

# Assessing smoking-related behaviours using massive online search query data

Transactions in Urban Data, Science,  
and Technology  
1–12

© The Author(s) 2023

Article reuse guidelines:

[sagepub.com/journals-permissions](http://sagepub.com/journals-permissions)

DOI: 10.1177/27541231231184591

[journals.sagepub.com/home/tus](http://journals.sagepub.com/home/tus)



**Ting Zhou\***

**Long Chen\***

School of Architecture, Tsinghua University, China

**Zhaoxi Zhang**

Department of Environmental Science, Aarhus University, Denmark

**Zhengying Liu**

School of Architecture, Tsinghua University, China

**Ying Long**

School of Architecture, Tsinghua University, China

Hang Lung Center for Real Estate, Tsinghua University, China

Key Laboratory of Eco Planning & Green Building, Ministry of Education, China

## Abstract

Online search queries have been used in various behaviour studies. As smoking has become a global health issue, studies that assess smoking-related behaviours using online search queries have faced limitations in data utilization and study design. This study conducts comparative analyses to investigate changes in smoking-related behaviours represented by online search volumes. Baidu search queries from 2013 to 2017 were used to examine the search volume containing four groups of smoking-related search keywords. A validation process was used to validate the proposed method by comparing changes in search queries with the observed tobacco consumption. The results show changes in smoking-related behaviours assessed by online search queries at the city level in China. The validation experiments illustrate the consistency between changes in search volumes and tobacco consumption. Thus, online search queries were verified to be an effective instrument for assessing smoking-related behaviours, and this study sheds light on broader behaviour studies and policy assessments.

## Keywords

Online search query, big data, smoking, tobacco control, behaviour study

---

\*These authors contribute equally to this work.

## Corresponding author:

Ying Long, School of Architecture, Tsinghua University, 1 Qinghuayuan, Beijing, 100084, China.

Email: [yulong@tsinghua.edu.cn](mailto:yulong@tsinghua.edu.cn)

The tobacco epidemic is the leading cause of preventable deaths globally (Danaei et al., 2009). The World Health Organization (WHO, 2019) has reported that more than 7 million preventable deaths are the result of direct tobacco use annually, while approximately 1.2 million non-smokers die due to second-hand smoke exposure each year. Considering the devastating consequences of tobacco use on public health, an increasing number of studies have paid attention to smoking-related behaviours, especially to examine the risk factors for cigarette smoking and to encourage behaviour changes, such as smoking cessation (Brewer et al., 2016; Christensen et al., 2014; Haardörfer et al., 2018; Lawless et al., 2015). Moreover, more than 132 economies have increased the level of attention given to the harm caused by tobacco use and have implemented various levels of tobacco control policies (WHO, 2019). The timely and accurate monitoring of smoking-related behaviours is imperative for policy development to reduce smoking prevalence. However, traditional surveillance techniques for tobacco use, such as online surveys, self-reporting, self-monitoring, and clinical trials, are inadequate to identify timely temporal and spatial health trends for public health, as well as surveillance for a very large population. In many countries, annual representative individual or household-level surveys of tobacco-related attitudes, beliefs, and behaviours are largely unavailable, and there are also considerable time lags between available data sources and the rapidly changing landscape of smoking and tobacco control.

With the advancement of the Internet and the development of big data, search query data from Internet search engines have become an attractive and promising alternative to traditional surveillance systems. The 2018 Global Digital Report from We Are Social and Hootsuite (2018) revealed that the world's Internet users now number more than 4 billion. Using web search engines to obtain information has become one of the most frequent online activities (Singer et al., 2015). Online search query data generated from online search engines, such as Google and Baidu, have been used to monitor and predict infectious diseases, such as influenza epidemics (Gahr et al., 2015; Ginsberg et al., 2009) and dengue fever (Li et al., 2017; Liu et al., 2016), to nowcast private consumption (Kholodilin et al., 2010), and to forecast tourism inflows and demands (Artola et al., 2015; Pan et al., 2012).

Although limited, online search query data have also been employed in smoking-related behaviour studies, such as tracking the popularity of electronic cigarettes (Ayers et al., 2011b; Paek et al., 2020) and water-pipe tobacco (Salloum et al., 2015), monitoring tax avoidance and smoking cessation after the tax increases (Ayers et al., 2011a), and assessing the impacts of tobacco control policies (Huang et al., 2013; Troelstra et al., 2016). However, most previous studies on smoking-related behaviours using online search queries have faced limitations in both data utilization and study design. Specifically, the search queries of these studies tend to contain a limited set of search keywords, which may overlook diversified smoking-related behaviours due to personal search habits. For example, Huang et al. (2013) used three key search terms, namely, "smoking ban(s)", "electronic cigarette(s)", and "quit smoking", along with news coverage to assess the impact of the National Smoking Ban. Troelstra et al. (2016) only used the term "quit smoking" on Google Trends to study the effect of tobacco control policies on information seeking for smoking cessation. Moreover, such studies lack the validation of their proposed method and data, which potentially reduces the validity of the research.

