, Bailey S, Sevilimedu V, London J, … Pechacek T (2015). Trends in U.S. newspaper and television coverage of tobacco. Tobacco Control, 24(1), 94–99. 10.1136/tobaccocontrol-2013-050963 [PubMed: 23864404]

Nelson L, Burk D, Knudsen M, & McCall L (2018). The future of coding: A comparison of hand-coding and three types of computer-assisted text analysis methods. Sociological Methods & Research, 0049124118769114. 10.1177/0049124118769114

Pedregosa F, Varoquaux G, Gramfort G, Michel V, Thirion B, Grisel O, … Duchesnay E (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

Prier KW, Smith MS, Giraud-Carrier C, & Hanson CL (2011). Identifying health-related topics on twitter. In Salerno J, Yang SJ, Nau D, & Chai S-K (Eds.), Social Computing, Behavioral-Cultural Modeling and Prediction (pp. 18–25). Springer Berlin Heidelberg. Retrieved from 10.1007/978-3-642-19656-0_4

Rooke C, & Amos A (2014). News media representations of electronic cigarettes: an analysis of newspaper coverage in the UK and Scotland. Tobacco Control, 23, 507–512. 10.1136/tobaccocontrol-2013-051043 [PubMed: 23884011]

Rose SW, Jo CL, Binns S, Buenger M, Emery S, & Ribisl KM (2017). Perceptions of menthol cigarettes among Twitter users: Content and sentiment analysis. Journal of Medical Internet Research, 19(2), e56. 10.2196/jmir.5694 [PubMed: 28242592]

Royal College of Physicians. (2016, April 28). Nicotine without smoke: Tobacco harm reduction Retrieved July 26, 2017, from https://www.rcplondon.ac.uk/projects/outputs/nicotine-without-smoke-tobacco-harm-reduction-0

Slater MD & Rouner D (2002). Entertainment-education and Elaboration Likelihood: Understanding the processing of narrative persuasion. Communication Theory 12(2), 173–191. 10.1111/j.1468-2885.2002.tb00265.x

Smith KC, Niederdeppe J, Blake KD, & Cappella JN (2013). Advancing cancer control research in an emerging news media environment. JNCI Monographs, 2013(47), 175–181. 10.1093/jncimonographs/lgt023

Smith KC, Wakefield MA, Terry-McElrath Y, Chaloupka FJ, Flay B, Johnston L, … Siebel C (2008). Relation between newspaper coverage of tobacco issues and smoking attitudes and behaviour among American teens. Tobacco Control, 17(1), 17–24. 10.1136/tc.2007.020495 [PubMed: 18218802]

Smith KC, Wakefield M, & Edsall E (2006). The good news about smoking: How do U.S. newspapers cover tobacco issues? Journal of Public Health Policy, 27(2), 166–181. [PubMed: 16961195]

Stone P, Dunphy D, & Smith M (1966). The general inquirer: A computer approach to content analysis Oxford, England: M.I.T. Press.

Stryker JE, Wray RJ, Hornik RC, & Yanovitzky I (2006). Validation of database search terms for content analysis: The case of cancer news coverage. Journalism and Mass Communication Quarterly; Columbia, 83(2), 413–426,428–430.

Truth Initiative. (2015, December). The truth about: Electronic nicotine delivery systems Retrieved July 26, 2017, from zotero://attachment/713/

U.S. Department of Health and Human Services. (2014). The health consequences of smoking-−-50 years of progress: A report of the Surgeon General Atlanta, GA: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health. Retrieved from http://www.cdc.gov/tobacco/data_statistics/sgr/2012/

U.S. Department of Health and Human Services. (2016). E-cigarette use among youth and young adults: A report of the Surgeon General Atlanta, GA: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office of Smoking and Health. Retrieved from https://e-cigarettes.surgeongeneral.gov/documents/2016_SGR_Full_Report_non-508.pdf

Wackowski OA, Giovenco DP, Singh B, Lewis MJ, Steinberg MB, & Delnevo CD (2017). Content analysis of US news stories about e-cigarettes in 2015. Nicotine & Tobacco Research 10.1093/ntr/ntx170

