A novel textual track-data-based approach for estimating individual infection risk of COVID-19

Lu Wei | Xiaojing Li | Zhongbo Jing | Zhidong Liu

School of Management Science and Engineering, Central University of Finance and Economics, Beijing, P. R. China

Correspondence
Zhidong Liu, School of Management Science and Engineering, Central University of Finance and Economics, No. 39 Xueyuan South Road, Haidian District, Beijing, P. R. China.
Email: liuzhidong@cufe.edu.cn

Abstract
With the recurrence of infectious diseases caused by coronaviruses, which pose a significant threat to human health, there is an unprecedented urgency to devise an effective method to identify and assess who is most at risk of contracting these diseases. China has successfully controlled the spread of COVID-19 through the disclosure of track data belonging to diagnosed patients. This paper proposes a novel textual track-data-based approach for individual infection risk measurement. The proposed approach is divided into three steps. First, track features are extracted from track data to build a general portrait of COVID-19 patients. Then, based on the extracted track features, we construct an infection risk indicator system to calculate the infection risk index (IRI). Finally, individuals are divided into different infection risk categories based on the IRI values. By doing so, the proposed approach can determine the risk of an individual contracting COVID-19, which facilitates the identification of high-risk populations. Thus, the proposed approach can be used for risk prevention and control of COVID-19. In the empirical analysis, we comprehensively collected 9455 pieces of track data from 20 January 2020 to 30 July 2020, covering 32 provinces/provincial municipalities in China. The empirical results show that the Chinese COVID-19 patients have six key features that indicate infection risk: place, region, close-contact person, contact manner, travel mode, and symptom. The IRI values for all 9455 patients vary from 0 to 43.19. Individuals are classified into the following five infection risk categories: low, moderate-low, moderate, moderate-high, and high risk.

KEYWORDS
COVID-19, infection risk, risk measurement, text mining, track data

1 | INTRODUCTION

In December 2019, an acute respiratory infectious disease, 2019-nCoV-infected pneumonia, broke out in Wuhan, Hubei, China. The World Health Organization (WHO) named this disease coronavirus disease 2019 (COVID-19). The WHO announced on March 11, 2020 that COVID-19 had become a global pandemic (Gautret et al., 2020).

COVID-19 is a highly transmissible coronavirus infection (Wu et al., 2020). In the past, there were three major, countrywide epidemics of coronavirus: the “severe acute respiratory syndrome” (SARS) outbreak in 2003 in mainland China, the “Middle East respiratory syndrome” (MERS) outbreak in 2012 in Saudi Arabia, and the MERS outbreak in 2015 in South Korea. This pattern indicates that infectious coronavirus diseases are a significant threat to the health of a large number of people (Wit et al., 2016). In the future, another coronavirus pandemic is very likely.

Thus, there is unprecedented urgency to devise an effective method to identify and assess who is most at risk of contracting the coronavirus disease. Since the first case of COVID-19 was reported, the WHO and its partners have been working with Chinese authorities and global experts to learn more about the highest-risk populations (WHO, 2020a). The assessment of individual infection risk can assist in reducing high-risk activities, lowering the infection risk, and containing the spread of infectious diseases.

The coronavirus disease spreads mainly through person-to-person contact during daily activities. It can be transmitted rapidly by a group of infectious agents through a variety of interactive ways (Medina, 2018). On February 3, 2020, during the early stage of the COVID-19 pandemic, the WHO
called for effective social distancing and movement restrictions to minimize human-to-human transmission of COVID-19 (Mizumoto & Chowell, 2020; WHO, 2020a). Early implementation of prevention and control measures can inhibit the spread of the infection (Kaur et al., 2020).

In the “big data” era, big data have advantages for infectious disease surveillance (Shweta et al., 2016). Track data consist of information on people’s daily activities that expose individuals to the risk of contracting infectious diseases. China has disclosed for the first time the textual track data of diagnosed patients during the COVID-19 pandemic. In addition to standard information available from mobile phone users, such as tracking information on location and time, the disclosed track data include information on modes of transportation, symptoms, contact manners, etc. Based on the valuable track data of diagnosed patients, China has taken many targeted effective measures, including nucleic acid testing, social distance control, isolation of diagnosed patients, and tracking and isolation of close contacts, which have successfully prevented and controlled the spread of COVID-19 in China. Thus, the valuable information on daily activities contained in the track data is of great importance for identifying and assessing who is most at risk of contracting the disease and controlling the spread of the disease.

However, in previous studies, only part of the information contained in track data (predominantly time and location information) was utilized for the analysis of disease spread, while other valuable information related to personal risk assessment contained in track data was ignored. Specifically, Wesolowski et al. (2016) used the location information contained in personal mobile phone data to study the spread of diseases. Lee et al. (2016) used location data contained in social media posts and web search information for infectious disease surveillance and inference. Au et al. (2018) collected data including disease type, geographic location, and timestamps through Bluetooth, which were used to conduct the real-time monitoring of epidemic outbreaks. Kuchler et al. (2020) found that COVID-19 is more likely to spread between regions with closer social connections in the United States, by combing information regarding the geographic location and social relationships on Facebook. Ren et al. (2020) utilized location information to assess the spread of COVID-19 in communities in Beijing, Guangzhou, and Shenzhen.

Thus, as discussed above, the previous studies mainly used only part of track data (typically location and time information) to analyze the spread of disease. Thus far, to the best of our knowledge, no study fully utilized the information in the textual track data related to behaviors of an individual immediately prior to infection, to calculate personal infection risk during a pandemic.

Based on an early effort made by Ginsberg et al. (2009), data analytics have played an important role in risk identification and assessment. Recently, with the development of text-mining technology (Cao et al., 2020; Feldman & Hart, 2018), analysis of massive amounts of textual data containing valuable risk information has been introduced into many fields of risk management, such as finance (Romqvist & Sarlin, 2014; Wei et al., 2019a) and energy (Li et al., 2018; Wei et al., 2019b), as well as health and safety (Ajayi et al., 2019). The use of textual data has enriched quantitative risk measures and improved the effectiveness of risk management.

Thus, this paper proposes a novel textual track data based infection risk measurement approach. This approach for the first time introduces the textual track data—which contains information on daily activities that expose an individual to the risk of contracting COVID-19—into an epidemiological analysis for personal infection risk assessment. It is applicable not only to individual risk measurement for COVID-19 infection but also to that of other epidemic sicknesses caused by coronaviruses with similar modes of transmission. This approach can enrich data-driven epidemiological risk assessment measures.

The principle of our proposed approach is that a person with movements similar to that of a diagnosed patient has a significantly higher risk of contracting COVID-19. Specifically, the proposed approach is divided into three steps. First, common features are identified from track data to build a general portrait of COVID-19 diagnosed patients. Then, the infection risk index (IRI) is constructed based on the identified common features to measure the individual infection risk of COVID-19. Finally, value-at-risk (VaR), a popularly used financial risk measurement method, is adopted to divide people into different COVID-19 infection risk categories based on the values of IRI. By doing so, the proposed approach is useful in identifying the high-risk groups to prevent the spread of COVID-19 and in guiding people to reduce high-risk activities to lower the risk of getting COVID-19.

This paper has two main contributions to the field of data-driven epidemiological risk measurement. First, compared with the previous studies, which used the track data for assessing the impact of travel on the spread of diseases, we found that the valuable information on daily activities contained in the track data is of great importance for identifying and assessing who is most at risk of contracting the disease and controlling the spread of disease. Therefore, this paper proposes a track data-driven approach for COVID-19 individual infection risk estimation. Furthermore, the track data (typically time and location information) used in previous studies are difficult to identify groups at high risk of getting COVID-19. Since there are many epidemiological characteristics of COVID-19 that determine the level of individual infection risk, it is difficult to measure individual infection risk only by relying on these two data of location and time. In addition to tracking information on location and time, the track data used in this paper also include information on modes of transportation, symptoms, contact manners, etc. The more complete track data we used can better capture relevant epidemiological characteristics of COVID-19, which can give a more accurate individual infection risk estimation result than using only location and time data. Thus, this paper proposes a novel approach for individual infection risk assessment during the disease pandemic based on a complete track data of diagnosed patients of COVID-19.
Second, this paper empirically illustrates the proposed approach based on 9455 pieces of textual track data of COVID-19 diagnosed patients during the period between January 20, 2020 and July 30, 2020. Since the patient data we collected covers 32 provincial-level administrative regions in China, in addition to studying the patient characteristics at the national level, we also study the patient characteristics of each province. The empirical results show that the relevant features of patients in different provincial-level administrative regions can vary significantly, indicating that context-specific regional risk control and prevention measures should be taken to contain the spread of COVID-19. Thus, using the proposed approach for other regions and countries can guide efforts for developing context-specific national and regional risk assessment and prevention plans since there may be differences in daily activities attributable to diverse cultural backgrounds worldwide. Furthermore, an empirical comparison is made between the personal infection risk assessment results and the clinical characteristics of diagnosed patients to prove the effectiveness of the proposed approach.

The rest of this paper is organized as follows. Section 2 introduces the proposed infection risk measurement approach based on textual track data. Section 3 describes the sample data. Section 4 presents the empirical results, including the portrait of COVID-19 patients, IRI, and infection risk categories. Finally, the conclusions are summarized in Section 5.

2 | METHODOLOGY

In this section, we describe the proposed approach for infection risk measurement based on textual track data of COVID-19 diagnosed patients. The proposed approach summarized the common features of patients to measure individual infection risk and classify individuals into different risk categories regarding risk of contracting COVID-19.

Figure 1 outlines the proposed approach to measure individual infection risk of COVID-19. Specifically, the input data were the textual track data of COVID-19 diagnosed patients. The proposed approach is divided into three steps. First, a general portrait of COVID-19 diagnosed patients is built by identifying the common features from the track data of diagnosed patients. Then, an IRI is constructed based on those identified common track features of patients to measure the individual infection risk of COVID-19. Finally, VaR, a popular financial risk measurement method, is adopted to divide people into different COVID-19 infection risk categories based on the IRI values. After these three steps, we can obtain three results, that is, the portrait of the COVID-19 diagnosed patient, the IRI values, and the individual infection risk categories. In the following subsections, we describe the three steps of the proposed approach in detail.

2.1 | Building portrait of COVID-19-diagnosed patients

The first step is to construct a portrait of COVID-19-diagnosed patients, which is composed of the common features among them. The textual track data contained information on where patients went, what they did, who they met, what symptoms they had, and so on. Thus, we can identify the common features from the track data of diagnosed patients to form a general portrait of the COVID-19 diagnosed patients.

To construct the general portrait of COVID-19 diagnosed patients, we first remove stop-words and segment text passages into words. Then, we select the high-frequency words. Finally, we classify the high-frequency words to identify common features and construct the portrait of COVID-19 diagnosed patients.

The proposed approach extracts high-frequency words based on term frequency (TF), one of the most commonly used high-frequency word extraction techniques (Rustam et al., 2019). TF is the ratio of the number of instances of a term to the total number of terms. The higher TF is for a particular term, the higher the frequency with which the term
occurs. TF can be expressed as

\[
tf_i = \frac{\text{count}(t_i)}{\sum_{j=1}^{n} \text{count}(t_j)}
\]  

(1)

where \(tf_i\) represents the value of TF for term \(t_i\). \(\text{count}(t_i)\) denotes the frequency of term \(t_i\), and \(n\) is the number of terms in total.

After selecting high-frequency words based on the calculated TF for each word, we then manually classify the selected high-frequency words into different classifications appropriately to determine the common features, using the domain knowledge of experts in the field and following the process designed by Huang and Li (2011). Although there are some automatic labeling methods (Mei et al., 2007), these are not suitable in cases where such labeling requires domain knowledge. It is customary to manually label classification results, ensuring high labeling quality (Chang et al., 2009). Huang and Li (2011) designed a manual labeling procedure that utilizes the domain knowledge of human experts; this process has also been adopted by Bao and Datta (2014) and Wei et al. (2019a, b).