By introducing online search query data, this article aims to propose an alternative tool for large-scale behavioural analysis with new, big data. Taking smoking-related behaviours as an example, this study adopted massive online search queries data from Baidu to track the behavioural changes in smoking. The proposed alternative tool is able to fulfill the literature gaps as (1) the Baidu search query data contains a massive amount of Internet users' daily search activities, which generates valuable data to analyse the behaviours of a large population; (2) a more comprehensive list of search keywords is compiled to capture smoking-related behaviours; (3) observed tobacco consumption data from statistical agencies are used to validate the online search queries.

A major premise of the tool is that searching for smoking-related keywords in an online search engine implies that the user requires such information, which may further induce the corresponding actions in real life. To elaborate on how the tool works, we used the search volume data to represent smoking-related behaviours and examined the behavioural change between 2013 and 2017. After revealing the behavioural changes in smoking, we compared them with the tobacco consumption data from local statistical agencies to validate our assumptions and the tool.

## Methods

### Data

China's Internet users numbered 772 million in 2017, and that number is still increasing (China Internet Network Information Center, 2018). More than 95% of Chinese Internet users have made Baidu their first choice for obtaining information among all the available search engines. Collaboration with Baidu enables us to obtain search volume data generated from massive search queries submitted on [www.baidu.com](http://www.baidu.com). Notably, although similar search query data are publicly accessible and can be downloaded from the Baidu Index (<http://index.baidu.com>), the available keywords from the Baidu Index are extremely limited because the site filters the keywords with relatively lower search volumes (Baidu Index value that is lower than 50).

Before proceeding with the data processing and analysis, we defined several key concepts that will be used in this study:

- (1) Search query: A search query is a phrase or a keyword combination users enter in search engines to find a particular content or information. It represents the search behaviour and refers to the willingness to engage in certain smoking-related behaviours in this study.
- (2) Search keyword: A search keyword refers to a word or group of words that are usually associated with particular content or topic. A keyword is an abstraction that we extracted from multiple search queries related to smoking behaviours.
- (3) Search volume: The search volume indicates the number of search queries for a specific search keyword in a search engine, such as Baidu in this case, within a given timeframe.

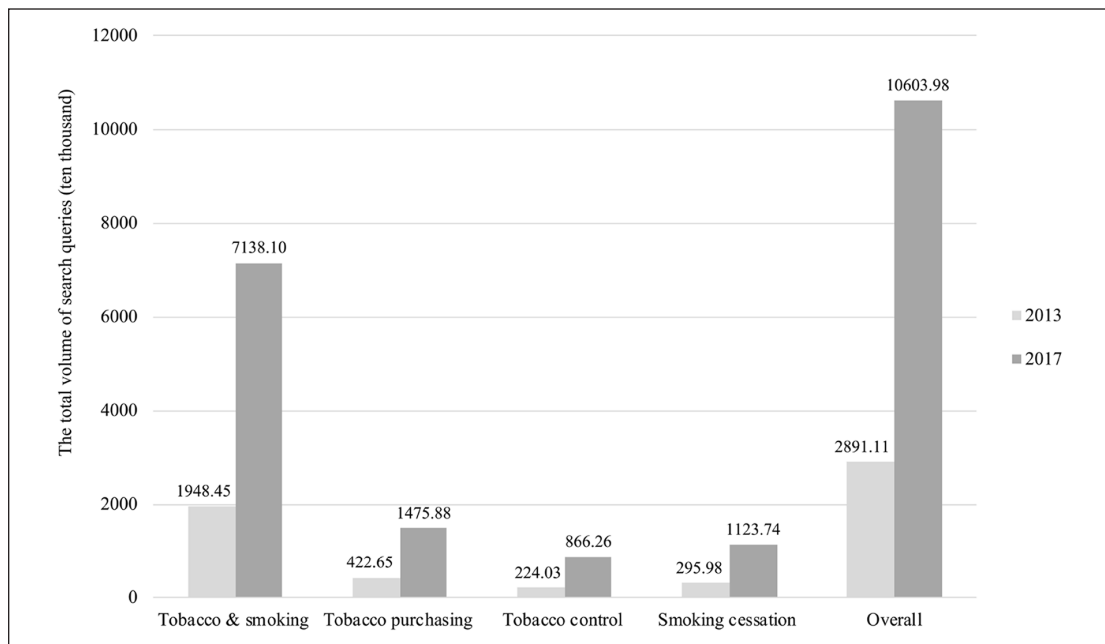
To comprehensively describe smoking-related behaviours, this study complies with a search keyword list that contains a total of 954 various keywords, based on a systematic literature review, interviews with smokers, and a digging for Baidu's search connection and recommendation algorithms. The original annual search volumes for the 954 search keywords in both 2013 and 2017 were collected at the city scale, which represent normalized numbers of unique searches containing these keywords.

