Check for updates

Health Informatics Journal

Original Article

Health Informatics Journal
1–11
© The Author(s) 2018
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/1460458217754242
journals.sagepub.com/home/jhi
SAGE

Online data to contextualize waterpipe tobacco smoking establishments surrounding large US universities

Jason B Colditz, Kar-Hai Chu, Galen E Switzer, Konstantinos Pelechrinis and Brian A Primack
University of Pittsburgh, USA

Abstract

Waterpipe tobacco smoking has grown in popularity among US college students and is associated with serious health risks. Much of the waterpipe tobacco smoking takes place in establishments such as "hookah bars" or in lounge settings. Web-based data platforms such as Yelp have demonstrated utility in locating these establishments but are prone to over- and underestimation. The purpose of this study was to optimize strategies for algorithmically estimating the prevalence of waterpipe tobacco smoking establishments. We conducted searches for potential waterpipe tobacco smoking establishments near highly residential US universities ($N=41$). Of 521 potential establishments, independent coders confirmed 257 as permitting waterpipe tobacco smoking. We compared four strategies for using Yelp metadata to estimate the number of confirmed waterpipe tobacco smoking establishments by location. An accuracy-weighted approach generated estimates that closely matched confirmed data without significant over- or underestimation. The use of algorithms such as these may dramatically improve the feasibility and efficacy of future research linking environmental data and health outcomes.

Keywords

epidemiology, Internet, methods, technology, tobacco

Introduction

A waterpipe (also called a hookah or narghile) consists of a metal bowl into which flavored tobacco and lit charcoal are placed, a glass body containing water for the smoke to bubble through, and a flexible hose to inhale smoke from the apparatus.[1] Waterpipe tobacco smoking (WTS) is associated with many of the same serious health risks as cigarette use[1–3] and is increasingly popular among university students.[4,5] University students generally perceive

Corresponding author:
Jason B Colditz, The Center for Research on Media, Technology, and Health and School of Medicine, University of Pittsburgh, 230 McKee Place, Suite 600, Pittsburgh, PA 15213, USA.
Email: colditzjb@pitt.edu

WTS—compared with cigarette smoking—as less addictive,[6] more socially acceptable,[7] and more pleasant.[8] While this population is knowledgeable about negative health effects of cigarettes, they underestimate adverse health effects of WTS.[9–11] Secondary exposure to WTS is also a concern, particularly in "hookah bars," which expose patrons to high levels of environmental carbon monoxide.[12–14] Establishments such as these have been gaining popularity in the United States and are particularly prevalent on university campuses.[15] In the United States, up to 41 percent of college-aged WTS takes place at these establishments.[16] Therefore, it is important to examine how and where WTS establishments function, especially near university campuses.

The online platform Yelp (www.yelp.com) may be particularly useful in understanding WTS establishments around university campuses. Yelp is a business directory and social site where individuals can search for and post ratings for public establishments such as restaurants, retailers, or event spaces. Of particular interest to this study, Yelp's categories of "hookah bars" and "tobacco shops" might be used to quickly identify WTS establishments. Patrons' ability to provide text narratives and images related to these establishments provides an opportunity to further characterize and confirm establishments' operational features (e.g. serves alcohol and sells tobacco). Data from Yelp have been used to effectively track unreported foodborne illnesses[17] and monitor consumer perceptions around and proliferation of shops specializing in electronic cigarette paraphernalia.[18,19] Specific to WTS, a 2015 study developed processes for estimating the operating status of establishments that Yelp tagged as "hookah bars" in the state of New York.[20] This study found that Yelp tended to overreport the number of open hookah bars.[20] The researchers devised an alternate method—classifying establishments as open if there had been a Yelp review within the previous 6 months—which tended to underestimate the number of open hookah bars.[20] In a broader study of WTS establishments surrounding 1454 US college campuses, Kates et al.[21] found that student population density was a strong predictor of WTS establishment proximity. WTS establishments cluster more closely around college campuses with denser student populations. The broad search strategy used yielded many false-positive results from Yelp, especially related to specialty tobacco retailers.[21]

Studies such as these have been foundational to building more accurate models of WTS prevalence estimation using primary Yelp data.[20,21] They highlight the potential importance of concurrently examining various Yelp search strategies. Optimization of a Yelp-related search strategy in this way would be extremely valuable for future research around environmental factors and tobacco-related health outcomes. This is because Yelp provides a robust publicly accessible Application Programming Interface (API), which provides data such as precise geolocation and business classification that can be used in rigorous analyses.