Specifically, the aforementioned scholars used the manual labeling procedure to identify different types of corporate risk factors from textual risk disclosures of annual reports of U.S. public companies. In essence, labeling the classification results of patients’ track data serves to identify the risk factors leading to COVID-19 exposure; this is similar to the application backgrounds of labeling corporate risk factors. Thus, the labeling procedure designed by Huang and Li (2011) was deemed suitable for our application of classifying high-frequency words into different categories with meaningful label names.

The process of using the manual labeling procedure designed by Huang and Li (2011) is as follows. Specifically, four experts on our research team undertook the work of labeling high-frequency words. Each expert labels half of the total high-frequency words and each high-frequency word is labeled by two experts. For a word that is assigned the same label by two experts, that label is then chosen as its meaningful label. For the word with no agreement on the label between two experts, four experts discuss and determine the final label. Thus, based on the four-expert domain knowledge, each high-frequency word is given a meaningful label name. The high-frequency words with the same label are merged to form the final categories of track data. Overall, through the manual labeling process, the final classifications with meaningful names are the common features extracted from textual track data, which form the general portrait of the COVID-19-diagnosed patients.

2.2 Constructing individual IRI

Next, we construct an individual IRI based on the common features identified in the first step. The principle of constructing the IRI is that a person with track data of daily activities most similar to those of an identified patient will be at a higher risk of infection. COVID-19 spreads mainly through person-to-person contact (Mizumoto & Chowell, 2020). Thus, if a person has visited a place where a diagnosed patient has been or was in close contact with someone that has been diagnosed, they have a higher risk of contracting COVID-19. In other words, the higher the degree of coincidence between the actions of a person and the common features of a patient, the greater the risk of infection for the person.

Thus, based on the common features of diagnosed patients, which comprehensively describe their daily behaviors from different perspectives, we develop an infection risk indicator system consisting of two levels of indicators. The first-level indicators are the common features of the diagnosed patients in the track data, which represent different types of infection risk factors that lead a person to contract COVID-19. The second-level indicators are the high-frequency words of each common feature, which represent the specific risky behaviors in each type of infection risk factor. The more frequently the secondary indicators appeared in the patients’ track data, the more frequently patients engage in the behaviors represented by the secondary indicators, thereby this behavior is a high-risk behavior leading a person to COVID-19 infection.

Thus, for each secondary indicator, we calculate a frequency weight to represent the risk importance of the secondary indicator, leading a person to suffer from COVID-19. A high frequency of occurrence for a secondary indicator represented a high likelihood of infection. The formula for weighting the secondary indicators is

\[
w_{ij} = \frac{\text{count}(x_{ij})}{\sum_{j=1}^{n_i} \text{count}(x_{ij})}
\]  

(2)

where \(w_{ij}\) and \(\text{count}(x_{ij})\) denote the weight and frequency of the \(j\)th secondary indicator under the \(i\)th primary indicator \(x_{ij}\), respectively, and \(n_i\) is the total number of secondary indicators under the \(i\)th primary indicator.

The IRI is calculated via weighted simple summation of the frequency of each secondary indicator:

\[
\text{IRI} = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \text{count}(x_{ij}) \times w_{ij}
\]  

(3)

where \(m\) is the number of primary indicators. The IRI quantifies the individual infection risk of COVID-19. A higher index represents a higher risk of becoming infected with COVID-19.

In summary, the infection risk level of an individual is determined by both the types of infection risk factors and the risk importance of the factors. The identified six common features from diagnosed COVID-19 patients are six infection risk factors leading to get COVID-19, which together determine the individual infection risk level. The risk importance
of factors is determined by the frequencies that occurred in the patients' track data. The higher the frequency of the factor, the higher the risk of the factor that leads a person to get infected with COVID-19.

Thus, under our developed infection risk indicator system, two people traveling on a single bus may have different infection risk levels. The infection risk is determined by not only the mode of transportation but also by whether they had close contact with others through sitting together or whether their seatmates contracted COVID-19. Compared with a person who rode the bus alone, a person who sat next to a COVID-19 patient on the bus will have a higher IRI value, which indicates that the latter has a higher risk of infection.

2.3 Determining individual infection risk category

Having obtained the IRI, we adopt the VaR method to classify the individuals into different infection risk categories with regard to contracting COVID-19. Pioneered by J.P. Morgan, VaR has become a standard measure used in financial risk measurement (Hashemi et al., 2019; Rosenberg & Schuermann, 2006). In this paper, we apply the VaR method to the field of infectious diseases to measure infection risk.

In the financial world, insufficient capital held by financial institutions to cover an extreme risk may bankrupt the institution should a loss occur. Therefore, for sound financial risk management, the measurement of extreme risk is particularly important. VaR is a widely used measure for extreme risk measurement. As shown in Figure 2, VaR is defined as a quantile of the financial risk distribution (Belles et al., 2014). It is the maximum possible loss suffered by the asset portfolio in the future holding period under a certain confidence level. VaR at a specific confidence level $\alpha \in (0, 1)$ is defined as the smallest number $l$ such that the probability of $L$ exceeding $l$ is not larger than $(1 - \alpha)$:

$$\text{VaR}(\alpha) = \inf \{ l : P(L \geq l) \leq 1 - \alpha \}$$  \hspace{1cm} (4)

where $\text{VaR}(\alpha)$ is the risk value under the confidence level $\alpha$, that is, we have a confidence of $\alpha$ that the loss will not exceed $\text{VaR}(\alpha)$. Thus, VaR essentially measures the tail risk, that is, the extreme risk to the right of the curve from the perspective of probability. Because of its convenience and effectiveness, VaR has been used in other fields of risk management as well, such as the chemistry (Xu et al., 2018), biology (Prettenthaler et al., 2015), engineering (Liu et al., 2018), and medicine (An et al., 2017).

In this study, we aimed to identify individuals at the highest risk of contracting an infectious disease, which is particularly important for successful disease prevention and control. In essence, we also aim to estimate the extreme value of individual infection risk. Thus, we adopted VaR, which herein refers to the maximum possible IRI for a subject over a certain period within a certain confidence level. In this paper, the value of the IRI is positive, and thus, a higher VaR indicates a higher risk of infection. We have a confidence of $\alpha$ that the IRI of a COVID-19 patient will not exceed $\text{VaR}(\alpha)$. The higher $\alpha$ is, the higher the personal infection risk represented by the $\text{VaR}(\alpha)$ becomes. Thus, by employing different confidence levels, we can obtain a series of VaR values, which are used as the basis to divide the IRI values of all the collected patients into different score ranges representing the different infection risk levels. Then, based on those score ranges, individuals with various IRI values are classified into different infection risk categories.

There are other classic clustering techniques, such as the widely used $K$-means approach (He et al., 2021), which divides a sample set into $K$ clusters according to the distance between samples, by minimizing the distance between data points to form clusters while maximizing the distance between clusters. However, clustering the IRI values through $K$-means analysis is based on the distance of sample data while the idea of extreme risk is not adequately considered. Thus, compared with the clustering techniques, the VaR measures extreme risk from the perspective of probability, which is more suitable in this paper. As a robustness test, we adopted the $K$-means clustering approach to divide the values of IRI into different risk categories. In summary, by completing the three aforementioned steps of the proposed approach, we can summarize that the common features of COVID-19 patients, measure the individual risk of contracting an infectious disease, and finally classify individuals into different infection risk categories based on the textual track data of diagnosed patients.

3 DATA

3.1 Data collection

To discover individuals at high risk of contracting COVID-19 and control the spread of the disease, China comprehensively disclosed the track data of diagnosed patients in detail for the first time. Every province conducted epidemiological investigations on the patients in the province to obtain the track data of them. Subsequently, the department of health commission in each province disclosed the track data of diagnosed patients to the public.
FIGURE 3 The number of daily cumulative diagnosed patients in China during 2020

Tencent News, one of the largest Internet portals in China (Dong & Wang, 2016; Zheng et al., 2019), summarized the disclosed track data of COVID-19 diagnosed patients in 32 provincial-level administrative regions of China to form a dataset that contains the track data of diagnosed COVID-19 patients in China. From January 20, 2020, China has begun to disclose the track data of diagnosed patients. According to the disclosure of the National Health Commission of the People’s Republic of China, the total number of cumulative diagnosed patients in China increased sharply in January and February 2020 and then remained essentially stable for 10 months since March 2020 (Figure 3). Thus, as COVID-19 was controlled in China with few newly diagnosed patients, Tencent News halted this data collection on July 30, 2020.

In total, Tencent News collected 9455 pieces of track data disclosed by 32 provincial-level administrative regions of China from January 20, 2020 to July 30, 2020. The track dataset constructed by Tencent News contained almost all the track data of diagnosed COVID-19 patients in China. From January 20, 2020 to July 30, 2020 for our empirical analysis, which is almost equal to the total sample of official disclosed track data of diagnosed COVID-19 patients in China.

Below, we provide an example of a piece of track data of a COVID-19 patient:

‘On January 21, four members of Luo’s family and Yu went to visit the patient’s home and had dinner at the patient’s home. On January 13, they drove to Jingzhou City, Hubei province to visit the father-in-law’s home and returned to Yueyang City one day later. From February 1 to February 7, Luo went to the Yueyang Hospital of traditional Chinese medicine because of “uremia.” On February 8, the novel coronavirus was detected as nucleic acid positive. Close contacts include Luo’s family members, patients and their families in the same consulting room, and some medical staff, a total of 73 people.’

Having obtained the original track data of patients, we then manually read all the collected 9455 pieces of track data. As shown in the example mentioned above, we found that the disclosed track data consist of information on the daily activities of patients, which mainly contain information on the regions and places where patients visited, the transportations patients took, the symptoms most patients had, who were the close contacts of diagnosed patients and the ways people had been in close contact with the patients. Thus, by manually performing exhaustive text perusal with experts’ domain knowledge, we summarized that the track data of COVID-19 patients included six aspects of information, that is, place, region, close-contact person, contact manner, travel mode, and symptom, which are useful for identifying individuals who at high risk of contracting COVID-19 and further for controlling the spread of COVID-19.

In addition, the collected 9455 pieces of track data cover diagnosed patients in 32 provincial-level administrative regions in China. Although there was no unified template for disclosing track data across provinces in China, the track data disclosed by each province contained these six aspects of information except Taiwan province. The two pieces of track data collected from the Taiwan province only contained the symptom information of patients.

3.2 Data description

In this section, we describe the collected textual track data of COVID-19 diagnosed patients from two aspects: regional distribution and age distribution.

3.2.1 Regional distribution of COVID-19 patients

The dataset contained 9455 diagnosed patients’ track data covering 32 provincial-level administrative regions. Table 1 presents the number of diagnosed cases in each province. It is noteworthy that Hubei province, which had the largest number of diagnosed cases, did not disclose track data, because the widespread nature of COVID-19 there made the number of diagnosed cases too excessive to disclose.