The 954 search keywords are classified into four groups based on the connotation of the keywords and the corresponding behaviours associated with them (Table 1). In each keyword group, the search volumes were summed up and further divided by the average annual number of Internet users in each city to generate per capita search volume of the smoking-related keywords. For example, search keywords related to tobacco and smoking, such as brands of cigarettes and how to start smoking, capture an overall interest in tobacco use. The tobacco purchasing group encompasses a wide range of search keywords related to cigarette prices and vendors. Tobacco control search keywords include news on tobacco control policies and smoking ban regulations. The smoking cessation group encompasses keywords related to smoking cessation methods and products, which provides indicative evidence of quit intentions and quit attempts among smokers who use Baidu to obtain quit smoking-related information.

Figure 1 provides an overview of the total search volumes of each group of smoking-related search queries. Specifically, in both 2013 and 2017, the tobacco and smoking search queries made up the largest share

**Table 1.** Smoking-related search keywords identified in this study.

Group of search keywords	Number of search keywords	Examples of search keywords	Description of search keywords
Tobacco and smoking search keywords	561	丽水烟草 (Lishui Tobacco), 双喜世纪经典 (Shuangxi Century Classics), 女性香烟 (female cigarettes), 抽烟手势 (smoking gestures), 教你如何科学地抽烟 (how to smoke scientifically)	Brand of cigarette, smoking tutorials, female tobacco brands, smoking gestures, smoking methods
Tobacco purchasing search keywords	120	烟草网上订货 (online tobacco ordering), 买烟去哪个网站 (websites to buy cigarettes), 合法买烟 (buying cigarettes legally)	Purchasing channels, vendors, prices
Tobacco control search keywords	154	无烟办公室 (smoke-free offices), 政府禁烟措施 (tobacco control policies), 禁烟活动方案 (tobacco control programmes), 控烟活动 (tobacco control activities), 无烟城市 (smoke-free cities)	Smoking bans, tobacco control policies, regulations, and penalties and smoke-free cities
Smoking cessation search keywords	119	有氧运动戒烟 (aerobic exercises to help quit smoking), 消除烟瘾 (eliminating smoking addiction), 轻松戒烟 (how to quit smoking easily), 有效的戒烟方法 (effective ways to quit smoking)	Cessation methods, institutions to help one quit smoking, treatment, e-cigarettes

**Figure 1.** The search volume of different search queries in 2013 and 2017.

of the total volume, as the group contains more than 60% of the keywords. The total search volumes of each group witnessed various degrees of increase from 2013 to 2017, which may indicate different increases in different smoking-related behaviours. Notably, the total volume of search queries reached over 1.06 billion in 2017, more than three times that in 2013, which may also be attributed to the ever-increasing number of

Internet users and online search activities. As tobacco use is harmful to general health, we define tobacco and smoking and tobacco purchasing queries as “negative” queries since they are potentially associated with behaviours that may result in smoking. However, tobacco control and smoking cessation queries are defined as “positive” groups due to their association with less smoking or cessation.

### *Study design*

The unit of analysis used in the present study consisted of administrative cities in mainland China, and the search volumes of the queries were aggregated at the city scale by the geotags embedded in the search activities. A total of 343 administrative cities were involved in the comparative analysis to examine changes in the search queries and the behaviours associated with them. In this study, smoking-related behaviours are calculated as the annual search volumes of the smoking-related search keywords divided by the number of Baidu users. To track the changes in behaviours between 2013 and 2017, a variable is defined as behaviour change and calculated using the difference of the search volumes between 2017 and 2013, and dividing by the search volume in 2013.

In addition, to explore how smoking-related behaviours vary among cities with different tobacco control policies, three types of cities were identified according to the tobacco control policies implemented by the local governments.<sup>1</sup> In Type A cities, which are represented by Beijing and Shanghai, smoking is strictly prohibited in all public places, indoor areas of the workplace, and on public transport. In addition, the outdoor areas of schools, cultural relics protection areas, stadiums and healthcare facilities are also designated as no-smoking zones. While in Type B cities, such as Dalian and Hohhot, smoking is not allowed in specific indoor areas, including stadiums, libraries, museums, stores larger than 200 square meters, the waiting rooms, consultation rooms, and wards in hospitals, as well as the classrooms and corridors in schools. Type C cities have no tobacco control policies enacted at the current time. The introduction of tobacco control policies in the analysis has at least two-fold meanings: first, as a policy intervention, the tobacco control policies would significantly influence smokers’ behaviours, and thus left noticeable marks in both search volume and consumption data, which helps us explain the changes in the data and supports the validation of our tool as well; second, it is our preliminary attempt to conduct a policy assessment using online search queries data, and this is also one of the major contributions of this paper, which is to advocate the use of online search queries data in behavioural studies and policy assessment, not only in tobacco use, but also in other fields that require massive behaviour surveys.