The richness of public-facing Yelp data may allow for additional insight into WTS establishments, particularly with understanding WTS establishment characteristics (e.g. serving food, alcohol, and age restrictions). These characteristics are important in forming public health policy, as clean indoor air regulations often hinge on the status of retail food and beverage sales.[22] Additionally, information about age restrictions, live entertainment, and promotional discounts may be important in understanding other appeals of these establishments.

The overarching purpose of this study was to build on previous research that utilized data from Yelp in examining WTS establishments on university campuses. In particular, we sought to achieve three primary aims: (1) advancing current methods of obtaining, parsing, and coding Yelp data; (2) comparing the utility of various Yelp search strategies and developing more accurate predictive models of WTS presence; and (3) examining characteristics of WTS establishments.

Materials and methods

Data collection

Yelp maintains an open API, which provides direct access to selected raw data for establishments listed on the site. We utilized Python 2.7 software to write an application built around the *yelp-python* package.[23] This allowed for reliable access to Yelp's Search API[24] for retrieving an array of public Yelp listings. Metadata included establishment name, link to the online Yelp listing, geolocation, and associated business categories (e.g. hookah bar, tobacco shop, Mediterranean restaurant, and dance club).

We collected all data from Yelp on 1 December 2015. As we were interested in any listing that may have been a WTS establishment, we used the broad search term "hookah" instead of more precise terms (e.g. hookah café) or Yelp's category of "hookah bars." To improve feasibility, we narrowed the scope to "4-year, large, highly residential" campuses, as defined by the Carnegie Classification Size and Setting criteria in the 2014–2015 academic year.[25] University classification data were obtained via the National Center of Education Statistics, Integrated Postsecondary Education Data System (IPEDS).[26] For the 41 matching institutions, we also collected zip codes and GPS campus location data via IPEDS so that we could effectively target our Yelp searches. Yelp API searches were completed using 5-mile search radii, centered on the official zip code for each university. However, to compensate for the imprecision of zip code data, we reduced our search radii to 2 miles using exact geolocations of identified establishments and institutions. A 2-mile radius is consistent with Yelp's "within biking distance" category.

Coding strategy

In order to contextualize WTS establishments, two independent coders followed a structured process of (1) coding establishment characteristics from the Yelp site, including reading reviews and viewing associated images; (2) verifying prior coding against each establishment's official website or social media accounts, where available; and (3) directly calling establishments for clarifications or to obtain missing information. Coders first assessed whether establishments permitted patron WTS on-site (designated as "WTS establishments"). This included, for example, restaurants that had a lounge or patio area designated for WTS but not establishments that were in close proximity to a separate WTS establishment (e.g. providing food to a separate WTS establishment located next door). A random subset of 100 establishments was selected to ascertain interrater reliability for coding WTS status and for coding whether establishments were specialty tobacco shops that sold WTS products or paraphernalia to consumers. For locations that were confirmed as WTS establishments, coders then assessed business status (open vs closed), availability of food (full menu vs limited menu or no food), availability of alcohol (including sales and bring your own), availability of entertainment (e.g. live music, karaoke, and belly dancers), discount availability (e.g. weekly specials and student discounts), and explicit age restrictions. Coding discrepancies were adjudicated in the presence of a supervising researcher. Consistent with previous research, we also recorded the dates of the most recent Yelp reviews of WTS establishments, prior to the date of data collection.

Estimation strategies

We examined approaches for determining whether Yelp metadata alone could predict the number of confirmed WTS establishments among campus locations. To this end, we examined the presence

or absence of categorical labels that are native to Yelp and conceptually relevant to identifying WTS establishments. Establishments can be ascribed up to three of such labels by confirmed owners or Yelp users/reviewers.