Wackowski OA, Lewis MJ, Delnevo CD, & Ling PM (2013). A content analysis of smokeless tobacco coverage in U.S. newspapers and news wires. Nicotine & Tobacco Research, 15(7), 1289–1296. 10.1093/ntr/nts332 [PubMed: 23288875]

Wakefield MA, Brennan E, Durkin SJ, McLeod K, & Smith KC (2011). Still a burning issue: Trends in the volume, content and population reach of newspaper coverage about tobacco issues. Critical Public Health, 21(3), 313–325. 10.1080/09581596.2010.502930

Walley SC, & Jenssen BP (2015). Electronic nicotine delivery systems. Pediatrics, 136(5), 1018–1026. 10.1542/peds.2015-3222 [PubMed: 26504128]

Weber RP (1990). Basic content analysis (2nd ed.). Newbury Park, CA: SAGE.

Wolfe RA, Gephart RP, & Johnson TE (1993). Computer-facilitated qualitative data analysis: Potential contributions to management research. Journal of Management, 19(3), 637–660. 10.1177/014920639301900307

World Health Organization. (2014). Electronic nicotine delivery systems Retrieved July 26, 2017, from http://apps.who.int/gb/fctc/PDF/cop6/FCTC_COP6_10-en.pdf

Yates K, Friedman K, Slater MD, Berman M, Paskett ED, & Ferketich AK (2015). A content analysis of electronic cigarette portrayal in newspapers. Tobacco Regulatory Science, 1(1), 94–102. 10.18001/TRS.1.1.9 [PubMed: 26229974]

Zhan Y, Liu R, Li Q, Leischow SJ, & Zeng DD (2017). Identifying topics for e-cigarette user-generated contents: A case study from multiple social media platforms. Journal of Medical Internet Research, 19(1), e24. 10.2196/jmir.5780 [PubMed: 28108428]