Tobacco consumption is an indicator that may represent the observed smoking-related behaviours, especially the changing trends. Such data are used to validate our data and method regarding proxying behaviours using online search queries data. In China, tobacco consumption data are collected at municipal level, but there is no statistical standard for these data among cities. For example, some cities use “total sales or total retail sales of tobacco products”, some cities use “taxes and other charges on principal business”, while others use “transaction amount, profit or output value”. One common attribute of these different statistics is that they can be used as proxies for tobacco consumption. Both tobacco consumption data and demographic data of the sample cities were collected from municipal statistical yearbooks, which are publicly available on the official Municipal Bureau Statistics website.

To validate our method and data, we first compared the change in smoking-related behaviours, which is measured by the search volume of the queries, with the tobacco consumption in each sample city. In addition, different proxies of tobacco consumption were further standardized to explore the heterogeneity between different types of cities, and sample cities with and without tobacco control policies were selected to explore how online search activities and tobacco consumption were influenced by policy interventions.

Only 17 cities were found to have publicly accessible tobacco consumption data, either in their yearbooks or other statistical datasets. Therefore, we could only use such limited observed data to validate our results. Type A cities include Beijing, Shanghai, and Shenzhen, and Type B cities include Nanning, Hohhot, Lanzhou, Nanjing, Hangzhou, Guangzhou, Changchun, Shijiazhuang, Fuzhou, and Dalian, while Type C cities include

**Table 2.** Smoking-related queries in different types of cities in China.

Type of cities	Year	Tobacco and smoking	Tobacco purchasing	Tobacco control	Smoking cessation	All queries
Type A (Cities with the strictest tobacco control policies)	2013	0.54	0.11	0.06	0.09	0.82
	2017 (adjusted)	0.61	0.13	0.09	0.13	0.96
	Change	0.13	0.11	<b>0.43</b>	<b>0.40</b>	0.18
Type B (Cities with slack tobacco control policies)	2013	0.55	0.12	0.06	0.09	0.82
	2017 (adjusted)	0.57	0.12	0.07	0.10	0.85
	Change	0.03	0.01	<b>0.11</b>	<b>0.11</b>	0.04
Type C (Cities without tobacco control policies)	2013	0.57	0.13	0.07	0.08	0.84
	2017 (adjusted)	0.57	0.12	0.07	0.08	0.83
	Change	0.00	-0.03	0.00	-0.03	-0.01
All cities	2013	0.57	0.13	0.07	0.08	0.83
	2017 (adjusted)	0.57	0.12	0.08	0.08	0.84
	Change	0.00	-0.03	<b>0.11</b>	-0.02	0.01

Correction factor=3% (The 2017 search query data were adjusted to minimize the influences from the change in Internet users' search behaviours. The correction factor is calculated as 3% considering that there was a 3% increase in per capita search volume in Baidu from 2013 to 2017.) The adjusted 2017 search queries=The original 2017 search queries/(1 + correction factor). The bold font emphasizes the changes in "positive" smoking-related queries that are more than 10%.

Chengdu, Chongqing, Suzhou, and Linyi. The tobacco consumption is also calculated at per capita basis for comparison purpose.

## Results

### *Changes in smoking-related behaviours*

Table 2 displays smoking-related behaviours calculated in the four groups of smoking-related keywords at the aggregate level of all cities in China in both 2013 and 2017, as well as the change between the two years. Specifically, we mainly focused on changes that are more than 10%, which is the second quartile of all the changes in each city. For the total condition, although the "negative" smoking-related behaviours given to the "negative" search keywords (tobacco and smoking and tobacco purchasing) remained higher than the "positive" keywords (tobacco control and smoking cessation) in 2017, tobacco control was the only group that increased significantly; the changes in other groups of smoking-related queries did not manifest as strongly as the overall search volume. On the one hand, Baidu search users remained interested in tobacco and smoking and tobacco purchasing; on the other hand, their awareness of tobacco control and legislation evidently increased from 2013 to 2017.