In particular, we first noted which establishments were categorized as "hookah bars." We anticipated that this categorization would have a high degree of specificity (i.e. most of these results are in fact WTS establishments) and lower sensitivity (i.e. WTS establishments which are not considered hookah bars will be missed). This was treated as a dichotomous predictor variable. Two additional categories, "tobacco shops" and "vape shops," were considered together as indicators that establishments were likely to be tobacco product retailers. This second dichotomous variable was reverse coded to reflect establishments which were neither tobacco shops nor vape shops. As such, this variable was expected to have relatively high sensitivity and low specificity (i.e. excluding retailers would remove false positives of tobacco retailers but miss true negatives such as restaurants). These two estimation strategies were used to develop two additional hybrid approaches to better estimate the number of confirmed WTS establishments. The third estimation approach was calculated per campus and relied on simple averages of frequencies using the two earlier estimates (see equation (1)). We anticipated that this approach would improve estimation by combining the two earlier approaches which were anticipated to have inverse relationships between sensitivity and specificity (i.e. a middling effect). The fourth approach was similar to the third approach, but it used weighted averages of the first two estimates. Weighting took into account the observed accuracy of two initial estimates (see equation (2)). By controlling for accuracy, we believed that this approach would provide the best estimate of ground truth (i.e. confirmed establishments)

$$n_3 = \frac{n_1 + n_2}{2} \tag{1}$$

$$n_4 = \frac{\left(n_1 * acc_1\right) + \left(n_2 * acc_2\right)}{acc_1 + acc_2} \tag{2}$$

These equations are applied at a university level, so *n*s relate to frequencies of establishments that match the "hookah" search term for each campus. The n_1 variable reflects the number of locations classified by Yelp as hookah bars. The n_2 variable reflects the number of locations not classified by Yelp as tobacco shops or vape shops. The *acc* variables reflect the accuracy of each of these two primary estimation strategies across all university campuses (see Baratloo et al.[27] for further explanation of accuracy calculations).

Statistical analysis

Statistical analyses were performed using Stata 12 software. We utilized frequencies, proportions, and measures of central tendency to contextualize basic WTS establishment characteristics. Interrater reliability of coders was computed using Cohen's kappa statistic to quantify coding agreement beyond that expected by random chance.[28] This allowed us to critically assess the consistency of coders in identifying WTS establishments and tobacco retailers. We calculated specificity, sensitivity, and overall accuracy for the first two estimation strategies (i.e. each WTS establishment predicted vs confirmed status).[27] These values were used both descriptively and to calculate the final model (see equation (2)). Spearman correlations were then used to examine the how all estimation strategies ranked campuses by the number of estimated versus confirmed WTS

Table 1. Descriptive characteristics of confirmed WTS establishments ($n = 257$).

Characteristic	Frequency	%
Alcohol permitted	214	83
Sold	196	76
Bring your own	18	7
Food sold	205	80
Full dinner menu	164	64
Limited menu	41	16
Entertainment	181	70
Discounts	136	53
General	126	49
Student	10	4

WTS: waterpipe tobacco smoking.

establishments. To further assess search strategies, we also calculated z coefficients and significance values using the Wilcoxon signed-rank test, to examine significant over- or underestimation of confirmed WTS establishments across locations. Nonparametric approaches (i.e. Spearman and Wilcoxon) were chosen over parametric approaches as the distribution of WTS establishments per location was not assumed to approximate a normal distribution. The α cutoff was set at 0.05 for all significance tests, as is customary in health science research.[29]