From Table 1, it is clear that the Heilongjiang province disclosed the largest number of track data of patients, with a total of 1641 cases, followed by the provinces of Henan (1208) and Zhejiang (1147). The provinces with the number of disclosed cases between 500 and 1000 include Chongqing (622), Guangdong (536), Hunan (515), and Shandong (503). The provinces with the number of disclosed cases between
TABLE 1 The number of diagnosed patients’ track data in each provincial-level administrative region

| Province                      | Number | Province                      | Number |
|-------------------------------|--------|-------------------------------|--------|
| Heilongjiang                  | 1641   | Jilin                         | 152    |
| Henan                         | 1208   | Shandong                      | 503    |
| Zhejiang                      | 1147   | Guangdong                     | 536    |
| Chongqing                     | 622    | Guangxi Zhuang Autonomous Region | 96  |
| Hunan                         | 515    | Guizhou                       | 105    |
| Henan                         | 1208   | Guangxi Zhuang Autonomous Region | 96  |
| Hunan                         | 515    | Guizhou                       | 105    |
| Shandong                      | 503    | Inner Mongolia Autonomous Region | 34  |
| Beijing                       | 396    | Ningxia Hui Autonomous Region  | 34     |
| Sichuan                       | 345    | Jiangxi                       | 17     |
| Hebei                         | 352    | Qinghai                       | 17     |
| Fujian                        | 271    | Shanghai                      | 9      |
| Anhui                         | 232    | Xinjiang Uygur Autonomous Region | 5   |
| Liaoning                      | 201    | Macao Special Administrative Region | 3 |
| Tianjin                       | 188    | Hong Kong Special Administrative Region | 3 |
| Jiangsu                       | 155    | Taiwan                        | 2      |

100 and 500 include Beijing (396), Hainan (353), Sichuan (345), Hebei (335), Fujian (271), Anhui (232), Liaoning (201), Tianjin (188), Jiangsu (155), Jilin (152), Shanxi (124), and Guizhou (105). The remaining provinces with the number of disclosed cases under 100 include Guangxi Zhuang Autonomous Region (96), Gansu (87), Shaanxi (62), Yunnan (57), Inner Mongolia Autonomous Region (34), Ningxia Hui Autonomous Region (34), Jiangxi (17), Qinghai (17), Shanghai (9), Xinjiang Uygur Autonomous Region (5), Macao Special Administrative Region (3), Hong Kong Special Administrative Region (3), and Taiwan (2).

3.2.2 Age distribution of COVID-19 patients

Most track data contained information about the age of the patient. Among the total 9455 samples, 7930 contained the valid age information of the patient. Patients were divided into nine age groups: 0–9, 10–19, 20–29, 30–39, 40–49, 50–59, 60–69, 70–79, and 80 years and above.

The age distribution of COVID-19 patients is shown in Figure 4. The patients were mainly aged between 30 and 60 years old, which accounted for 62.94% of the total. Among those, the number of patients aged 40–50 years was the highest (1703), followed by those aged 30–40 years (1661) and 50–60 years (1627). For other age groups, the group aged 20–30 and 60–70 had 1091 and 910 diagnosed cases, respectively. There were 351 diagnosed patients aged between 70 and 80 years. The numbers of diagnosed cases in the groups of people aged 10–20, 0–10, and above 80 are less than 300, which were 292, 159, and 136, respectively.

4 EMPIRICAL ANALYSIS

In this section, based on the collected 9455 pieces of textual track data of COVID-19 diagnosed patients in China, we estimate the individual infection risk of COVID-19 using the proposed approach. Specifically, we first build a general portrait of the diagnosed patients in China and compare the difference of patients’ portraits in different provinces by extracting common features from textual track data. Then based on the extracted common features of the diagnosed patients, we construct an infection risk indicator system and calculate the IRI for each of the 9455 cases. Finally, by adopting the VaR method, we divide the individual infection risk categories based on the result of the IRI. In the following subsections, we describe the empirical results in more detail.
### 4.1 Portraits of Chinese COVID-19 diagnosed patients

As discussed in Section 2, the portrait of the COVID-19 diagnosed patients is composed of the common features, which are the classification results of high-frequency words extracted from the textual track data. For the collected original-track data, we first adopt the Harbin Institute of Technology stop-word list to remove the meaningless stop-words. It is one of the most widely used stop-word lists in the field of Chinese natural language processing and has been adopted in many studies (Gao et al., 2019; Zhong et al., 2020). Then we use the “jieba” module in Python, a popular Chinese word augmentation tool for Chinese word segmentation (Bharti & Singh, 2016).

#### 4.1.1 General portrait of nationwide COVID-19 diagnosed patients

Based on the total nationwide 9455 pieces of COVID-19 diagnosed patients’ track data, we obtain 129,093 words; of those, 9383 were unique. We regard the 9383 different words as candidates for high-frequency words selection. To select the high-frequency words, we rank candidate words based on their frequency and take the cumulative proportion of frequency reaching 80% as the boundary. Among the 9383 candidate words, the frequency of the top 1200 words accounts for 81.2% of all the words. Thus, the top 1200 words are high-frequency words extracted from the track data of the diagnosed patients.

As presented in Section 3.1, after manually examining all the original collected 9455 pieces of track data of diagnosed COVID-19 patients, we predefine six categories of topics contained in the patient’s track data: places and regions visited, close contacts, contact manner, travel mode, and symptoms. Then, as discussed in Section 2, we adopt the manual labeling process proposed by Huang and Li (2011) to manually classify the 1200 high-frequency words into the predefined six categories based on the domain knowledge of four experts. Specifically, four experts on our research team undertook the work of classifying and labeling these 1200 high-frequency words. Each expert labels 600 out of the 1200 high-frequency words by giving one of the names of six categories and each high-frequency word is labeled by two persons. For each half of 1,200 high-frequency words, 98% and 97% of the initial labels were agreed upon without further discussion. For the word that two experts give the same label, its meaningful label name is the same label given by two experts. For the remaining high-frequency words without the agreement on labels, all four experts discussed and decided upon the final category.

Finally, after classifying the high-frequency words with consensus labels into corresponding categories, we manually classify 1200 high-frequency words into six categories. The final six categories with meaningful names are the common features extracted from track data of diagnosed COVID-19 patients, that is, place, travel mode, contact manner, region, symptom, and close-contact person, which are used to construct the general portrait of nationwide COVID-19 diagnosed patients. As shown in Figure 5, the proportions of high-frequency words classified into six common features of place, region, travel mode, contact manner, symptom, and close-contact person are 26.77%, 24.17%, 14.54%, 14.71%, 10.38%, and 9.43%, respectively.

Figure 6 shows the word clouds of the six common features identified from the track data of COVID-19 patients, where the font size corresponds to the frequency of occurrence for each word in the common feature. The larger the font size of the word as presented, the more frequently it appears. Figure 7 visually presents the general portrait of the COVID-19 diagnosed patients in China, which is characterized by six common features with typical high-frequency words classified into each feature. Below, we describe the six identified common features that constitute the general portrait.

The place feature represents the commercial, institutional, and residential locations where the patients had frequently been within the suspected infection window. The high-frequency words classified into the place feature are the high-risk places where the diagnosed patient frequently appeared. As shown in Figure 6, the high-risk places, those shown in larger fonts, included hospitals, residential districts, centers for disease control, supermarkets, COVID-19 designated hospitals, ports, airports, and hotels. Frequent visits to these high-risk places during this period without proper protective equipment likely increased COVID-19 infection risk.

The region feature represents the administrative regions (province, city, and district) that the patients had visited. Regions appearing frequently in the dataset represented the high-risk regions visited by the patients. In Figure 6, Wuhan city of Hubei province appeared much more frequently than other regions, which showed that a large number of patients had traveled to these locations during the suspected infection window. Besides, some patients in Russia flew from Moscow to Vladivostok and then entered Suifenhe city in the

![Figure 5](image-url)

**Figure 5** The proportions of high-frequency words classified into six common features.
The third feature is close-contact person, which refers to people who had close contact with diagnosed patients. Individuals coming in close contact with infected people are at a high risk of COVID-19 infection. Since COVID-19 initially broke out during the Spring Festival in China, typical holiday gatherings meant that the close contacts were mainly family members, such as spouses, children, and parents (Figure 6). If an individual had close contact with a patient, their risk of contracting COVID-19 was higher.

The contact manner feature denotes the behavioural attributes of the patients or the manner in which they came in contact with others. As shown in Figure 6, the high-frequency contact manners included gathering, dining together, shopping (specifically grocery shopping), going to work, riding together, and other close contact ways. Thus, a person in one of these types of close contact with a patient had a higher risk of contracting COVID-19.

The next feature is the travel mode, which represents how the patients had traveled within the suspected infection window. As shown in Figure 6, the high-frequency travel modes

southeast of Heilongjiang Province. Thus, Russia and its Vladivostok and Moscow city and Suifenhe city (a border city in China) were high-risk regions with larger fonts in word clouds. Additionally, Beijing city, Haikou city in Hainan province, Chongqing city and its Wanzhou district were also found to be high-risk regions. People should avoid going to these high-risk regions. If an individual had traveled to these high-risk regions without proper protection, their risk of contracting COVID-19 was higher than when traveling to other areas.

The third feature is close-contact person, which refers to people who had close contact with diagnosed patients. Individuals coming in close contact with infected people are at a high risk of COVID-19 infection. Since COVID-19 initially broke out during the Spring Festival in China, typical holiday gatherings meant that the close contacts were mainly family members, such as spouses, children, and parents (Figure 6). If an individual had close contact with a patient, their risk of contracting COVID-19 was higher.

The contact manner feature denotes the behavioural attributes of the patients or the manner in which they came in contact with others. As shown in Figure 6, the high-frequency contact manners included gathering, dining together, shopping (specifically grocery shopping), going to work, riding together, and other close contact ways. Thus, a person in one of these types of close contact with a patient had a higher risk of contracting COVID-19.

The next feature is the travel mode, which represents how the patients had traveled within the suspected infection window. As shown in Figure 6, the high-frequency travel modes
were private vehicle, plane, taxi, train, bus, coach, and online car-hailing. With the exception of private vehicle is private transportation. To contain the COVID-19 pandemic, public transport should be avoided, and masks should be worn. If an individual frequently used public transportation without proper protective equipment, the risk of infection COVID-19 was relatively high.

The last feature is the symptoms of the diagnosed patients. As shown in Figure 6, the common symptoms of diagnosed patients included fever, cough, indisposition, fatigue, headache, sore throat, and so on. The common symptoms can be used to assist the preliminary judgment of suspected cases. Thus, if a person had these high-risk symptoms, the likelihood of infection was high and he should seek medical treatment in time.

4.1.2 Differences among patient portraits in different provinces

After constructing a general portrait based on track data of nationwide diagnosed patients, we further compare the portraits of patients in different provinces to analyze any differences between provinces. As presented in Section 3.2.1, the sample data covered diagnosed patients in 32 provincial-level administrative regions. Among them, the track data of diagnosed patients in Taiwan province only included symptom information. Thus, due to the limited information, the Taiwan province data were discarded from this analysis.

By constructing a unique COVID-19 patient portrait for each province, we found that similar to the general portrait constructed in Section 4.1.1, the provincial unique portrait of diagnosed patients is also composed of the six common risk features, that is, place, region, close-contact person, contact manner, travel mode, and symptom. Patients in different provinces exhibited similar symptoms. However, for the remaining five features, there are clear differences among patients in different provincial-level administrative regions.

Specifically, as shown in Table 2, with regard to the place feature, there were differences in places of living, entertainment, and public transportation for patients in different provinces. Places of living refer to the places that residents frequent in their daily lives, such as communities, supermarkets, hospitals, and companies; entertainment places are those related to leisure activities, including shopping malls, hotels, restaurants, and scenic spots; public transportation places are bus stations, railway stations, airports, etc.