Figure A-1 in the Supplemental material gives a full picture of the spatial distribution of the changes related to smoking-related queries. For the "negative" queries, most of the cities with an evidently increased search volume were located in the southeastern and southwestern areas of China, including the cities of Hangzhou, Shenzhen, Shaoyang, and Chongqing, with a few cities positioned in the central and northern parts of China, including Lanzhou and Beijing. Additionally, the cities where tobacco and smoking search volume remained stable numbered more than their tobacco purchasing search volume counterparts, which were mainly located in the western and northeastern areas of China. With regard to the "positive"

smoking-related queries, cities with increased search volume given to tobacco control or smoking cessation were mainly situated in the southeastern, southwestern, and central areas of China. In Western China, many cities witnessed a spike in tobacco control search volume but a decrease in smoking cessation search volume. Notably, compared with the spatial distribution of changes in “negative” smoking-related queries, more cities showed increases in the changes in “positive” queries, especially in the western and central areas of China. Regarding overall search volume, most cities remained stable. Specifically, cities with an increased search volume were mainly located on the southeastern coast and central and western areas of China, while most of the northern and western areas of China experienced a drop in the overall search volume.

Based on previous studies, geographical context exerts an obvious effect on smoking-related behaviours. At the national scale (between cities), the prevalence of smoking may be affected by various policies related to tobacco taxation (Callison and Kaestner, 2014), the advertising of tobacco products (Hanewinkel et al., 2010), and intervention policies such as “smoking bans” (Catalano and Gilleskie, 2021). At the local level (within cities), area-based housing improvement (Bond et al., 2013), social capital and cohesion (Andrews et al., 2014), neighbourhood crime (Shareck and Ellaway, 2011), and population structure such as age and gender (Grøtvedt and Stavem, 2005) have influenced smoking behaviour evidently.

Although we tried to summarize a spatial pattern among cities, we noticed that smoking-related behaviours, either in the positive direction (tobacco control and tobacco cessation) or in the negative (smoking and tobacco purchasing) do not show significant spatial autocorrelation. However, what we found in the spatial distributions is that cities experienced increased attention to tobacco control and smoking cessation covers most Type A and B cities (cities with strict or moderate tobacco control policies) and these cities could also be found to have increased overall attention. Besides, those cities that experienced a decrease in overall attention are mainly concentrated in Type C cities. Our findings are consistent with the literature at the national level, but due to the lack of data at a finer scale, how smoking behaviours distributed within each city is not included in this study.

### *Changes in smoking-related behaviours in different types of cities in China*

Table 2 displays per capita search volumes calculated in the four groups of smoking-related search keywords at the aggregated level of three types of cities in both 2013 and 2017, as well as the changes between the two years. As Table 2 shows, all groups of smoking-related queries and the overall search volume showed various degrees of increase from 2013 to 2017 in Type A cities. In particular, the change in tobacco control search volume was the most evident, which was followed by the change in smoking cessation search volume. For Type B cities, “positive” smoking-related queries increased noticeably, whereas their “negative” counterparts and the level of overall search volume remained stable. For Type C cities, all types of smoking-related queries remained stable.

Notably, tobacco control policies exerted more impact on “positive” smoking-related queries, which include tobacco control and smoking cessation search volume. Although “positive” smoking-related search volumes increased markedly in both Type A and Type B cities, the changes related to these queries were much more obvious in Type A cities. These results illustrated that in cities with more stringent tobacco control policies, Baidu search users’ interest in tobacco control and smoking cessation increased. In addition, their awareness of related legislation and quitting smoking was greatly influenced by the intensity of tobacco control policies.


## **Validation results**

### *Comparison of online search query data and municipal statistics*

Figure A-2 in the Supplemental material illustrates the comparison of smoking-related behaviours assessed by online search query data and the municipal statistics for each city, and Table 3 provides a summary of the

**Table 3.** The validation results between online search volume and consumption.

City		Volume of positive queries	Volume of negative queries	Tobacco consumption	Validation result
Type A (Cities with the strictest tobacco control policies)	<b>Beijing</b>				<b>Y</b>
	Shanghai				<b>Y</b>
	Shenzhen				<b>Y</b>
Type B (Cities with slack tobacco control policies)	Guangzhou				<b>Y</b>
	Changchun				<b>Y</b>
	Shijiazhuang				<b>Y</b>
	Fuzhou				<b>Y</b>
	Dalian				<b>Y</b>
	Nanning				<b>Y</b>
	Hangzhou				<b>Y</b>
	Hohhot				<b>N</b>
	Lanzhou				<b>N</b>
	Nanjing				<b>N</b>
Type C (Cities without tobacco control policies)	Chengdu				<b>Y</b>
	Suzhou				<b>Y</b>
	Chongqing				<b>Y</b>
	Linyi				<b>N</b>

 means an increase in the search queries or consumption,  means a decrease. For smoking-related queries, the darker the colour, the more the behaviour changed.

validation result. We expected that changes in online search volumes would be associated with actual tobacco consumption, for example that a more evident increase in the volume of “positive” search queries would be associated with decreased tobacco consumption, and vice versa. Notably, 13 out of 17 sample cities showed consistent changes in both their municipal statistics and online search volumes, as expected; however, the other four cities failed to meet the expectation.