Results

Sample characteristics

Searching within a 2-mile radius of the 41 targeted institutions resulted in 521 Yelp records that matched the "hookah" search term. Nine of the institutions yielded no relevant records, leaving 32 campuses in our sample. Among the locations which yielded Yelp records, there were a mean of 9.0 (standard deviation (SD) = 14.1, median = 3.5, interquartile range (IQR): 1–8, range: 1–65) results per campus. Several institutions were located in metropolitan areas such as New York City and Washington, DC, where overlapping search radii resulted in duplicate establishment listings. After de-duplicating the data, there were 437 unique business listings that matched the "hookah" search term. Of these, 257 (59%) were manually confirmed as establishments where WTS was available for patrons. Of the 180 remaining non-WTS establishments, 79 (44%) were confirmed as specialty tobacco retailers selling WTS paraphernalia. Interrater reliability was excellent for identifying WTS establishments (Cohen's $\kappa = 0.93$) and WTS retailers (Cohen's $\kappa = 0.96$). The 101 remaining false positives in the search results arose from the words "hookah" or "hookahs" appearing tangentially within a Yelp review (e.g. a reviewer mentions going to a hookah bar after visiting the identified establishment and the establishment was noted as neighboring a hookah bar). In some cases, false positives arose from mentions of "hookah pens" or similar paraphernalia relating to electronic nicotine delivery systems. The defining characteristics of confirmed WTS establishments are listed in Table 1.

Alcohol was sold by a significantly lower proportion of establishments that were categorized as "hookah bars" on Yelp (63%) as compared to WTS establishments that were not categorized as such (87%; $\chi^2 = 20.10$, $p < 0.001$). The establishments designated as hookah bars were also less likely to have a full menu (53% vs 73%; $\chi^2 = 10.89$, $p = 0.001$) and less likely to provide live entertainment (62% vs 78%; $\chi^2 = 8.20$, $p = 0.004$). No significant differences were found among age restrictions or availability of discounts.

Table 2. Most recent Yelp reviews per confirmed WTS establishment ($n = 257$) and performance of recency-based cutoff criteria on predicting confirmed open status of establishments.

Recency in months	*n*	Cutoff performance[a]	
		Sensitivity	Specificity
1	139	0.59–0.63	0.65–0.71
2	26	0.71–0.74	0.58–0.69
3	24	0.82–0.85	0.51–0.67
4	14	0.88–0.89	0.43–0.62
5	6	0.90–0.91	0.40–0.60
6	4	0.92–0.93	0.38–0.58
7	7	0.92–0.94	0.32–0.47
8	4	0.94–0.95	0.31–0.44
9	3	0.95–0.95	0.27–0.42
10	1	0.95–0.96	0.27–0.42
11	5	0.97–0.97	0.23–0.40
12	2	0.97–0.97	0.21–0.36
>12	22	1.00	0.00

WTS: waterpipe tobacco smoking.

[a]Cutoffs are inclusive of all reviews predating the recency date. Ranges indicate values when treating all establishments of ambiguous operating status ($n = 36$) as closed (resulting in higher sensitivity and lower specificity) versus treating them as open (resulting in lower sensitivity and higher specificity).

Assessing open status

Of the 257 confirmed WTS establishments, 108 (42%) were confirmed as open based on phone calls (primary criteria) and 68 (26%) were confirmed as open based on having updated their websites or social media accounts within the past 6 months (secondary criteria). An additional 36 (14%) were considered ambiguous due to Yelp user activity in the past 6 months but no other indicators of being open and 45 (18%) were coded as closed based on all criteria. Of the 176 WTS establishments confirmed as open by primary and secondary criteria, 13 (7%) had most recent Yelp reviews that predated our data collection by more than 6 months. Thus, if we had used "any reviews in the past 6-months" criteria, we would have measurement sensitivity of 0.93 in predicting confirmed open status of establishments. However, if we assume that all unconfirmed locations were in fact closed, the 6-month criterion misclassifies 50 (19%) establishments as open. Thus, the lowest possible specificity using a 6-month criterion was 0.38. Using a less conservative approach, whereby ambiguous cases were treated as open, specificity would have reached 0.58. Table 2 displays sensitivity and specificity using incremental month of most recent review as predictive criteria for confirmed open status of WTS establishments.

Estimation strategy performance

Our first estimation strategy utilized Yelp's "hookah bars" designation, as has been used as a primary search strategy in previous research.[20] Of the 257 confirmed WTS establishments, 117 (46%) were categorized as hookah bars on Yelp. This strategy significantly underestimated WTS establishment prevalence among campuses (Wilcoxon $z = -4.65$, $p < 0.001$) as it lacked sensitivity (0.45), though specificity was high (0.97). Overall accuracy was 0.67. The estimates produced by this approach were highly correlated with the number of confirmed WTS establishments across the identified campuses (Spearman $\rho = 0.91$). See the top-left panel of Figure 1.