In particular, the high-risk places were mainly public transportation places, accounting for as much as 60.00, 45.45, 40.66, 36.83, and 34.81% of all words classified into the place feature for Xinjiang, Macao, Shaanxi, Guangdong, and Yunnan, respectively. Meanwhile, living places accounted for above 90% of all words classified under the place feature for patients in Shanghai (100.00%), Henan (97.44%), Jiangsu (96.62%), Beijing (96.54%), Hunan (96.27%), Chongqing (96.23%), Anhui (94.64%), Guangxi (94.47%), Hebei (94.34%), Jiangxi (92.41%), Sichuan (90.00%)

| High-frequency places | Province (proportion of subtype place to total places) |
|-----------------------|------------------------------------------------------|
| Public transportation | Xinjiang (60.00%), Macao (45.45%), Shaanxi (40.66%), Guangdong (36.83%), Yunnan (34.81%) |
| Living                | Shanghai (100.00%),Henan (97.44%),Jiangsu (96.62%),Beijing (96.54%), Hunan (96.27%), Chongqing (96.23%), Anhui (94.64%), Guangxi (94.47%), Hebei (94.34%), Jiangxi (92.41%), Sichuan (90.00%) |
| Entertainment         | Xinjiang (40%), Ningxia (24.18%), Gansu (18.26%), Hainan (17.83%), Yunnan (17.04%), Heilongjiang (14.12%), Qinghai (10.53%) |

Table 2 Differences in high-frequency places by region
### Table 3 Differences in high-frequency travel modes by region

| High-frequency travel modes | Province (proportion of short or long-distance transportation) |
|----------------------------|---------------------------------------------------------------|
| Short-distance transportation (taxi, bus, subway, etc.) | Beijing (100.00%), Shanghai (100.00%), Liaoning (100.00%), Chongqing (82.28%), Xinjiang (71.43%), Ningxia (65.00%), Gansu (63.45%), Shaanxi (57.02%), Shanxi (55.28%), Inner Mongolia (52.17%) |
| Long-distance transportation (plane, train, ferry, etc.) | Macao (100.00%), Jiangsu (100.00%), Qinghai (100.00%), Zhejiang (100.00%), Tianjin (93.62%), Fujian (92.95%), Shandong (82.55%), Henan (83.26%), Guangdong (80.38%), Jilin (80.34%), Guizhou (80.00%), Guangxi (78.85%), Sichuan (77.85%), Hainan (77.21%), Yunnan (75.36%), Hunan (75.00%), Hebei (65.88%), Anhui (63.64%), Heilongjiang (63.01%) |

### Table 4 Differences in high-frequency contact manners by region

| High-frequency contact manner | Province (proportion of subtype activity to total activities) |
|------------------------------|---------------------------------------------------------------|
| Gathering entertainment      | Gansu (73.08%), Shanghai (72.73%), Inner Mongolia (67.50%), Chongqing (60.90%) |
| Travelling                   | Yunnan (15.22%), Guangxi (9.59%), Zhejiang (7.74%), Guangdong (7.38%) |
| Going to work and meetings   | Beijing (31.53%), Shanghai (27.27%), Shaanxi (14.29%), Tianjin (10.94%), Jiangsu (10.77%), Liaoning (10.47%), Sichuan (10.05%) |
| Visiting relatives           | Qinghai (50.00%), Hebei (29.76%), Ningxia (27.45%), Guangxi (19.18%), Guangdong (17.51%), Henan (14.10%) |

as they entered the province by trains or planes, so these imported patients did not take short-distance transportations to spread the disease within the province.

Table 4 presents the differences in high-risk contact manner between different provinces. The proportion of each type of activity in relation to all activities is listed in parentheses after the name of each province. We found that diagnosed patients in different provinces had different preferences in terms of gathering entertainment, traveling, visiting relatives, and going to work or meetings.

Specifically, in Gansu (73.08%), Shanghai (72.73%), Inner Mongolia (67.50%), and Chongqing (60.90%), participating in group entertainment activities was the main mode of coming into contact with patients. The patients in Yunnan (15.22%), Guangxi (9.59%), Zhejiang (7.74%), and Guangdong (7.38%) were more likely to be exposed during travel. In Beijing (31.53%), Shanghai (27.27%), Shaanxi (14.29%), Tianjin (10.94%), Jiangsu (10.77%), Liaoning (10.47%), and Sichuan (10.05%), going to work and meetings were high-risk activities. An important contact mode in Qinghai (50.00%), Hebei (29.76%), Ningxia (27.45%), Guangxi (19.18%), Guangdong (17.51%), and Henan (14.10%) was visiting relatives.

The differences in high-frequency close contact by region are listed in Table 5. Relatives accounted for more than 90% of close contacts in many provinces, including Anhui (90.99%), Heilongjiang (95.39%), Guizhou (95.06%), Chongqing (94.65%), Liaoning (93.99%), Shanxi (93.98%), Fujian (93.85%), Shanghai (93.55%), Hebei (93.53%), Tianjin (93.44%), and Guangdong (93.10%).

However, in other provinces, there were additional types of high-frequency close contact. Specifically, operators and suppliers of companies accounted for 64.71% of close contacts in Beijing, the capital of China and a developed city. For Yunnan, a typical tourist city, tour guides, passengers, and company staff accounted for 21.62, 13.51, and 10.81% of close contacts, respectively. Tourists and overseas students, with proportions of 30 and 10%, respectively, were the two most frequent types of close contact in Zhejiang.

Table 6 lists the differences between high-risk provinces in terms of the region feature. For some provinces, especially border provinces with imported cases, apart from the domestic high-risk regions identified, there were some overseas high-risk regions as well.

Specifically, the Philippines accounted for 50% and 6.96% of high-risk regions for Macao and Fujian, respectively; these regions are separated from the Philippines by sea. For Guangdong province, London, UK accounted for 3.75% of high-risk regions. Russia accounted for 43.16% of the high-risk regions for Heilongjiang province, a large province bordering Russia. For the provinces of Liaoning and Zhejiang, Tokyo, Japan, and Moscow, Russia, accounted for 4.21% and 8.51% of the high-risk regions, respectively. For Tianjin, 3.67% and 3.21% of the high-risk regions were Paris, France, and Russia, respectively. Thus, for these provinces, special attention should be paid to the detection and isolation of imported cases of COVID-19.

In summary, we conclude that the relevant features of patients in different provincial-level administrative regions can vary significantly. Thus, context-specific regional risk control and prevention measures should be taken to contain the spread of COVID-19.

### 4.2 IRI results

Having built the portrait of diagnosed patients by extracting common features from track data, in this section, we
TABLE 5 Differences in high-frequency contact manners by region

| Province                                      | Proportion of high-frequency close contacts |
|-----------------------------------------------|--------------------------------------------|
| Anhui, Heilongjiang, Guizhou, Chongqing, Liaoning, Shanxi, Fujian, Shanghai, Hebei, Tianjin, Guangdong | Relatives and family members: 90.99, 95.39, 95.06, 94.65, 93.99, 93.98, 93.85, 93.55, 93.53, 93.44, and 93.10%, respectively. |
| Beijing                                      | Relatives and family members (31.77%), operators and suppliers (64.71%) |
| Yunnan                                       | Relatives and family members (43.24%), tour guides (21.62%), passengers (13.51%), company staffs (10.81%) |
| Zhejiang                                     | Relatives and family members (50.00%), tourists (30%), overseas students (10%) |

TABLE 6 Differences in high-frequency regions between different provinces

| Province | High-frequency overseas region (proportion of overseas region) |
|----------|---------------------------------------------------------------|
| Macao    | Philippines (50%)                                             |
| Fujian   | Philippines (6.96%)                                            |
| Guangdong| London, UK (3.75%)                                             |
| Heilongjiang| Vladivostok, Russia (20.48%), Moscow, Russia (14.67%), Russia (8.01%) |
| Liaoning | Tokyo, Japan (4.21%)                                           |
| Tianjin  | Paris, France (3.67%), Russia (3.21%)                          |
| Zhejiang | Moscow, Russia (8.51%)                                         |

construct an infection risk indicator system to calculate the IRI value based on the identified common features of the diagnosed patients.

4.2.1 Infection risk indicator system

Using the common features extracted from the track data of COVID-19 patients, we construct an infection risk indicator system. Specifically, as discussed in Section 2.2, the infection risk indicator system consists of two levels of indicators. The first-level indicators are six common features (place, region, close-contact person, contact manner, travel mode, and symptom). As discussed in section 4.1, after manually classifying 1200 different words recorded in track data into different categories, 246, 543, 105, 153, 69, 84 words are classified into features of place, region, close-contact person, contact manner, travel mode, and symptom, respectively.

For each first-level indicator, we select the top 10 high-frequency words as second-level indicators, to obtain a total of 60 second-level indicators. Then, for each second-indicator level, we calculate the weight, which represents the importance of the indicator in the calculation of the IRI. Based on the constructed infection risk indicator system, we aim to use the top 10 important daily risk activities recorded in patients’ track data of each feature, which are important for infection risk estimation to measure individual infection risk levels.

To support the rationality of selecting the top 10 high-frequency words for each feature category into the infection risk indicator system, we provide both numbers and probabilities of the top 10 high-frequency words under each feature in Table A1 in Appendix A. Table A1 shows that for features of place, region, close-contact person, contact manner, travel mode and symptom, the accumulative word frequencies of top 10 high-frequency words are 16,795, 7543, 5260, 8931, 11,602, and 7507, accounting for 59.82, 50.04, 53.17, 50.63, 55.21, and 56.99% of the frequency of total words under the corresponding feature, respectively.

We can see that the accumulative frequency proportion of the top 10 high-frequency words for each feature ranges from 50.04 to 59.82%, which is relatively close, and all more than 50% of the total frequency of words are classified into the corresponding feature category. Thus, for each feature, the amount of information reflected by the top 10 high-frequency words is relatively close and all exceed 50% of the total information contained in their corresponding features. On the whole, the number of high-frequency words selected into the infection risk indicator system accounts for 5% (60/1200) of the total different words recorded in track data, while the accumulative word frequency of the selected 60 high-frequency words accounts for 54.96% of the frequency of the total words. It shows that the small number of high-frequency words selected by us can reflect most of the track information of diagnosed COVID-19 patients.

Besides, Table A1 also shows that for each feature, the word frequency of the tenth high-frequency word accounts for about 1% of the frequency of the total words under the corresponding feature. Thus, when adding additional words into the infection risk indicator system, the increase of word frequency proportion brought by each word will be about 1% or less. Therefore, in addition to the selected 60 high-frequency words, adding other words into the indication system will bring less information increase, but will make the
indicator system more complex, which makes it unable to calculate the individual infection risk simply and efficiently. Therefore, it is reasonable to select the top 10 high-frequency words for each feature into the infection risk indicator system. Table 7 shows the infection risk indicator system, which records the word frequency and the risk weight for each second-level indicator. Specifically, the secondary indicators and their risk weights (in parentheses) under the first-level indicator of place were hospital (0.1357), residential district (0.0248), center for disease control (0.0222), supermarket (0.0217), COVID-19 designated hospital (0.0214), port (0.0189), airport (0.0156), hotel (0.0147), clinic (0.0089), and market (0.0081). For the first-level indicator of contact manner, the secondary indicators and their risk weights included gathering (0.0347), close contact (0.0322), shopping (0.0222), dining together (0.0180), in contact with others (0.0107), riding together (0.0095), going to work (0.009), dining out (0.0073), having meals (0.0064), and grocery shopping (0.0052).

For the first-level indicator of symptom, the secondary indicators and their risk weights were fever (0.0646), cough (0.0174), abnormal temperature (0.0102), indisposition (0.0097), asymptomatic (0.0083), fatigue (0.0067), medication (0.0044), expectoration (0.0033), dry cough (0.003), and sore throat (0.0027). The secondary indicators and their risk weights under the first level indicator of close-contact person were relatives (0.0127), husband (0.0124), family member (0.0101), wife (0.0089), parents (0.0085), son (0.0082), household (0.0081), infected person (0.008), mother (0.0078), and father (0.0068). For the first-level indicator of region, the secondary indicators and their risk weights were Wuhan (0.0313), Suifenhe (0.0299), Vladivostok (0.0174), Moscow (0.0154), Russia (0.0081), Beijing (0.0063), Hubei (0.0057), Haikou (0.0055), Chongqing (0.0052), and Wanzhou district of Chongqing city (0.0046).