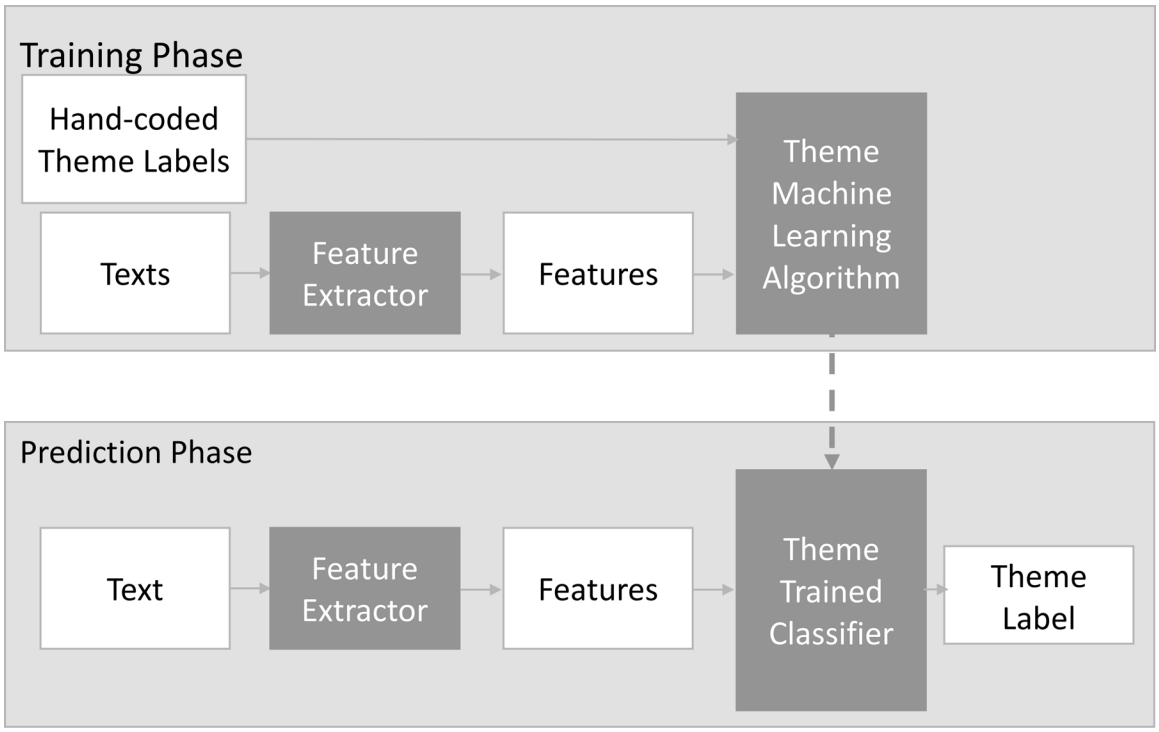

Figure 1.

Conceptual flow chart of algorithm development for all sources. The training phase relies on hand-coding of a small sample of texts. The prediction phase is used for both small test samples (to check the quality of the classifiers) and coding entire unlabeled databases. Once a classifier is finalized, the feature extractor and classifier are identical in the training and prediction phases.

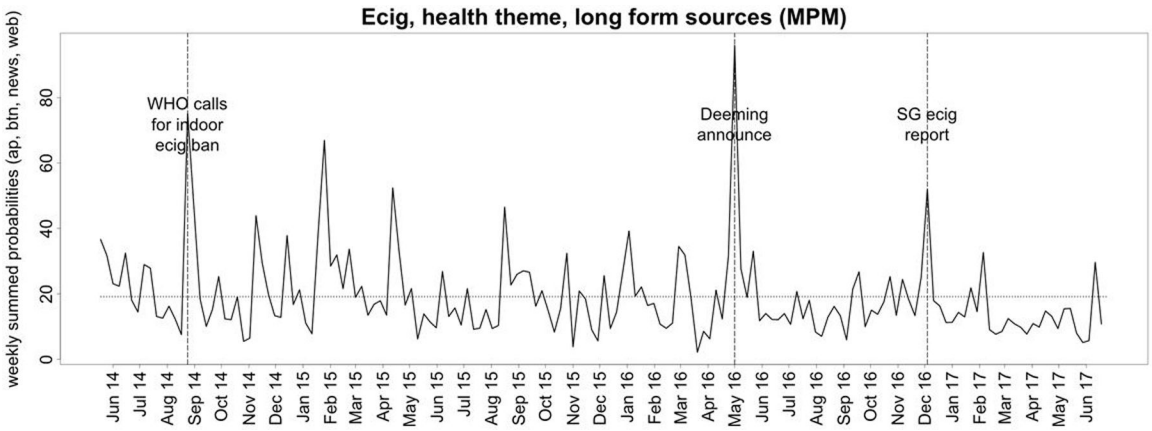

Figure 2.

MPM (more than passing mention) long-form e-cigarette coverage. *Note*. The horizontal dotted lines depict the mean coverage over the entire period for each graph. Vertical dashed lines are labeled events corresponding to weeks with days that received a lot of coverage (top 3) overall or within each theme. Ecig=e-cigarettes. WHO calls for indoor ecig ban=the World Health Organization (WHO) called for banning e-cigarettes indoors & the American Heart Association called for regulating e-cigarettes like tobacco. Deeming announce=the FDA announced they would regulate e-cigarettes, cigars, and all other tobacco products. SG ecig report=the Surgeon General released a report on e-cigarette use in youth and young adults.

Table 1.

Test Set Precision and Recall of Themes

	Health		Policy		Addiction		Youth	
	Precision	**Recall**	**Precision**	**Recall**	**Precision**	**Recall**	**Precision**	**Recall**
Long-form texts (*n*s range 335–397)	.94	.88	.96	.90	.76	.97	.81	.85
Tweets (*n*=651)	.77	.81	.82	.85	.79	.96	.73	.89
YouTube videos (*n*=590)	.73	.55	.65	.45	.82	.39	.59	.33

Note: Precision and Recall for the Long-form texts was calculated by binning the probabilities from the long-form classifiers into those probabilities >.50 versus the others.