Eleven cities (Beijing, Shanghai, Shenzhen, Guangzhou, Changchun, Shijiazhuang, Fuzhou, Dalian, Chengdu, Suzhou, and Chongqing) showed a similar pattern, that is, their increases in “positive” smoking-related behaviours were much larger than that of their “negative” ones from 2013 to 2017. Take Beijing as an example, the red colour in the search volume of positive and negative search queries implies both of them experienced an increase between 2013 and 2017, while the blue colour refers to a decrease in tobacco consumption. The dark shade in the positive queries indicates the increase in positive queries is greater than the negative ones. In Nanning, the “positive” smoking-related behaviours experienced an increase, whereas the “negative” behaviours declined slightly. Correspondingly, the observed tobacco consumption per capita in the aforementioned 12 cities experienced varying degrees of shrinkage. Conversely, in Hangzhou, the increase in “negative” smoking-related behaviours was much more evident, and the tobacco consumption per capita was on the rise.

In addition, the situations in the other four cities (Hohhot, Lanzhou, Nanjing, and Linyi) were entirely different. In Hohhot, “positive” smoking-related behaviours spiked rapidly, while “negative” behaviours dropped noticeably. In Nanjing, the increase in “positive” smoking-related behaviours was much more evident. Conversely, the decrease in “negative” smoking-related behaviours was much more obvious in Linyi. Nevertheless, tobacco consumption per capita in the preceding three cities increased. In addition, in Lanzhou,



**Table 4.** The changes in online search queries and tobacco consumption from 2013 to 2017.

Type of city	Search queries				Municipal statistics
	Tobacco and smoking	Tobacco purchasing	Tobacco control	Tobacco cessation	Tobacco consumption
Type A average	0.13	0.11	0.43	0.40	-0.05
Type B average	0.06	0.05	0.12	0.15	-0.02
Type C average	0.06	0.02	0.11	0.10	0.03
All cities average	0.07	0.05	0.17	0.18	-0.01

Type A refers to cities with the strictest tobacco control policies; Type B refers to cities with slack tobacco control policies; Type C refers to cities without tobacco control policies. Comparison of the impacts of tobacco control policies on online search query data and municipal statistics.

the rise in “negative” smoking-related behaviours was more remarkable, but the observed tobacco consumption per capita saw a decrease. However, the contradictory results of Hohhot, Lanzhou, and Linyi can be explained by the *Statistical Report on China’s Internet Development Status* published by China Internet Network Information Center (CINIC) in 2017. It was reported that the Internet penetration rates in Inner Mongolia and the Gansu and Shandong Provinces were 52.2%, 42.4%, and 52.9%, respectively, which were lower than that of the whole nation, i.e., 53.2%. Thus, in cities where people have limited access to the Internet, the online search query data may underestimate the actual behaviours.

Additionally, Table 4 shows a comparison of the changes in online search queries and tobacco consumption at the aggregate level of the three types of cities. The increase in “positive” smoking-related behaviours was the largest in Type A cities, followed by Type B cities. Correspondingly, the descent in tobacco consumption was the most significant in Type A cities. Notably, Type C cities were the only cities in which the observed tobacco consumption increased slightly.

Six cities, including two Type A cities (Beijing and Shanghai) and four Type B cities (Nanning, Lanzhou, Changchun, and Fuzhou), were further selected to examine their tobacco control policies and their impacts on both online search queries and tobacco consumption. Different tobacco control policies were implemented between 2013 and 2017 in these selected cities, thereby offering us the chance to assess these policy interventions. As Figure A-3 in the Supplemental material shows, tobacco control policies exerted a consistent impact on both municipal statistics and online search queries in all of the cities except Lanzhou. After the adoption of tobacco control policies, tobacco consumption began to decline in Beijing and Nanning; this decrease was aggravated in Shanghai and Fuzhou, and the increase in consumption became less evident in Changchun. Correspondingly, the increase in “positive” smoking-related behaviours was much larger in these five cities. However, Lanzhou was an exception. After issuing tobacco control policies, the observed tobacco consumption decreased, but the increase in “negative” smoking-related behaviours was more marked than that of “positive” behaviours. The fact that the Internet penetration rate was only 42.4% in Gansu Province, as reported by the CINIC in 2017, might be an explanation. However, undeniably, tobacco control policies exerted a positive impact on smoking control and cessation.