For the secondary indicators under the first level indicator of travel mode, we found that four high-frequency words represented the same meaning. Specifically, in Chinese, “自驾”，“自驾车”，“开车,” and “驾车” are four different words that describe the behavior of someone driving a vehicle. Using these four high-frequency words with the same meaning as four different secondary indicators would not satisfy the condition of mutual exclusion between indicators.

Thus, we combine these four words with the same meaning to form a combined secondary indicator termed “self-driving,” which represents the behavior of driving a vehicle oneself. Accordingly, the risk weight of the “self-driving” secondary indicator was set as the sum of the frequency proportions of the four individual words (“自驾”, “自驾车”, “开车,” and “驾车”) in the track data of the COVID-19 patients. Overall, the final secondary indicators and their risk weights were self-driving (0.0458), plane (0.0392), walking (0.0296), private vehicle (0.0254), taxi (0.0183), bus (0.0134), coach (0.0119), train (0.0064), high-speed rail (0.0064), and online-car hailing (0.0052).

To summarize, when the secondary indicators were summed, the weight rankings of the first-level indicators were place (0.2920), travel mode (0.2016), contact manner (0.1552), symptom (0.1303), region (0.1294), and close contact (0.0915). This ranking indicates the relative importance of the common features extracted from textual track data of patients in terms of contributing to the IRI of COVID-19.

4.2.2 | High-risk activities for contracting COVID-19

Based on the constructed infection risk indicator system, we further summarize the activities that put a person at risk of getting COVID-19.

In practice, the Texas medical association (TMA) ranked a series of activities from high-risk to low-risk in terms of contracting COVID-19, based on the expert judgment of physicians from the TMA COVID-19 Task Force and the TMA Committee on Infectious Diseases (TMA, 2020). The physicians were asked to assign a risk level of 1 (least risk) to 10 (highest risk) to each activity. The risk score and risk category for each risk event are listed in Table B1 in Appendix B. Specifically, activities of opening the mail, ordering restaurant takeout, and pumping gasoline are ranked in the low-risk category, whereas activities like grocery shopping, going for a walk, and eating outside at a restaurant are listed in the moderate-low risk category. Activities such as having dinner at the home of someone else, shopping at a mall, and working in an office building are ranked in the moderate risk category. Moderate-high activities include going to a hair salon, eating inside at a restaurant, and traveling by plane. High-risk activities include working out in a gym, going to a large concert or sports stadium, and drinking at a bar.

In contrast to predicting the infection risk associated with social activities based on expert opinions, we opt to identify activities exposing people to the risk of contracting COVID-19 based on track data of COVID-19 patients. Specifically, the secondary indicators in the infection risk indicator system were the high-frequency words that appeared in the track data of COVID-19 patients, and these indicators represent the high-risk activities with respect to contracting COVID-19. Thus, by combining similar indicators and expanding each indicator into an event, we obtain a total of 46 different risk events. The risk weight for each event is the sum of the weights of the underlying secondary risk indicators.

As shown in Figure 8, these 46 risk events are divided into three risk categories by weights of 0.020 and 0.008. Specifically, risk events including staying at a hospital, having a fever, staying with family, driving a vehicle, taking a plane, staying at a crowded place, having close contact with others, going to Wuhan, going to Suifenhe, going for a walk, taking a private vehicle, staying in a community with diagnosed cases, shopping at a mall, going to a market, and having a cough will put people at extremely high risk of getting COVID-19. Activities such as going to a port, taking a taxi, gathering for dinner, going to Vladivostok, staying at an airport, going to Moscow, staying at a hotel, dining out, taking a bus,
## Table 7 COVID-19 Infection Risk Indicator System Based on Textual Track Data

| First-level indicators | Second-level indicators | Frequency | Weight |
|------------------------|-------------------------|-----------|--------|
| Place                  | Hospital                | 7807      | 0.1357 |
|                        | Residential district    | 1426      | 0.0248 |
|                        | Centre for disease control | 1277    | 0.0222 |
|                        | Supermarket             | 1250      | 0.0217 |
|                        | COVID-19 designated hospital | 1230   | 0.0214 |
|                        | Port                    | 1085      | 0.0189 |
|                        | Airport                 | 897       | 0.0156 |
|                        | Hotel                   | 845       | 0.0147 |
|                        | Clinic                  | 512       | 0.0089 |
|                        | Market                  | 466       | 0.0081 |
| Travel mode            | Self-driving            | 2634      | 0.0458 |
|                        | Plane                   | 2258      | 0.0392 |
|                        | Walking                 | 1703      | 0.0296 |
|                        | Private vehicle         | 1462      | 0.0254 |
|                        | Taxi                    | 1056      | 0.0183 |
|                        | Bus                     | 773       | 0.0134 |
|                        | Coach                   | 687       | 0.0119 |
|                        | Train                   | 368       | 0.0064 |
|                        | High-speed rail         | 365       | 0.0064 |
|                        | Online car-hailing      | 296       | 0.0052 |
| Contact manner         | Gathering               | 1997      | 0.0347 |
|                        | Close contact           | 1852      | 0.0322 |
|                        | Shopping                | 1277      | 0.0222 |
|                        | Dining together         | 1038      | 0.0180 |
|                        | In contact with others  | 615       | 0.0107 |
|                        | Riding together         | 544       | 0.0095 |
|                        | Going to work           | 518       | 0.009 |
|                        | Dining out              | 422       | 0.0073 |
|                        | Having meals            | 367       | 0.0064 |
|                        | Grocery shopping        | 301       | 0.0052 |
| Symptom                | Fever                   | 3717      | 0.0646 |
|                        | Cough                   | 1002      | 0.0174 |
|                        | Abnormal temperature    | 589       | 0.0102 |
|                        | Indisposition           | 559       | 0.0097 |
|                        | Asymptomatic            | 480       | 0.0083 |
|                        | Fatigue                 | 383       | 0.0067 |
|                        | Medication              | 256       | 0.0044 |
|                        | Expectoration           | 190       | 0.0033 |
|                        | Dry cough               | 175       | 0.0030 |
|                        | Sore throat             | 156       | 0.0027 |
| Close-contact person   | Relatives               | 732       | 0.0127 |
|                        | Husband                 | 714       | 0.0124 |
|                        | Family member           | 583       | 0.0101 |
|                        | Wife                    | 510       | 0.0089 |
|                        | Parents                 | 491       | 0.0085 |

(Continues)
visiting relatives, taking a coach, having contact with others, feeling unwell, taking transportation with others, going to the workplace, being symptomatic, going to a market, going to Russia, and staying with infected people are ranked in the high-risk category. Activities such as feeling tired, going to Beijing, going to Hubei, going to Haikou, online-car hailing, going grocery shopping, going to Chongqing, going to Wanzhou District, taking medication, having an expectation, or headache are listed in the moderate-high risk category.

Compared with the risk events identified by the TMA (Table B1), we found some similarities and differences in the risk categories of social activities between China and the United States during the COVID-19 pandemic. In both the United States and China, activities of eating at a restaurant and traveling by plane are high-risk events. However, activities such as meeting relatives and friends, shopping at a mall, working in an office building, and staying at a hospital or hotel, grocery shopping which are ranked in the moderate- or moderate-low-risk category in the United States, were moderate-high risk, high-risk, and even extremely-high-risk events in China. The high-risk activities of working out in a gym, going to a large concert or sports stadium, and drinking at a bar in the United States were not high-risk events in China. The possible reason for this is that during the Spring Festival, people mainly visit relatives and friends rather than going to recreational places (i.e., gym, concert, sports stadium, bar), many of which are closed during the holiday.

In summary, based on the identified high-risk activities in China, awareness regarding how to lower the personal risk of COVID-19 infection can be spread. The Chinese public should try to avoid high-risk activities to protect themselves. If their employment involves high-risk activities, they should wash their hands frequently and wear masks.

### 4.2.3 Infection risk index for COVID-19 diagnosed patients

Based on the constructed infection risk indicator system, we calculate the infection risk index for each of the 9455 COVID-19 diagnosed patients in China. The probability density distribution of the infection risk shown in Figure 9, and Table 8 lists the statistics of the infection risk distribution.

Specifically, the IRI values of diagnosed patients range from 0 to 43.19. A higher value corresponds to a higher statistical risk of contracting COVID-19. The standard deviation is 0.03. The values of the mean and median are 2.67 and 2.27, respectively. The infection risk distribution is skewed to the right, with sharp peaks and thick tails (skewness 2.40, kurtosis 20.68), which indicates that there exist cases with extremely high IRI values.

We further analyze the differences in the individual IRI for various age groups. As discussed in Section 3.2.3, among the total 9455 pieces of track data, 7930 contain the valid age information regarding the patients. We divide these 7930 patients into nine age groups: 0–9, 10–19, 20–29, 30–39, 40–49, 50–59, 60–69, 70–79, and 80 years and above. We classify people aged 0–29, 30–59, and >60 years as young, middle-aged, and elderly, respectively.

By setting 10 as the score of the infection risk index as the boundary, we analyze which age groups people with higher infection risk index belong to. Table 9 lists the number of patients with an index value larger than 10 in various age groups. The number of such middle-aged patients was 51 (22 patients aged 30–39, 12 patients aged 40–49, 17 patients aged 50–59), which was much higher than the corresponding numbers of elderly people (12) (12 patients aged 60–69, 0 patients aged above 70) and young people (11) (0 patients aged 0–19, 11 patients aged 20–29). Thus, higher IRI values corresponded to middle-aged people, which indicates

### Table 7 (Continued)

| First-level indicators | Second-level indicators  | Frequency | Weight |
|------------------------|--------------------------|-----------|--------|
| Son                    |                          | 471       | 0.0082 |
| Household              |                          | 466       | 0.0081 |
| Infected person        |                          | 459       | 0.008  |
| Mother                 |                          | 446       | 0.0078 |
| Father                 |                          | 388       | 0.0068 |
| Region                 |                          |           |        |
| Wuhan                  |                          | 1804      | 0.0313 |
| Suifenhe               |                          | 1722      | 0.0299 |
| Vladivostok            |                          | 1000      | 0.0174 |
| Moscow                 |                          | 887       | 0.0154 |
| Russia                 |                          | 469       | 0.0081 |
| Beijing                |                          | 363       | 0.0063 |
| Hubei                  |                          | 327       | 0.0057 |
| Haikou                 |                          | 317       | 0.0055 |
| Chongqing              |                          | 297       | 0.0052 |
| Wanzhou District       |                          | 267       | 0.0046 |
that middle-aged people were at a higher risk of COVID-19 infection.

Compared with young and old people, who mainly stay at school or home, the middle-aged may have more contact with others and a wider range of travel due to work. Since the spread of COVID-19 is mainly through person-to-person contact during daily activities, middle-aged people have a higher risk of contracting COVID-19. By analyzing the high-risk activities shown in Figure 8, we also found that some of these activities are performed primarily by middle-aged people, such as gathering for dinner, traveling in a taxi or train, and going to the workplace. Thus, middle-aged people engaged in more high-risk activities and were consequently at a higher risk of contracting COVID-19.

By delineating the clinical characteristics of patients in China, experts led by Zhong Nanshan arrived at a similar conclusion, that is, the diagnosed cases were mainly middle-aged people (Guan et al., 2020). Figure 4 in Section 3.2.2 also shows that 62.94% of the collected diagnosed cases with age information are middle-aged people. The clinical characteristic of more middle-aged patients proves the effectiveness of our proposed approach in measuring individual infection risk to a great extent.

In addition, a previous study found that age is one of the key factors associated with death from COVID-19 (Williamson et al., 2020). Although infection risk for the elderly is relatively low, older people had a greater risk of COVID-19-related death than young people did. Once an old person is diagnosed, the possibility of death is relatively high. WHO also determined that protecting vulnerable groups, including the elderly and people with underlying diseases, is one of the priorities of COVID-19 prevention (WHO,
FIGURE 9 The probability density distribution of infection risk. Note: The part of the curve corresponding to values less than 0 on the horizontal axis is not accurate in reality, because the IRI only had positive values.