Table 2.

Dates with High Amounts of Theme Coverage Compared to Average Daily Theme Coverage

		Health			Policy		Addiction			Youth	
Source	**Event Name**	**Gwynn (tob) 6/16/14**	**Nimoy (tob) 2/27/15**	**Processed meats (tob) 10/26/15[a]**	**CVS Health (tob) 9/3/14**	**Deeming announce (ecig) 5/5/16[a]**	**CVS Health (tob) 9/3/14**	**Deeming announce (ecig) 5/5/16**	**Student ecig use up (ecig) 4/16/15[a]**	**Deeming announce (ecig) 5/5/16[a]**	**SG ecig report (ecig) 12/8/16**
Broadcast news	Event	**4.8**	**5.6**	**9.6**	**6.6**	**9.6**	**5.2**	**5.4**	**6.3**	**8.6**	**4.1**
	All		0.4		0.2	0.1	0.4	0.1		0.1	
Long-form (other 3)	Event	**29.6**	**32.3**	**71.5**	**34.8**	**70.7**	**33.5**	**63.6**	**25.5**	**60.3**	**30.8**
	All		15.3		14.0	4.8	11.6	3.1		3.3	
Twitter	Event	**1,857**	**1,471**	**4,118**	**9,295**	**15,449**	**2,084**	**482**	**7,526**	**5,576**	**10,083**
	All		1,352		1,231	699	1,115	194		747	
YouTube	Event	**NA**	**0.0**	**1.5**	**0.0[b]**	**6.0**	**0.0**	**0.0[c]**	**0.0**	**0.5**	**1.0**
	All		0.4		0.1	1.6	0.2	0.0		0.1	

Note. All long-form values are summed theme probabilities; all social media values are counts. Both are daily sums averaged over the relevant time period. Shaded values indicate that source did not have high amounts of theme coverage for that event compared to the overall average. Event dates were chosen when broadcast news theme coverage was > .50 in at least 5 shows that day because broadcast news time limits only allow coverage of very salient events. Consecutive dates are grouped into single events. Long-form estimates exclude Broadcast News since that source was used to choose event dates. Ecig=e-cigarettes. Tob=Other tobacco products (excludes e-cigarettes). CVS Health=CVS announced they would stop selling tobacco products. Deeming announce=the FDA announced they would regulate e-cigarettes, cigars, and all other tobacco products. Gwynn=baseball's Tony Gwynn died of smoking-related causes. Nimoy=actor Leonard Nimoy died from smoking-related causes. Processed meats=the World Health Organization ranked processed meats in the same cancer risk category as smoking. Student ecig use up=the CDC released a report that e-cigarette use tripled in middle and high school students. SG ecig report=the Surgeon General released a report on e-cigarette use in youth and young adults. Event=day of the event. All=all days. NA=data not available for YouTube on this day.

[a]This event covered two days.

[b]A Youtube video relevant to CVS Health (tob) aired the following day, 9/4/14. It was coded as Policy, but not Addiction.

[c]A Youtube video relevant to the deeming announcement (ecig) aired the following day, 5/6/16. It was coded as Addiction.

Table 3.

Spearman-Brown Reliabilities for Weekly Estimates within E-cigarette Items Versus Items Only about Other Tobacco Products

		E-cigarette items						Other tobacco items					
		Items	**Ecig**	**Hea.**	**Pol.**	**Add.**	**You.**	**Items**	**Other**	**Hea.**	**Pol.**	**Add.**	**You.**
Long-form	MPM	7,139	.92	.90	.94	.92	.94	19,939	.88	.87	.86	.84	.79
	PM	3,460	.85	.31	.64	.56	.64	105,226	.94	.84	.83	.79	.76
Social Media	Twitter	24.3M	>.99	>.99	>.99	>.99	>.99	51.0M	>.99	>.99	>.99	>.99	>.99
	YouTube	9,168	.92	.36	.86	.15	.32	3,094	.85	.38	.17	.34	.30

Note. Values are Spearman Brown reliabilities = (2*split-half correlation)/(1+split-half correlation) calculated from the average of 100 boot-strapped random split-half correlations with week as the unit of analysis (N=162 weeks, except YouTube where N=155). This estimates how reliable the full dataset (not just half of it) distinguishes between weekly estimates. Values for long-form texts are averaged across all four sources (weighted by their item frequencies). Ecig=E-cigarette items; Hea.=Health; Pol.=Policy; Add.=Addiction; You.=Youth; Other=Other tobacco items. Shaded reliabilities are lower than our threshold of .70.