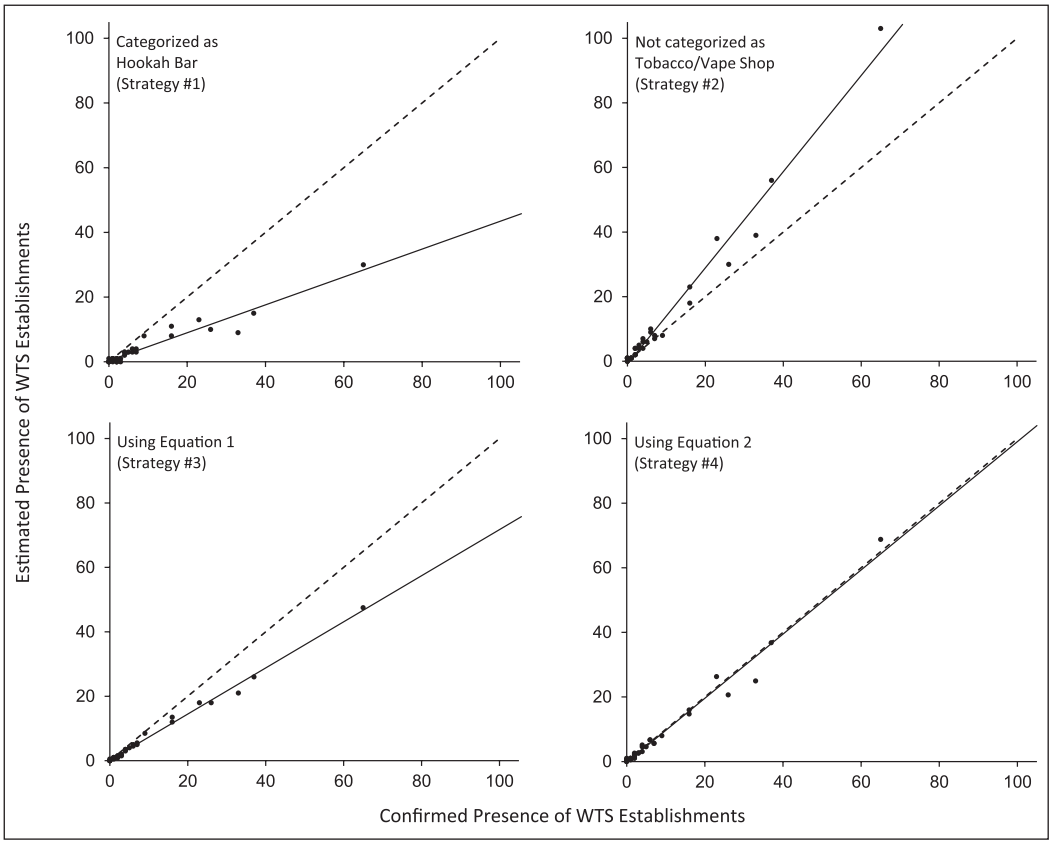

Figure 1. Comparisons of estimated versus confirmed presence of WTS establishments.
The *y*-axis indicates the estimated frequency of WTS establishments per location using each of the four estimation strategies. The *x*-axis indicates the frequency of WTS establishments confirmed by human coders. The dashed line demonstrates a slope of perfect agreement ($y = x$) between estimated and confirmed WTS establishment frequency. The solid lines represent linear slopes of observed results for each estimation strategy, to illustrate over- and underestimation.

The second search strategy subtracted the number of establishments that Yelp categorized as "tobacco shops" or "vape shops" from the overall search results. Of the 85 confirmed specialty tobacco retailers, 69 (81%) were correctly categorized as such on Yelp. This strategy significantly overestimated the number of WTS establishments among campuses (Wilcoxon $z = 4.04$, $p < 0.001$) as it lacked specificity (0.46), though it demonstrated high sensitivity (0.98). Overall accuracy was 0.76. Estimates produced by this approach had a slightly stronger correlation with the number of observed WTS establishments (Spearman $\rho = 0.98$), as compared to the first strategy. See the top-right panel of Figure 1.