TABLE 8 Descriptive statistics for infection risk distribution

| Statistics          | Value |
|---------------------|-------|
| Minimum             | 0.00  |
| Maximum             | 43.19 |
| Mean                | 2.67  |
| Median              | 2.27  |
| Standard deviation  | 0.03  |
| Skewness            | 2.40  |
| Kurtosis            | 20.68 |

As shown in Figure 10, the five risk categories from risk-level of low to high are low risk (0, 5.68), moderate-low risk (5.68, 7.02), moderate high (7.02, 9.92), moderate-high (9.92, 19.36), and high risk (19.36 and above). Thus, based on the individual IRI, the risk category of an individual can be determined. For example, at the 99.9% confidence level, the VaR value of infection risk is 19.36, which indicates that we have 99.9% confidence to say that the infection risk index of a diagnosed patient will not exceed 19.36. Thus, if a person’s infection risk index exceeds 19.36, the risk of getting COVID-19 is extremely high.

TABLE 9 The number of patients in various age groups under different risk index ranges

| Risk index age group (years) | ≥20 | 15–19 | 10–14 | Total |
|------------------------------|-----|-------|-------|-------|
| Young                        | 0   | 0     | 0     | 0     |
| 10–19                        | 0   | 0     | 0     | 0     |
| 20–29                        | 0   | 1     | 10    | 11    |
| Middle-aged                  |     |       |       |       |
| 30–39                        | 1   | 2     | 19    | 22    |
| 40–49                        | 0   | 0     | 12    | 12    |
| 50–59                        | 0   | 4     | 13    | 17    |
| Elderly                      |     |       |       |       |
| 60–69                        | 3   | 1     | 8     | 12    |
| 70–79                        | 0   | 0     | 0     | 0     |
| ≥80                          | 0   | 0     | 0     | 0     |
VaR. Thus, the conclusion drawn from this study about the classification of different infection risk levels is reliable.

### 4.4 Robustness test results based on unsupervised topic model latent Dirichlet allocation

Discovering and identifying topics from large amounts of unstructured text is a nontrivial task for social science researchers (Bao & Datta, 2014). Classifying texts into different categories based on manual reading can yield highly accurate results when using the domain knowledge of experts (Huang & Li, 2011). So in this paper, we classify track data of diagnosed COVID-19 patients into six categories by manually performing exhaustive text perusal with experts’ domain knowledge. However, some important topics may be left out by identifying topics from texts through manual reading (Bao & Datta, 2014). Furthermore, manually identifying topics from a mass of text documents is time-consuming and different. It is difficult, indeed infeasible, to manually identify topics from large amounts of texts (Mirakur, 2011).

In this scenario, it is tempting to apply automated text analysis to this important problem. Latent Dirichlet allocation (LDA), which uses the technique of topic modeling, is a popular unsupervised clustering method (Agarwal et al., 2016). It has been widely used to identify different topics from textual data (Huang et al., 2017). Compared with manual identification of topics, using unsupervised LDA to identify topics from mass texts is rapid, convenient, and requires no predefined topic types. Thus, LDA can comprehensively discover text topics, which ensures that no important topics are left out (Bao & Datta, 2014). Therefore, in this paper, we adopt the unsupervised topic model LDA to conduct the robustness test to verify whether there are other categories of patients’ common features that have not been identified from track data based on the manual reading of experts.

However, the unsupervised LDA makes researchers stay away from data, which may lead to biased classification results without appropriate human interpretation and adjustment (Miller, 2017). In this study, we adopt two commonly used measures, precision, and recall, to evaluate the classification accuracy of LDA (Powers, 2011; Wei et al., 2019a). As discussed previously, through manual labeling of high-frequency words, whose cumulative frequency proportions accounted for 81.2% of the total sample data, the high-frequency words are classified into six categories. Classifications obtained through manual labeling are considered the most accurate and are used as the benchmark to measure the classification accuracy of LDA. For convenience, the classification results obtained through manual labeling of the track data are referred to as manual features, and the common patient features identified using the LDA model are termed LDA features.

By observing the words classified into one type of LDA feature, the manual labels of these words are consistent with the label of the LDA feature indicating that these words contained in the LDA feature are correct. Thus, for one type of LDA feature, the precision rate is used to reflect the proportion of words in which the manual labels are consistent with the label of the LDA feature. Mathematically,

$$PR_i = \frac{M_i}{N_i}$$

where $PR_i$ denotes the precision rate of LDA feature $i$, $M_i$ represents that the number of words within LDA feature $i$, whose manual labels are consistent with the label of LDA feature $i$, $N_i$ stands for the number of total words classified into LDA feature $i$. The larger $PR_i$ is, the higher is the proportion of correct words contained in LDA feature $i$, and the higher is the accuracy of the classification result obtained using the unsupervised LDA model.

From the perspective of manual labels of words, words with the same manual label may be classified into different LDA features. Words are considered to be correctly classified when the label of the LDA feature is consistent with the manually assigned label. Thus, for a type of manual feature, the recall rate can be used to measure the proportion of words that are classified into the correct LDA feature. The recall rate is written as

$$RR_i = \frac{W_i}{K_i}$$
where $RR_j$ denotes the recall rate of manual feature $j$, $W_j$ represents the number of words under manual feature $j$ that are classified into the correct LDA feature. $K_j$ denotes the number of total words contained in manual feature $j$. The larger $RR_j$ is, the higher is the proportion of words that are classified into the correct LDA feature, and the higher is the accuracy of the classification result.

Overall, the precision and recall metrics can be used to measure the accuracy of the automatic classification results. Using a confusion matrix, we can calculate the precision and recall rates for each of the identified common patient features, according to Equations (5) and (6), respectively. Then, by averaging the precision rate and recall rate of all the identified features, we can obtain the averaged precision rate and averaged recall rate, respectively. The precision and recall rates of each feature are used to reflect the classification accuracy of the corresponding feature. Thus, the averaged precision and averaged recall rates reflect the average level of classification accuracy when using the unsupervised LDA approach to identify common features of COVID-19 patients.

The robustness test results for LDA are as follows. In the empirical analysis, after removing stop-words as described in Section 4.1, we obtained a total of 129,093 words, among which 9383 were unique. We selected the top 1200 words based on frequency, which appeared a total of 104,866 times and whose cumulative frequency accounted for 81.2% of the total words obtained, to perform manual classification. To make the LDA classification results can be compared with the results of the manual classification in terms of classification accuracy, we also use the 104,866 words, which account for 81.2% of the total sample and have been labeled manually to conduct LDA analysis.

To determine the appropriate number of classifications, which affects the precision of classification, LDA uses the perplexity obtained via 10-fold cross-validation, as in Blei and Lafferty (2007), to reflect the precision of classification. As shown in Figure 11, the perplexity tends to converge since 15. By analyzing the classification results with different numbers of topics, we found that the LDA model performs best for providing appropriate and clear classifications with 25 topics. Thus, the number of topics is set to 25.

We give these 25 topics meaningful labels using the manual labeling procedure mentioned in Section 2.1. The 25 automatic classifications include COVID-19 designated hospital, airport, private vehicle, drive, Suifenhe city, hospital, train, Fuzhou city, customs, close contact, fever, hotel, taxi, relative, Wuhan city, high-speed rail, center for disease control, airplane, supermarket, bus, unwell, port, gather together, walk, and cough, respectively. The word clouds of these 25 classifications are visualized shown in Figure C1 of Appendix C, in which the font size corresponds to the frequency of the word occurring in the classification.

Then, based on the domain knowledge of experts, the discovered 25 classifications with the same label or similar meaning are merged. Specifically, the topics of private vehicle, drive, train, taxi, high-speed rail, airplane, bus, and walk are combined to form a common feature termed travel mode. The topics of COVID-19 designated hospital, airport, hotel, hospital, port, customs, center for disease control, and supermarket are combined to form the common feature of place. The common feature of region is obtained by combining the topics of Suifenhe city, Fuzhou city, and Wuhan city. The topics of close contact and gathering together are combined to form the common feature referred to as contact manner. Relative is the common feature of close-contact person. The topics of fever, unwell, and cough are combined together to form the common feature of symptom. The numbers of words classified into the common features of place, travel mode, region, contact manner, symptom, and close-contact person are 36,535, 27,569, 16,558, 12,710, 9711 and 1783, which respectively account for 34.84, 26.29, 15.79, 12.12, 9.26, and 1.70% of the total sample (Figure 12).

Finally, we obtain six classifications, that is, place, travel mode, contact manner, region, symptom, and close-contact person, which are consistent with the result of manual classification. The word clouds of these identified six common features of COVID-19 patients are shown in Figure C2 in Appendix C. Thus, we conclude that the six manually identified common features of COVID-19 patients are comprehensive and that there are no other categories of patients’ common features that have not been identified from track data.
TABLE II The confusion matrix and the results of precision and recall rates of the LDA model

| Feature      | Place       | Travel mode | Contact manner | Symptom     | Close-contact person | Region      | Recall rate |
|--------------|-------------|-------------|----------------|-------------|----------------------|-------------|-------------|
| Place        | 27,525      | 3435        | 675            | 326         | 195                  | 2788        | 78.77%      |
| Travel mode  | 928         | 17,509      | 427            | 98          | 78                   | 1475        | 85.35%      |
| Contact manner | 837       | 2040        | 10,668         | 89          | 90                   | 1659        | 69.35%      |
| Symptom      | 5,053       | 703         | 108            | 8865        | 32                   | 956         | 56.40%      |
| Close-contact person | 617 | 1583      | 368            | 56          | 1300                 | 358         | 30.36%      |
| Region       | 1575        | 2299        | 464            | 277         | 88                   | 9322        | 66.47%      |
| Precision rate | 75.34% | 63.51%      | 83.93%         | 91.29%      | 72.93%               | 56.3%       | 64.45%      |
|              |             |             |                |             |                      |             | 73.88%      |

Having obtained the classification results using the LDA model, then we calculate the two measures of precision rate and recall rate to assess the accuracy of the LDA classification results based on the confusion matrix (Table II).

From Table II, we can see that the averaged values of precision rate and recall rate are 73.88 and 64.45%, respectively, indicating that the classification accuracy of the LDA model is relatively satisfactory. Then, by analyzing the precision and recall rates for each common feature, we found that the accuracies of the features of place, travel mode, contact manner, and region are relatively high, indicating that the classification accuracies of these four features are satisfactory. Specifically, the precision and recall rates of the place feature are 75.34 and 78.77%, respectively. The travel mode feature has precision and recall rates of 63.51 and 85.35%, respectively. For the contact manner feature, the precision and recall rates are 83.93 and 69.35%, respectively. The precision rate and recall rate of the region feature are 56.30 and 66.47%, respectively.

Moreover, the classification accuracy of the unsupervised LDA model depends on the size of the sample. The larger the sample size is, the more accurate the classifications become. Thus, when processing a large number of patients’ track data, the classification accuracy of LDA will be further improved.

To sum up, by conducting a robustness test using the unsupervised LDA model, we found that the common features identified using the LDA model are consistent with those identified based on manual classification. This indicates that the six manually identified common features of COVID-19 patients are comprehensive and that there are no other categories of common features that have not been identified from track data.

5 | CONCLUSIONS

The main contribution is that this paper introduces the track data into the personal infection risk assessment for the first time by proposing a textual track-data-based infection risk measurement approach. By calculating the IRI values based on common features extracted from track data of patients, we determine the risk categories for individuals with regard to contracting COVID-19. The proposed approach can effectively identify people at a high risk of COVID-19 infection and provide guidance on reducing high-risk daily activities to further lower the infection risk.