## Discussion

This study assessed smoking-related behaviours using Baidu search query data and found that (1) although the level of attention given to all four groups of smoking-related queries increased, the increase in tobacco control search volume was the most significant, and that (2) differences in tobacco control policies are reflected in both the online smoking-related search queries and the observed tobacco consumption. Notably, tobacco control policies exerted the largest impact on smoking-related search queries and tobacco consumption in Type A cities, which means that the intensity of tobacco control policies greatly influences Baidu

users' awareness of tobacco control and cessation. (3) In addition, the information acquisition preferences represented by Baidu online search queries are consistent with observed levels of tobacco consumption, thereby achieving scientific and objective assessment for smoking-related behaviours. The validation process demonstrates that the assessed behaviours based on Baidu search queries are well associated with observed levels of tobacco consumption in sample cities, and both the online search queries and levels of consumption are sensitive to the implementation of tobacco control policies. Moreover, our validation sheds light on the possibility of using online search query data in policy assessment.

The added value of this study includes (1) expanding the available research on behaviour studies using online search query data through the inclusion of many more search queries and the supplementation of the validation process, (2) defining four categories of smoking-related queries to represent more diversified smoking-related behaviours, and (3) further demonstrating the possibility of using online search queries as a new instrument for policy assessment, especially for evaluating large-scale population behaviours in the public health domain.

Some limitations of this study are noteworthy. First, the online search query data omitted some population segments, such as elderly individuals who seldom use search engines and low-income individuals who have limited access to the Internet. Second, Baidu only offers search query data for two separate years and fails to examine continuous changes in search queries and statistical tobacco consumption. Future studies are encouraged to assess the robustness of this method by expanding the panel data to a time series analysis, especially to examine how online search queries and tobacco consumption respond to local tobacco control policies. Finally, only a limited number of cities in China provide publicly accessible data on tobacco consumption, and alternative validation data are needed to further examine the validity and robustness of the proposed method and search query data. In addition, the development of big data on social media also provides a platform that allows people to share their interests on certain topics and their behavioural preferences. Running content analyses on such social media data can be useful for developing crossover analyses with online search queries and offers a new lens to trace smoking-related information and behaviour changes.

### **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) received no financial support for the research, authorship, and/or publication of this article.

### **Supplemental material**

Supplemental material for this article is available online.

### **Note**

1. The relevant tobacco control policies were derived from the official government documents, which could be found on cities' government websites.

### **References**

- Andrews JO, Mueller M, Newman SD, et al. (2014) The association of individual and neighborhood social cohesion, stressors, and crime on smoking status among African-American women in southeastern US subsidized housing neighborhoods. *Journal of Urban Health* 91(6): 1158–1174.
- Artola C, Pinto F and de Pedraza García P (2015) Can internet searches forecast tourism inflows? *International Journal of Manpower* 36(1): 103–116.
- Ayers JW, Ribisl K and Brownstein JS (2011a) Using search query surveillance to monitor tax avoidance and smoking cessation following the United States' 2009 "SCHIP" cigarette tax increase. *PLoS One* 6(3): e16777.