The third strategy estimated the number of WTS establishments using the simple mean of the two earlier estimation strategies (see equation (1)). This strategy demonstrated nearly identical underestimation as the first approach when tested using the Wilcoxon method ($z = -4.16$, $p < 0.001$) and a similar correlation to that of the second method (Spearman $\rho = 0.98$). See the bottom-left panel of Figure 1.

The fourth strategy estimated WTS establishment presence using a composite score of the first two search strategies, weighted by our observed search accuracies (see equation (2)). This approach was able to estimate the approximate number of confirmed WTS establishments among college

campuses without statistically significant differences among estimated and confirmed WTS status (Wilcoxon $z = 0.17$, $p = 0.87$). The weighted estimation demonstrated similarly strong correlations as the previous strategies (Spearman $\rho = 0.97$). See the bottom-right panel of Figure 1.

In short, all strategies performed well when comparing campuses ranked by estimated number of WTS establishments versus ranking by confirmed WTS establishments (i.e. Spearman correlation). Our final strategy (weighted by search accuracy) was superior to all others when estimating the actual number of confirmed WTS establishments per campus (i.e. Wilcoxon signed-rank test). Filling in the observed accuracy parameters from the first two approaches, this optimal estimation approach is demonstrated in equation (3)

$$n_4 = \frac{(n_1 * 0.67) + (n_2 * 0.76)}{1.43} \tag{3}$$

Equation (3) is the simplified form of equation (2), with calculated accuracy parameters provided. These parameters are based on our sample of university campuses, and parameter adjustments should be considered if applying this approach to other search contexts (e.g. metropolitan areas).

Discussion

Previous research has utilized Yelp data as a primary data source for establishing the prevalence of WTS establishments in US localities. However, as noted in these studies, limitations included the unknown accuracy of various search strategies.[21] While using the "hookah" search term resulted in extraneous information, relying on Yelp's "hookah bar" categorization missed relevant WTS establishments. The hookah bar classification, while highly specific, was not sensitive enough to accurately estimate the number of WTS establishments. Categorization on Yelp may reflect a more general social concept of what constitutes a hookah bar, as opposed to our health-oriented concept of whether or not an establishment permits WTS. In our data, establishments that were categorized as hookah bars were substantially fewer in number than confirmed WTS establishments. Hookah bars were also significantly less likely to sell alcohol, provide a full menu, or have live entertainment. As such, we conclude that hookah bars are both quantitatively and qualitatively distinct from the broader concept of WTS establishments. These distinctions are important, as clean indoor air policies often hinge on business classifications, particularly relating to characteristics such as food and alcohol sales.[22]

A primary aim of the current research was to test various strategies of using Yelp metadata to estimate the number of confirmed WTS establishments by geographic location. All strategies performed reasonably well across campus locations when looking at ranked correlations (i.e. Spearman ρ). As such, any estimation approach may be adequate when the goal is to rank order locations by number of proximal WTS establishments. However, if the goal is to estimate the number of WTS establishments surrounding a location, then using our fourth approach would be more appropriate. This was the only approach that did not systematically over- or underestimate the number of confirmed WTS establishments. Data from this approach would be particularly useful to include in epidemiologic studies of WTS use. To our knowledge, this is the first systematized approach able to accurately predict ground truth of localized WTS establishments, while relying primarily on publicly available metadata.

There are immediate opportunities to utilize the current approaches to model prevalence of WTS establishments in a broader epidemiologic scope. This would allow for the inclusion of WTS establishment prevalence estimates when analyzing secondary data about WTS use patterns (e.g.

National College Health Association data, nationally representative surveys). To date, such estimates of WTS prevalence have been difficult to obtain. However, these data would help to control for this important environmental factor (i.e. local availability of WTS) in more complex social-ecological frameworks of WTS use.