Through experiments based on the collected textual track data of 9455 COVID-19 patients in China over the period of January 20, 2020 to July 30, 2020, we construct a general portrait of patients, which is composed of six common features: place, region, close-contact person, contact manner, travel mode, and symptom. Based on the IRI values for all 9455 patients, which range from 0 to 43.19, the individual infection risk levels are classified into five categories: low (0, 5.68), moderate-low (5.68, 7.02), moderate (7.02, 9.92), moderate-high (9.92, 19.36), and high risk (19.36 and above). Individuals with an IRI of more than 19.36 are classified into the high-risk group; these individuals should avoid high-risk activities to lower the risk of contracting COVID-19.

By comparing the high-risk activities identified in our study with those identified by TMA experts, we found that...
there are some differences between the high-risk activities in China and the United States. In addition, middle-aged people were found to be at a higher risk of getting COVID-19, which is consistent with the clinical characteristics of more middle-aged patients; this further proves the effectiveness of the proposed approach in measuring individual infection risk of getting COVID-19.

Thus, with regard to the novel coronavirus diseases that have recently become a significant threat to humanity, this paper proposes a unique approach to effectively identify individuals at high risk of COVID-19 infection, which solves the important problem of identifying high-risk populations in epidemic prevention and control. The proposed approach is highly significant in terms of policy formulation and practical applications. This approach can be used to identify individuals who are at high risk of contracting coronavirus disease as well as high-risk daily activities that can lead to infection. Thus, this approach can allow governments to adopt effective measures, such as nucleic acid testing, isolation of high-risk groups, and appealing to people to avoid high-risk activities, to effectively prevent and control the spread of the coronavirus disease. Besides, the proposed approach is generalizable and can be used for identifying high-risk populations of other countries or other coronavirus infectious diseases with similar transmission modes as COVID-19, which has practical significance for the prevention and control of the spread of coronavirus infectious diseases.

Nevertheless, this study has certain limitations. One is that the disclosure format of track data varied by province. The amount of information contained in the disclosed track data may be detailed or rough, which influenced the effectiveness of the proposed approach to some extent. In the future, the standardized disclosure of track data of patients would enhance the effectiveness of our approach and assist in providing better support for disease prevention and control. The other is that the influence of vaccination injection on individual infection risk of COVID-19 has not been analyzed since the large-scale vaccination has not been implemented when the study was conducted. Someone with movements similar to that of a diagnosed patient will have a significantly higher risk of contracting COVID-19, which would be less likely to happen if the person has taken the shot. Therefore, in future research, we will take the effect of people getting vaccinated into account in measuring the individual infection risk of COVID-19.

ACKNOWLEDGMENTS

This research was funded by grants from the National Natural Science Foundation of China (NSFC Grant Numbers: 72001223, 71850008) and the Engineering Research Center of National Financial Security, Ministry of Education.

REFERENCES

Agarwal, S., Chen, V. Y. S., & Zhang, W. (2016). The information value of credit rating action reports: A textual analysis. Management Science, 62(8), 2218–2240. https://doi.org/10.1287/mnsc.2015.22432016:mnsc. 2015.2243

Ajai, A., Oyedele, L., Owolabi, H., Akinade, O., & Akanbi, L. (2019). Deep learning models for health and safety risk prediction in power infrastructure projects. Risk Analysis, 40(10), 2019–2039. https://doi.org/10.1111/ risa.13425

An, Y., Liang, J., Schild, S. E., Bues, M., & Liu, W. (2017). Robust treatment planning with conditional value at risk chance constraints in intensity-modulated proton therapy. Medical Physics, 44(1), 28–36. https://doi.org/10.1002/mp.12001

Au, A., Moser, N., Rodriguez-Manzano, J., & Georgiou, P. (2018). Live demonstration: A mobile diagnostic system for rapid detection and tracking of infectious diseases. Paper presented at the 2018 IEEE International Symposium on Circuits and Systems (ISCAS), Florence, Italy.

Bao, Y., & Datta, A. (2014). Simultaneously discovering and quantifying risk types from textual risk disclosures. Management Science, 60(6), 1371–1391. https://doi.org/10.1287/mnsc.2014.1930

Belles, J., Guillem, M., & Santolino, M. (2014). Beyond value-at-risk: Glue-Var distortion risk measures. Risk Analysis, 34(1), 121–134. https://doi. org/10.1111/risa.12080

Bharti, K. K., & Singh, P. K. (2016). Opposition chaotic fitness mutation based adaptive inertia weight BPSO for feature selection in text clustering. Applied Soft Computing, 43, 20–34. https://doi.org/10.1016/j.asoc.2016.01.019

Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of science. The Annals of Applied Statistics, 1(1), 17–35. https://doi.org/10.1214/07-AOAS114

Cao, T., Mu, W., Gou, J., & Peng, L. (2020). A study of risk relevance reasoning based on a context ontology of railway accidents. Risk Analysis, 40(8), 1589–1611. https://doi.org/10.1111/risa.13506

Chang, J., Boyd-Graber, J. L., Gerrish, S., Wang, C., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, & A. Culotta (Eds.), Advances in neural information processing systems (pp. 288–296). Curran Associates.

Dong, D., & Wang, Y. (2016). Challenges of rare diseases in China. Lancet, 387(10031), 1906–1906. https://doi.org/10.1016/S0140-6736(16)30418-4

Feldman, L., & Hart, P. S. (2018). Is there any hope? how climate change news imagery and text influence audience emotions and support for climate mitigation policies. Risk Analysis, 38(3), 585–602. https://doi.org/10.1111/risa.12868

Gao, M., Li, T., & Huang, P. (2019) Text classification research based on improved Word2vec and CNN. In X. Liu, Mrissa, M., Zhang, L., Benslimane, D., Ghose, A., Wang, Z., Bucchiarone, A., Zhang, W., Zou, Y., & Yu, Q. (Eds.), Service-oriented computing – ICSOC 2018 Workshops. ICSOC 2018. Lecture Notes in Computer Science, (Vol. 11434). Springer. https://doi.org/10.1007/978-3-030-17642-6_11

Gautret, P., Lagier, J. C., Parola, P., Hoang, V. T., & Raoult, D. (2020). Hydroxychloroquine and azithromycin as a treatment of covid-19: Results of an open-label non-randomized clinical trial. International Journal of Antimicrobial Agents, 56, 105949. https://doi.org/10.1016/j. ijantimicag.2020.105949

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. Nature, 457(7232), 1012–1014. https://doi.org/10. 1038/nature07634

Guan, W., Ni, Z., Hu, Y., Liang, W., Ou, C., He, J., Liu, L., Shan, H., Lei, C., Hui, D., Du, B., Li, L., Zeng, G., Yuen, K., Chen, R., Tang, C., Wang, T., Chen, P., Xiang, J., & Zhong, N. (2020). Clinical characteristics of 2019 novel coronavirus infection in China. medRxiv, https://doi.org/10.1101/ 2020.02.06.20020974

Hashemi, S. J., Khan, F., & Ahmed, S. (2019). An insurance model for risk management of process facilities. Risk Analysis, 39(3), 713–728. https:// doi.org/10.1111/risa.13179

He, Y. Y., Lindbergh, S., Graves, C., & Rakas, J. (2021). Airport exposure to lightning strike hazard in the contiguous United States. Risk Analysis, 41(8), 1323–1344. https://doi.org/10.1111/risa.13630

Huang, A. H., Lehavy, R., Zang, A. Y., & Zheng, R. (2017). Analyst information discovery and interpretation roles: A topic modeling approach.
Management Science, 64(6), 2833–2855. https://doi.org/10.1287/mnsc.2017.2751

Huang, K. W., & Li, Z. L. (2011). A multilabel text classification algorithm for labelling risk factors in SEC form 10-K. ACM Transactions on Management Information Systems (TMIS), 2(3), 1–19. https://doi.org/10.1145/2019618.2019624

Kaur, S., Bhervhani, H., Gulia, S., Vijay, R., & Kumar, R. (2020). Understanding COVID-19 transmission, health impacts and mitigation: Timely social distancing is the key. Environment, Development and Sustainability, 23(5), 6681–6697. https://doi.org/10.1007/s10668-020-00884-x

Kuchler, T., Russel, D., & Stroebel, J. (2020). The geographic spread of covid-19 correlates with structure of social networks as measured by Facebook. CSESifo Working Paper Series. NBER.

Lee, E. C., Asher, J. M., Goldlust, S., Kraemer, J. D., Lawson, A. B., & Bansal, S. (2016). Mind the scales: Harnessing spatial big data for infectious disease surveillance and inference. The Journal of Infectious Diseases, 214, 409–413. https://doi.org/10.1093/infdis/jiw544

Li, X., Shang, W., & Wang, S. (2018). Text-based crude oil price forecasting: A deep learning approach. International Journal of Forecasting, 35(4), 1548–1560. https://doi.org/10.1016/j.ijforecast.2018.07.006

Liu, M., Liang, B., Zheng, F. F., & Chu, F. (2018). Stochastic airline fleet assignment with risk aversion. IEEE Transactions on Intelligent Transportation Systems, 20(8), 3081–3090. https://doi.org/10.1109/TITS.2018.2871969

Medina, R. A. (2018). 1918 influenza virus: 100 years on, are we prepared against the next influenza pandemic? Nature Reviews Microbiology, 16(2), 61–62. https://doi.org/10.1038/nrmicro.2017.174

Meli, Q. Z., Shen, X. H., & Zhai, C. X. (2007). Automatic labelling of multinomial topic models. In P. Berkhin, R. Caruana, & X. Wu (Eds.), The 13th ACM SIGKDD International conference on knowledge discovery and data mining (pp. 490–499). ACM.

Miller, G. S. (2017). Discussion of “the evolution of 10-K textual disclosure: Evidence from latent Dirichlet allocation”. Journal of Accounting & Economics, 64(2-3), 246–252. https://doi.org/10.1016/j.jacceco.2017.07.004

Miralurk, Y. (2011). Risk disclosure in SEC corporate filings. Working paper, University of Pennsylvania, Philadelphia. http://repository.upenn.edu/wharton_research_scholars/85

Mizumoto, K., & Chowell, G. (2020). Estimating risk for death from coronavirus disease, China, January–February 2020. Emerging Infectious Disease, 26(6), 1251–1256. https://doi.org/10.3201/eid2606.200233

Powers, D. M. W. (2011). Evaluation: From precision, recall and F-factor to ROC, informedness, markedness & correlation. Journal of Machine Learning Technologies, 2, 2229–3981. https://doi.org/10.9735/jemales.2010.3981

Pretenthaler, F., Köberl, J., & Bird, D. L. (2015). ‘Weather value at risk’: A uniform approach to describe and compare sectoral income risks from climate change. Science of The Total Environment, 543, 1010–1018. https://doi.org/10.1016/j.scitotenv.2015.04.035

Ren, H., Zhao, L., Zhang, A., Song, L., & Cui, C. (2020). Early forecasting of the potential risk zones of covid-19 in China’s megacities. Science of The Total Environment, 729, 138995. https://doi.org/10.1016/j.scitotenv.2020.138995

Ronnqvist, S., & Sarlin, P. (2014). Bank networks from text: Interrelations, centrality and determinants. Quantitative Finance, 15(10), 1619–1635. https://doi.org/10.1080/14697688.2015.1071076

Rosenberg, J. V., & Schuermann, T. A. (2006). General approach to integrated risk management with skewed, fat-tailed risks. Journal of Financial Economics, 79(3), 569–614. https://doi.org/10.1016/j.jfineco.2005.03.001

Rustam, F., Ashraf, I., Id, Mehmoond, A., & Khan, Y. (2019). Tweets classification on the base of sentiments for us airline companies. Entropy, 21(11), 1078. https://doi.org/10.3390/e21111078