- Ayers JW, Ribisl KM and Brownstein JS (2011b) Tracking the rise in popularity of electronic nicotine delivery systems (electronic cigarettes) using search query surveillance. *American Journal of Preventive Medicine* 40(4): 448–453.
- Bond L, Egan M, Kearns A, et al. (2013) Smoking and intention to quit in deprived areas of Glasgow: Is it related to housing improvements and neighbourhood regeneration because of improved mental health? *Journal of Epidemiology & Community Health* 67(4): 299–304.
- Brewer NT, Hall MG, Noar SM, et al. (2016) Effect of pictorial cigarette pack warnings on changes in smoking behaviour: A randomized clinical trial. *JAMA Internal Medicine* 176(7): 905–912.
- Callison K and Kaestner R (2014) Do higher tobacco taxes reduce adult smoking? New evidence of the effect of recent cigarette tax increases on adult smoking. *Economic Inquiry* 52(1): 155–172.
- Catalano MA and Gilleskie DB (2021) Impacts of local public smoking bans on smoking behaviours and tobacco smoke exposure. *Health Economics* 30(8): 1719–1744.
- China Internet Network Information Center (2018) *Statistical Report on Internet Development in China*. Beijing, China: China Internet Network Information Center.
- Christensen T, Welsh E and Faseru B (2014) Profile of e-cigarette use and its relationship with cigarette quit attempts and abstinence in Kansas adults. *Preventive Medicine* 69: 90–94.
- Danaei G, Ding EL, Mozaffarian D, et al. (2009) The preventable causes of death in the United States: Comparative risk assessment of dietary, lifestyle, and metabolic risk factors. *PLoS Med* 6(4): e1000058.
- Gahr M, Uzelac Z, Zeiss R, et al. (2015) Linking annual prescription volume of antidepressants to corresponding web search query data: A possible proxy for medical prescription behaviour? *Journal of Clinical Psychopharmacology* 35(6): 681–685.
- Ginsberg J, Mohebbi MH, Patel RS, et al. (2009) Detecting influenza epidemics using search engine query data. *Nature* 457(7232): 1012–1014.
- Grøtvedt L and Stavem K (2005) Association between age, gender and reasons for smoking cessation. *Scandinavian Journal of Public Health* 33(1): 72–76.
- Haardörfer R, Kreuter M, Berg CJ, et al. (2018) Cessation and reduction in smoking behaviour: Impact of creating a smoke-free home on smokers. *Health Education Research* 33(3): 256–259.
- Hanewinkel R, Isensee B, Sargent JD, et al. (2010) Cigarette advertising and adolescent smoking. *American Journal of Preventive Medicine* 38(4): 359–366.
- Huang J, Zheng R and Emery S (2013) Assessing the impact of the national smoking ban in indoor public places in China: Evidence from quit smoking related online searches. *PLoS One* 8(6): e65577.
- Kholodilin KA, Podstawski M and Siliverstovs B (2010) Do Google searches help in nowcasting private consumption? A real-time evidence for the US. KOF Swiss Economic Institute Working Paper No. 256, 1 April. Zurich: KOF Swiss Economic Institute.
- Lawless MH, Harrison KA, Grandits GA, et al. (2015) Perceived stress and smoking-related behaviours and symptomatology in male and female smokers. *Addictive Behaviours* 51: 80–83.
- Li Z, Liu T, Zhu G, et al. (2017) Dengue Baidu search index data can improve the prediction of local dengue epidemic: A case study in Guangzhou, China. *PLoS Neglected Tropical Diseases* 11(3): e0005354.
- Liu K, Wang T, Yang Z, et al. (2016) Using Baidu search index to predict dengue outbreak in China. *Scientific Reports* 6: 38040.
- Paek HJ, Baek H, Lee S, et al. (2020) Electronic cigarette themes on Twitter: Dissemination patterns and relations with online news and search engine queries in South Korea. *Health Communication* 35(1): 1–9.
- Pan B, Wu DC and Song H (2012) Forecasting hotel room demand using search engine data. *Journal of Hospitality and Tourism Technology* 3(3): 196–210.
- Salloum RG, Thrasher JF, Kates FR, et al. (2015) Water pipe tobacco smoking in the United States: Findings from the National Adult Tobacco Survey. *Preventive Medicine* 71: 88–93.
- Shareck M and Ellaway A (2011) Neighbourhood crime and smoking: The role of objective and perceived crime measures. *BMC Public Health* 11(1): 1–10.
- Singer G, Prullmann-Vengerfeldt P, Norbistrath U, et al. (2015) The relationship between Internet user type and user performance when carrying out simple vs. complex search tasks. *arXiv preprint arXiv:1511.05819*.
- Troelstra SA, Bosdriesz JR, De Boer MR, et al. (2016) Effect of tobacco control policies on information seeking for Smoking Cessation in the Netherlands: A Google Trends Study. *PLoS One* 11(2): e0148489.
- We Are Social and Hootsuite (2018) *Digital in 2018*. London: We Are Social and Hootsuite.

World Health Organization (WHO) (2019) *WHO Report on the Global Tobacco Epidemic, 2019*. Geneva, Switzerland: World Health Organization.

### **Author biographies**

**Ting Zhou** is a PhD student in the Department of Built Environment in Eindhoven University of Technology. Her research area focuses on smart urban environments that promote healthy living and well-being, involving employing advanced machine learning approaches in urban research.

**Long Chen** was a post-doc research associate at the School of Architecture, Tsinghua University and now is a Lecturer at Beijing University of Technology. His research interests include the association between human behaviours and the built environment, big data in urban planning and transportation.

**Zhaoxi Zhang** is a PhD student in the Department of Environmental Science at Aarhus University. She integrates multiple personal sensors to measure people's exposure in public open spaces and human physiological stress responses and analyses the association between urban features and mental health in the built environment.

**Zhengying Liu** is a research fellow at Tsinghua University. He holds a PhD in urban planning from the Eindhoven University of Technology, the Netherlands. His area of expertise focuses on smart urban environments that promote healthy living and well-being.

**Ying Long** is now working in the School of Architecture, Tsinghua University, as a tenured associate professor. His research focuses on urban science, including applied urban modeling, urban big data analytics and visualization, data augmented design and future cities.