Research on Risk Level Assessment of Respiratory Infectious Disease Transmission: Taking Public Places with Novel Coronavirus Infection as an Example Postprint

Authors: Hu Jian, Liu Min, Zengtao Jiao, Jing Wenzhan, Liang Wannian, Liang Wannian

Date: 2026-02-12T16:44:08+00:00

Abstract

Background During major epidemic prevention and control processes, rapid assessment of infectious disease transmission risk levels serves as an important foundation for subsequent emergency response measures. Objective Based on machine learning algorithms, this study aims to develop a risk assessment model capable of rapidly identifying high-risk venues for respiratory infectious disease transmission, thereby facilitating rapid and precise epidemic prevention and control. Methods Data were collected from 201 venues associated with 43 COVID-19 cases in a district of Beijing during the implementation of China' s "dynamic zero-COVID" policy in April 2022, including venue type, area, ventilation conditions, and personnel flow patterns. Machine learning algorithm models were constructed using the secondary transmission risk value at case activity venues as the dependent variable to assess the risk levels of epidemic-associated venues. Results The performance of Random Forest (RForest), Gradient Boosting Decision Tree (GBDT), and Extreme Gradient Boosting (XGBoost) models for predicting secondary cases among close contacts at venues demonstrated that the GBDT model achieved the best performance with an area under the receiver operating characteristic curve (AUC) = 0.78 and sensitivity = 0.647. The initial Ct value (minimum value) of the index case, Health Kit scan record queries, mask-wearing status of the index case, and duration of stay of the index case were identified as important features in the model. Among specific venue types, Chinese restaurants, fast food/bento establishments, and bars with relatively high foot traffic showed higher relative importance. Conclusion The GBDT assessment model established in this study can provide effective support for risk assessment of venues during respiratory infectious disease outbreaks, assisting policymakers in prioritization and decision-making under resource-limited condi-

tions. Given the potentially similar transmission mechanisms among respiratory infectious diseases, this model may serve as a reference for risk assessment of other respiratory infectious diseases.

Full Text

Research on Risk Level Assessment of Respiratory Infectious Disease Transmission: A Case Study of COVID-19-Infected Public Places

HU Jian[1], LIU Min[2], JIAO Zengtao[3], JING Wenzhan[4], LIANG Wannian[4,5*]

[1]School of Biomedical Engineering, Tsinghua University, Beijing 100084, China [2]Department of Epidemiology and Health Statistics, School of Public Health, Peking University, Beijing 100191, China [3]Yidu Cloud (Beijing) Technology Co., Ltd., Beijing 100083, China [4]Vanke School of Public Health, Tsinghua University, Beijing 100084, China [5]Institute for Health China, Tsinghua University, Beijing 100084, China

*Corresponding author: LIANG Wannian, Professor; E-mail: liangwn@tsinghua.edu.cn

Abstract

Background: During major epidemic prevention and control efforts, rapid assessment of infectious disease transmission risk levels is crucial for implementing emergency response measures.

Objective: To develop a machine learning-based risk assessment model for rapid identification of high-risk venues for respiratory infectious disease transmission, thereby facilitating rapid and precise epidemic prevention and control.

Methods: Data were collected from 201 public places associated with 43 COVID-19 cases in a district of Beijing during the "dynamic zero-COVID" policy implementation in April 2022, including venue type, area, ventilation conditions, and foot traffic. Machine learning models were constructed using the secondary attack risk value of case activity venues as the dependent variable to assess risk levels.

Results: Among the Random Forest (RForest), Gradient Boosting Decision Tree (GBDT), and Extreme Gradient Boosting (XGBoost) models evaluated for predicting close-contact conversion, the GBDT model performed best with an area under the receiver operating characteristic curve (AUC) of 0.78 and sensitivity of 0.647. Important features included the first case' s initial Ct value (minimum), Health Code scan records, mask-wearing status, and dwell time. Stratified by venue type, high-traffic areas such as Chinese restaurants, fast-food outlets, and bars showed higher relative importance.

Conclusion: The GBDT evaluation model established in this study provides an effective tool for risk assessment of respiratory infectious disease transmission in public venues, aiding policymakers in prioritization and decision-making under resource constraints. Given potential similarities in transmission mechanisms

among respiratory infectious diseases, this model may serve as a reference for risk assessment of other respiratory pathogens.

Keywords: Respiratory infectious diseases; Risk assessment; Machine learning; Public health

Introduction

Globally, the prevention and control of respiratory infectious diseases remain a significant challenge in public health. According to WHO data, these diseases cause millions of illnesses annually, imposing substantial disease burdens and socioeconomic losses. Over the past two decades, outbreaks of respiratory infectious diseases such as influenza, severe acute respiratory syndrome (SARS), Middle East respiratory syndrome (MERS), and coronavirus disease 2019 (COVID-19) have not only threatened global health but also profoundly impacted social order and economic development.

Respiratory pathogens spread rapidly and infect large numbers of people, easily causing local outbreaks, epidemics, or pandemics. Therefore, scientific, rapid, and effective control measures are essential to prevent further spread. During epidemics or pandemics of respiratory infectious diseases, rapidly identifying and controlling high-risk venues is critical for curbing disease transmission. With the development of big data and artificial intelligence technologies, integrating AI into epidemic prevention and control can improve the speed and accuracy of risk assessment, better predict epidemic trends, optimize resource allocation, and formulate more effective prevention strategies.