Future studies within the field health informatics may benefit from incorporating data from the Yelp API and accuracy-weighted estimates into other novel frameworks. For example, future work may include bars or alcohol retailers around college campuses, the emergence of vape shops and vape lounges, or other types of establishments that may contribute to alcohol or nicotine exposure. Additionally, it may be worthwhile to examine prevalence of establishments that may encourage positive health such as gymnasiums, parks, health clinics, or grocers. These types of data make it increasingly possible to model geographic and community-level contributors that may impact health outcomes. Such approaches present new opportunities for the field of health informatics as a whole.

Limitations

As our primary data were derived from the Yelp platform, we cannot account for WTS establishments that are not listed on Yelp. This may result in slightly conservative estimates of WTS prevalence. However, Yelp tends to be one of the highest-yielding platforms for WTS establishment records,[21] and the availability of its API makes it an ideal venue for large-scale metadata extraction. Additionally, our focus on categorically large, highly residential campus locations has implications for external validity. We acknowledge that WTS establishments may be qualitatively different in areas with smaller populations of residential students. Yet, even the campuses from our sample yielded diverse numbers of WTS establishments, with many campuses having a relatively small number of establishments. As such, we believe that it would be reasonable to use the proposed estimation strategy to assess other US residential areas. Finally, we acknowledge that we were unable to replicate the high level of sensitivity (0.90) reported in previous research that used 6-month cutoff criteria to predict WTS establishment operating status (see Table 2).[20] Additional investigations will be required in order to further assess the viability of such an approach.

Conclusion

WTS is an important public health issue, particularly in college student populations. Recent research has begun to monitor WTS proliferation using data derived from the online Yelp platform. However, establishing ground truth for these data can be a time-consuming and arduous process. As such, we developed a new approach for using publicly accessible metadata from the online Yelp platform to provide accurate and expedient estimates of WTS establishment prevalence. Information from this approach may be of great use in epidemiologic studies that focus on rates of WTS use, which have previously been unable to quantify WTS establishment presence at scale. The present research also provides a framework for understanding important considerations for understanding "hookah bars" and other types of WTS establishments. In particular, we recognize that (1) WTS establishments extend beyond the realm of hookah bars, including restaurants and other types of establishments with unique characteristics; (2) as compared to other WTS establishments, hookah bars are significantly less likely to serve alcohol or food or provide live entertainment; and (3) these unique characteristics provide opportunities to further consider policy interventions that may impact WTS establishment proliferation. For example, policymakers may wish to consider clearly defining and setting minimum standards that apply differentially to different types of WTS establishments. This may include limiting WTS at establishments that are primarily restaurants or bars,

minimizing specials and discounts of WTS products/services, and setting minimum standards for air quality across various types of WTS establishments. Future studies should focus on refining and validating semi-autonomous approaches for obtaining and screening publicly available data to track and characterize WTS establishments. This would allow for not only better-informed epidemiologic studies but also timely and well-tailored policy interventions.

Acknowledgements

We would like to acknowledge Erica Barrett and Daria Williams, who assisted with coding and preliminary analysis of data. We also thank Michelle Woods for editorial assistance.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the National Cancer Institute (NCI) under grant R01-CA140150.

Supplementary materials

Code related to this project can be found at https://github.com/CRMTH/YelpSearch/