Shweta, B., Gerardo, C., Lone, S., Alessandro, V., & Viboud, C. (2016). Big data for infectious disease surveillance and modeling. Journal of Infectious Diseases, 214( Suppl 4), 8375–8379. https://doi.org/10.1093/infdis/jiw400

Texas Medical Association (TMA). (2020). COVID-19: Know and lower your risk. https://texmed.inreachce.com/Details/Information/2f05eb43-b28d-4be6-beba5-6c91056a011

Wei, L., Li, G., Zhu, X., & Li, J. (2019a). Discovering bank risk factors from financial statements based on a new semi-supervised text mining algorithm. Accounting & Finance, 59(3), 1519–1552. https://doi.org/10.1111/afi.12453

Wei, L., Li, G., Zhu, X., Sun, X., & Li, J. (2019b). Developing a hierarchical system for energy corporate risk factors based on textual risk disclosure. Energy Economics, 80, 452–460. https://doi.org/10.1016/j.eneco.2019.01.020

Wesolowski, A., Buckee, C. O., Engo-Monsen, K., & Metcalf, C. J. E. (2016). Connecting mobility to infectious diseases: The promise and limits of mobile phone data. The Journal of Infectious Diseases, 214, 414–420. https://doi.org/10.1093/infdis/jiw273

Williamson, E. J., Walker, A. J., Bhaskaran, K., Bacon, S., & Goldacre, B. (2020). Opensively: Factors associated with covid-19 death in 17 million patients. Nature, 584(7821), 430–436. https://doi.org/10.1038/s41586-020-2521-4

Wit, E. D., Van Doremalen, N., Falzarano, D., & Munster, V. J. (2016). SARS and MERS: Recent insights into emerging coronaviruses. Nature Reviews Microbiology, 14(8), 523. https://doi.org/10.1038/nrmicro.2016.81

World Health Organization (WHO). (2020a, February 03). 2019 novel coronavirus (2019-nCoV): Strategic preparedness and response plan. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/strategies-and-plans

World Health Organization (WHO). (2020b, August 03). WHO COVID-19 preparedness and response progress report–1 February to 30 June 2020. https://www.who.int/publications/m/item/who-covid-19-preparedness-and-response-progress-report-1-february-to-30-june-2020

Wu, A., Peng, Y., Huang, B., Ding, X., & Jiang, T. (2020). Genome composition and divergence of the novel coronavirus (2019-ncov) originating in China. Cell Host & Microbe, 27(3), 325–328. https://doi.org/10.1016/j.chom.2020.02.001

Xu, J. P., Hou, S. H., Xie, H. P., Lv, C. W., & Yao, L. M. (2018). Equilibrium approach towards water resource management and pollution control in coal chemical industrial park. Journal of Environmental Management, 219(1), 56–73. https://doi.org/10.1016/j.jenvman.2018.04.080

Zheng, J., Qi, Z., Dou, Y., & Tan, Y. (2019). How mega is the Mega? Exploring the spillover effects of WeChat using graphical model. Information Systems Research, 30(4), 1343–1362. https://doi.org/10.1287/isre.2019.0865

Zhong, B. T., Pan, X., Love, P. E. D., Sun, J., & Tao, C. J. (2020). Hazard analysis: A deep learning and text mining framework for accident prevention. Advanced Engineering Informatics, 46, 101152. https://doi.org/10.1016/j.aei.2020.101152

How to cite this article: Wei, L., Li, X., Jing, Z., & Liu, Z. (2023). A novel textual track-data-based approach for estimating individual infection risk of COVID-19. Risk Analysis, 43, 156–182. https://doi.org/10.1111/risa.13944
**APPENDIX A**

**TABLE A1** The word frequency and proportion of the top 10 high-frequency words for each feature category

| Feature                          | Frequency of total words | Top 10 high-frequency words | Frequency of top 10 high-frequency words | Frequency proportion of top 10 high-frequency words |
|----------------------------------|--------------------------|-----------------------------|------------------------------------------|---------------------------------------------------|
| Place                            | 28,076                   | Hospital 7807               | 27.81%                                   |                                                   |
|                                  |                          | Residential district 1426   | 5.08%                                    |                                                   |
|                                  |                          | Centre for disease control 1277 | 4.55%                                |                                                   |
|                                  |                          | Supermarket 1250            | 4.45%                                    |                                                   |
|                                  |                          | COVID-19 designated hospital 1230 | 4.38%                                |                                                   |
|                                  |                          | Port 1085                   | 3.86%                                    |                                                   |
|                                  |                          | Airport 897                 | 3.19%                                    |                                                   |
|                                  |                          | Hotel 845                   | 3.01%                                    |                                                   |
|                                  |                          | Clinic 512                  | 1.82%                                    |                                                   |
|                                  |                          | Market 466                  | 1.66%                                    |                                                   |
|                                  |                          | Top 10 accumulative total 16,795 | 59.82%                               |                                                   |
| Region                           | 15,074                   | Wuhan 1814                  | 12.03%                                   |                                                   |
|                                  |                          | Sui fenhe 1732             | 11.49%                                   |                                                   |
|                                  |                          | Vladivostok 1010           | 6.70%                                    |                                                   |
|                                  |                          | Moscow 897                 | 5.95%                                    |                                                   |
|                                  |                          | Russia 479                 | 3.18%                                    |                                                   |
|                                  |                          | Beijing 373                | 2.47%                                    |                                                   |
|                                  |                          | Hubei 337                  | 2.24%                                    |                                                   |
|                                  |                          | Haikou 327                 | 2.17%                                    |                                                   |
|                                  |                          | Chongqing 307              | 2.04%                                    |                                                   |
|                                  |                          | Wanzhou district (Chongqing city) 267 | 1.77%                                |                                                   |
|                                  |                          | Top 10 accumulative total 7543 | 50.04%                               |                                                   |
| Close-contact person             | 9893                     | Relatives 732              | 7.40%                                    |                                                   |
|                                  |                          | Husband 714                | 7.22%                                    |                                                   |
|                                  |                          | Family member 583          | 5.89%                                    |                                                   |
|                                  |                          | Wife 510                   | 5.16%                                    |                                                   |
|                                  |                          | Parents 491                | 4.96%                                    |                                                   |
|                                  |                          | Son 471                    | 4.76%                                    |                                                   |
|                                  |                          | Household 466              | 4.71%                                    |                                                   |
|                                  |                          | Infected person 459        | 4.64%                                    |                                                   |
|                                  |                          | Mother 446                 | 4.51%                                    |                                                   |
|                                  |                          | Father 388                 | 3.92%                                    |                                                   |
|                                  |                          | Top 10 accumulative total 5260 | 53.17%                               |                                                   |

(Continues)
| Feature       | Frequency of total words | Top 10 high-frequency words | Frequency of top 10 high-frequency words | Frequency proportion of top 10 high-frequency words |
|---------------|--------------------------|-----------------------------|------------------------------------------|-------------------------------------------------|
| Contact manner| 17,638                   | Gathering                   | 1997                                     | 11.32%                                          |
|               |                          | Close contact               | 1852                                     | 10.50%                                          |
|               |                          | Shopping                    | 1277                                     | 7.24%                                           |
|               |                          | Dining together             | 1038                                     | 5.89%                                           |
|               |                          | In contact with others      | 615                                      | 3.49%                                           |
|               |                          | Riding together             | 544                                      | 3.08%                                           |
|               |                          | Going to work               | 518                                      | 2.94%                                           |
|               |                          | Dining out                  | 422                                      | 2.39%                                           |
|               |                          | Having meals                | 367                                      | 2.08%                                           |
|               |                          | Grocery shopping            | 301                                      | 1.71%                                           |
|               |                          | Top 10 accumulative total   | 8931                                     | 50.63%                                          |
| Travel mode   | 21,013                   | Self-driving                | 2634                                     | 12.54%                                          |
|               |                          | Plane                       | 2258                                     | 10.75%                                          |
|               |                          | Walking                     | 1703                                     | 8.10%                                           |
|               |                          | Private vehicle             | 1462                                     | 6.96%                                           |
|               |                          | Taxi                        | 1056                                     | 5.03%                                           |
|               |                          | Bus                         | 773                                      | 3.68%                                           |
|               |                          | Coach                       | 687                                      | 3.27%                                           |
|               |                          | Train                       | 368                                      | 1.75%                                           |
|               |                          | High-speed rail             | 365                                      | 1.74%                                           |
|               |                          | Online car-hailing          | 296                                      | 1.41%                                           |
|               |                          | Top 10 accumulative total   | 11,602                                   | 55.21%                                          |
| Symptom       | 13,172                   | Fever                       | 3717                                     | 28.22%                                          |
|               |                          | Cough                       | 1002                                     | 7.61%                                           |
|               |                          | Abnormal temperature        | 589                                      | 4.47%                                           |
|               |                          | Indisposition               | 559                                      | 4.24%                                           |
|               |                          | Asymptomatic                | 480                                      | 3.64%                                           |
|               |                          | Fatigue                     | 383                                      | 2.91%                                           |
|               |                          | Medication                  | 256                                      | 1.94%                                           |
|               |                          | Expectoration               | 190                                      | 1.44%                                           |
|               |                          | Dry cough                   | 175                                      | 1.33%                                           |
|               |                          | Sore throat                 | 156                                      | 1.18%                                           |
|               |                          | Top 10 accumulative total   | 7507                                     | 56.99%                                          |
| Total         | 104,866                  | Top 60 accumulative total   | 57,638                                   | 54.96%                                          |
### APPENDIX B

#### TABLE B1  Risk categories of activities in the United States divided by the Texas medical association

| Risk category | Risk event                                                                 | Risk score |
|---------------|---------------------------------------------------------------------------|------------|
| **High risk** | Going to a bar                                                             | 9          |
|               | Attending a religious service with 500+ worshipers                        | 9          |
|               | Going to a sports stadium                                                 | 9          |
|               | Attending a large music concert                                           | 9          |
|               | Going to a movie theater                                                  | 8          |
|               | Going to an amusement park                                                | 8          |
|               | Working out at a gym                                                     | 8          |
|               | Eating at a buffet                                                        | 8          |
| **Moderate-high** | Hugging or shaking hands when greeting a friend                           | 7          |
|               | Playing football                                                          | 7          |
|               | Playing basketball                                                        | 7          |
|               | Traveling by plane                                                        | 7          |
|               | Attending a wedding or funeral                                            | 7          |
|               | Eating in a restaurant (inside)                                           | 7          |
|               | Going to a hair salon or barbershop                                        | 7          |
| **Moderate risk** | Visiting an elderly relative or friend in their home                       | 6          |
|               | Swimming in a public pool                                                 | 6          |
|               | Working a week in an office building                                      | 6          |
|               | Sending kids to school, camp, or day care                                  | 6          |
|               | Shopping at a mall                                                        | 5          |
|               | Going to a beach                                                          | 5          |
|               | Attending a backyard barbecue                                             | 5          |
|               | Having dinner at someone else’s house                                     | 5          |
| **Moderate-low** | Spending an hour at a playground                                          | 4          |
|               | Walking in a busy downtown                                                | 4          |
|               | Eating in a restaurant (outside)                                          | 4          |
|               | Going to a library or museum                                               | 4          |
|               | Sitting in a doctor’s waiting room                                        | 4          |
|               | Staying at a hotel for two nights                                         | 4          |
|               | Playing golf                                                              | 3          |
|               | Going for a walk, run, or bike ride with others                           | 3          |
|               | Grocery shopping                                                          | 3          |
| **Low risk**  | Going camping                                                              | 2          |
|               | Playing tennis                                                            | 2          |
|               | Pumping gasoline                                                           | 2          |
|               | Getting restaurant takeout                                                | 2          |
|               | Opening the mail                                                          | 1          |
FIGURE C1  Word clouds of 25 automatic classifications obtained using the LDA model
**FIGURE C2** Word clouds of six common features identified based on the LDA model