Table 4.

Theme Coverage within E-cigarette Texts Versus Texts Only about Other Tobacco Products

		E-cigarette texts						Other tobacco texts					
		N	**Health**	**Policy**	**Addiction**	**Youth**	**Ave # Themes**	**N**	**Health**	**Policy**	**Addiction**	**Youth**	**Ave # Themes**
Long-form	MPM	7,139	.44	.68	.48	.50	2.1	19,939	.29	.32	.26	.13	1.0
	PM	3,460	.05	.22	.07	.09	0.4	105,226	.11	.09	.08	.04	0.3
Social Media	Twitter	24.3M	.02	.03	.01	.03	0.1	51.0M	.03	.03	.02	.04	0.1
	YouTube	9,168	.03	.19	<.01	.01	0.2	3,094	.14	.02	.06	.05	0.3

Note. Values for long-form texts are mean predicted probabilities averaged across all four sources (weighted by their item frequencies); for Twitter and YouTube, values are proportions. Ave # Themes=Average number of themes per item (i.e., text, tweet, video). MPM=More than passing mention. PM=Passing mention. M=Million.

Table 5.

Weekly Theme Coverage within E-cigarette Items Versus Items Only about Other Tobacco Products

			E-cigarette items					Other tobacco items				
			Ecig	**Health**	**Policy**	**Addiction**	**Youth**	**Other**	**Health**	**Policy**	**Addiction**	**Youth**
Long-form	MPM	Mean	43.9	19.1	30.0	20.9	22.0	122.6	36.2	39.6	32.1	15.8
		SD	23.9	12.9	20.7	14.8	18.1	32.5	14.9	15.6	11.6	7.0
		CV	54%	67%[a]	69%	71%	82%	26%	41%	40%[a]	36%	44%[a]
	PM	Mean	21.2	1.1	4.7	1.5	2.0	646.5	74.2	60.1	52.1	26.9
		SD	11.8	0.7	2.9	0.9	1.5	104.3	16.2	13.3	8.8	5.1
		CV	56%	62%[a]	63%	58%[a]	76%	16%	22%[ab]	22%[ab]	17%[ab]	19%[ab]
Social Media	Twitter	Mean	149,590	3,279	4,903	1,358	5,231	313,158	9,473	8,635	7,819	11,266
		SD	47,856	3,520	3,586	1,047	4,505	77,310	3,899	4,244	2,741	9,749
		CV	32%	107%	73%	77%	86%	25%	41%	49%	35%	87%
	YouTube[c]	Mean	59.0	1.8	11.3	0.3	0.9	19.9	2.8	0.4	1.2	0.9
		SD	27.7	1.6	8.7	0.6	1.1	11.4	2.1	0.7	1.3	1.1
		CV	47%	93%	77%	203%	128%	58%	76%	164%	111%	125%

Note. N=162 weeks, except YouTube where N=155. Values for long-form texts are mean weekly summed predicted probabilities across all four sources (weighted by their item frequencies); for Twitter and YouTube, values are weekly counts. Ecig=E-cigarette items. Other=Other tobacco items. MPM=More than passing mention. PM=Passing mention. CV=Coefficient of variation.

[a]Less variation than Twitter according to the modified signed-likelihood ration test (SLRT) at $p<.05$

[b]Less variation than MPM according to the modified signed-likelihood ration test (SLRT) at $p<.05$

[c]We did not test variation against YouTube because of its low reliability and validity numbers.