ORCID iD

Jason B Colditz https://orcid.org/0000-0002-2811-841X

References

1. Maziak W. The waterpipe: an emerging global risk for cancer. *Cancer Epidemiol* 2013; 37: 1–4.
2. Rakower J and Fatal B. Study of narghile smoking in relation to cancer of the lung. *Br J Cancer* 1962; 16: 1–6.
3. Munshi T, Heckman CJ and Darlow S. Association between tobacco waterpipe smoking and head and neck conditions: a systematic. *J Am Dent Assoc* 2015; 146: 760–766.
4. Primack BA, Shensa A, Kim KH, et al. Waterpipe smoking among U.S. university students. *Nicotine Tob Res* 2013; 15: 29–35.
5. Barnett TE, Smith T, He Y, et al. Evidence of emerging hookah use among university students: a cross-sectional comparison between hookah and cigarette use. *BMC Public Health* 2013; 13: 302.
6. Primack BA, Sidani JE, Agarwal AA, et al. Prevalence of and associations with waterpipe tobacco smoking among U.S. university students. *Ann Behav Med* 2008; 36: 81–86.
7. Smith-Simone SY, Curbow BA and Stillman FA. Differing psychosocial risk profiles of college freshmen waterpipe, cigar, and cigarette smokers. *Addict Behav* 2008; 33: 1619–1624.
8. Maziak W. The global epidemic of waterpipe smoking. *Addict Behav* 2011; 36: 1–5.
9. Sidani JE, Shensa A, Shiffman S, et al. Behavioral associations with waterpipe tobacco smoking dependence among US young adults. *Addiction* 2016; 111: 351–359.
10. El-Zaatari ZM, Chami HA and Zaatari GS. Health effects associated with waterpipe smoking. *Tob Control* 2015; 24(Suppl. 1): i31–i43.
11. Nuzzo E, Shensa A, Kim KH, et al. Associations between hookah tobacco smoking knowledge and hookah smoking behavior among US college students. *Health Educ Res* 2013; 28: 92–100.
12. Cobb CO, Vansickel AR, Blank MD, et al. Indoor air quality in Virginia waterpipe cafes. *Tob Control* 2013; 22: 338–343.

13. Barnett TE, Curbow BA, Soule EK, et al. Carbon monoxide levels among patrons of hookah cafes. *Am J Prev Med* 2011; 40: 324–328.
14. Martinasek MP, Ward KD and Calvanese AV. Change in carbon monoxide exposure among waterpipe bar patrons. *Nicotine Tob Res* 2014; 16: 1014–1019.
15. Eissenberg TE, Ward KD, Smith-Simone S, et al. Waterpipe tobacco smoking on a U.S. college campus: prevalence and correlates. *J Adolesc Heal* 2008; 42: 526–529.
16. Cobb CO, Ward KD, Maziak W, et al. Waterpipe tobacco smoking: an emerging health crisis in the United States. *Am J Health Behav* 2010; 34: 275–285.
17. Harrison C, Jorder M, Stern H, et al. Using online reviews by restaurant patrons to identify unreported cases of foodborne illness—New York City, 2012-2013. *MMWR-Morbid Mortal W* 2014; 63: 441–445.
18. Dai H and Hao J. Geographic density and proximity of vape shops to colleges in the USA. *Tob Control* 2017; 26: 379–385.
19. Sussman S, Garcia R, Cruz T, et al. Consumers' perceptions of vape shops in Southern California: an analysis of online Yelp reviews. *Tob Induc Dis* 2014; 12: 22.
20. Cawkwell PB, Lee L, Weitzman M, et al. Tracking hookah bars in New York: utilizing Yelp as a powerful public health tool. *JMIR Public Health Surveill* 2015; 1: e19.
21. Kates FR, Salloum RG, Thrasher JF, et al. Geographic proximity of waterpipe smoking establishments to colleges in the U.S. *Am J Prev Med* 2016; 50: e9–e14.
22. Colditz JB, Ton JN, James AE, et al. Toward effective water pipe tobacco control policy in the United States: synthesis of federal, state, and local policy texts. *Am J Heal Promot* 2017; 31: 302–309.
23. Mitton K. Yelp-python: a Python library for the Yelp API, https://github.com/Yelp/yelp-python (2016).
24. Yelp Documentation: Search API version 2, https://www.yelp.com/developers/documentation/v2/search_api (2016).
25. Carnegie Foundation for the Advancement of Teaching. The Carnegie Classification of Institutions of Higher Education, http://carnegieclassifications.iu.edu (2011).
26. National Center for Education Statistics. Integrated postsecondary education data system: compare institutions, https://nces.ed.gov/ipeds/datacenter/login.aspx (2015)
27. Baratloo A, Hosseini M, Negida A, et al. Part 1: simple definition and calculation of accuracy, sensitivity and specificity. *Emerg* 2015; 3: 48–49.
28. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas* 1960; 20: 37–46.
29. Schumm WR, Pratt KK, Hartenstein JL, et al. Determining statistical significance (alpha) and reporting statistical trends: controversies, issues, and facts. *Compr Psychol*. Epub ahead of print 1 January 2013. DOI: 102466/03.CP.2.10.