Currently, machine learning algorithms have been widely applied internationally to infectious disease risk assessment and prediction. Researchers abroad have used machine learning models to construct diagnostic models for pediatric community-acquired lower respiratory tract infections, achieving AUROC values of 0.953 [10]. Others have applied machine learning to risk assessment and early warning of bacterial pneumonia, *Mycoplasma pneumoniae* infection, tuberculosis, and other respiratory diseases, demonstrating good predictive performance and accuracy in infectious disease surveillance [7,9-10]. During COVID-19, Greece classified travelers based on country of origin, age, sex, and entry time, updating risk estimates for each category based on recent batch test results [3-4]. In China, increasing attention has been paid to machine learning applications in epidemic prevention and control. For example, studies have constructed risk assessment models by analyzing epidemiological characteristics and activity trajectories of cases [1-2]. During COVID-19, many local governments combined household registration information, personal location, medical records, medication purchase records, travel history, and self-reported health status to create personal infection risk scoring systems (Health Codes), enabling real-time risk assessment and differentiated management measures based on risk levels [5-6]. Although existing research has made progress in risk assessment for specific diseases and scenarios, there remains a lack of universal risk assessment mod-

els applicable to various respiratory infectious diseases. Existing models often rely on specific data sources such as medical health data or social media data, limiting their application to other types of respiratory infectious diseases.

This study aims to develop a universal respiratory infectious disease transmission risk assessment model based on machine learning algorithms. By analyzing characteristic data of cases in different activity venues, the model can rapidly assess the transmission risk level of involved venues, providing policymakers with a rapid and precise decision-support tool for optimizing prioritization and resource allocation under limited resources to address evolving infectious disease challenges.

Methods

1.1 Study Subjects

This study focused on public places associated with COVID-19 cases reported in a district of Beijing during China' s implementation of the "dynamic zero-COVID" policy in April 2022. Specific public places were defined as venues open to the public within a certain time and space range for social activities, exchanges, and consumption (Table). These venues typically feature high population density and mobility, potentially becoming important risk points for epidemic transmission.

Specific public places included but were not limited to: Chinese and Western restaurants, fast-food outlets, cafes, and bars providing catering services; supermarkets, shopping centers, department stores, grocery stores, and farmers' markets for shopping; hotels, hostels, guest houses, homestays, and collective dormitories providing accommodation; kindergartens, primary and secondary schools, universities, and extracurricular training institutions for education; hospitals, clinics, and pharmacies providing medical services and drug sales; office buildings, factories, and construction sites as workplaces; cinemas, theaters, KTVs, gyms, amusement parks, and parks for entertainment; government service departments, community service centers, post offices, banks, and telecommunications offices providing public services; and subway stations, railway stations, bus stations, and airports for transportation.

Inclusion and exclusion criteria for study subjects are shown in Table . Based on these criteria, 201 involved venues were included in the analysis.

1.2 Study Design

Combining field epidemiological investigation, gene sequencing, artificial intelligence technology, and machine learning algorithms, we constructed and validated a risk assessment model for public places involved in epidemics of respiratory infectious diseases, particularly COVID-19. The specific workflow was as follows:

(1) Through field epidemiological investigations, detailed information was collected on COVID-19 patients, including clinical symptoms, activity history, health status, and personal protective equipment usage (data sourced from the National Infectious Disease Surveillance Reporting System, field investigation forms, and epidemiological reports [12]). Gene sequencing technology was used to determine transmission chains between infected individuals and close contacts, calculating the secondary attack rate among close contacts as a key indicator for assessing venue risk.

(2) Using artificial intelligence technologies such as machine vision and facial recognition, activity trajectories and dwell times of infected individuals in public places were recorded in detail. Potential risk factors affecting epidemic transmission were identified through literature review; based on expert discussion, a set of representative venue characteristic indicators were selected to quantify venue risk levels.

1.3 Multi-source Data Fusion Methods

This study employed a multi-source data fusion approach to comprehensively capture and analyze the activity characteristics of COVID-19 cases in specific public places and their potential transmission risks.

Infector characteristic data: Through the "Infectious Disease Reporting System" —a subsystem of the China Information System for Disease Control and Prevention [12]—detailed epidemiological information was collected on infected individuals, including onset time, clinical symptoms, symptom occurrence time, and viral load. Field epidemiological investigation forms and reports were reviewed to obtain case activity trajectories, health status, personal protective measures, and close contact information during the 14 days before onset.

Public place activity data: Using AI machine vision and facial recognition technology, surveillance videos of venues visited by cases during the 14 days before onset were analyzed. This technology identified and calculated contact frequency between individuals, dwell time, and personal protective measure adoption within venues.

Venue characteristic data: Through review of field epidemiological investigation reports and site visits, characteristic data were collected on public places involved in case activity trajectories, including venue type, area, and ventilation conditions.

Venue foot traffic data: Health Code scan data were used as a proxy indicator for venue foot traffic. For venues where scan data could not be obtained, estimates were made using proportionality coefficients determined by experts to obtain relatively accurate foot traffic data.

1.4 Quality Control

All data underwent rigorous preprocessing including data cleaning, missing value imputation, and standardization to ensure data quality and model training accuracy. Numerical variables were uniformly imputed using mean values; for non-numerical variables, the most frequent value within the "venue type" category was used for imputation. To facilitate model training, symptoms at the time of the first infected individual' s visit were converted to "symptomatic" and "asymptomatic," with a new variable "whether the first case was symptomatic" added; discrete variables were converted to One-hot encoding. Additionally, all personally identifiable information was de-identified in compliance with relevant data protection regulations and ethical standards.

During data preprocessing, missing values and outliers were systematically handled. First, for numerical variables (e.g., venue area, dwell time, Health Code scan frequency), missing values were imputed using the mean of that venue category; for categorical variables (e.g., mask-wearing status, venue type), the most frequent value within that category was used to maintain data distribution representativeness. Second, for extreme outliers, the boxplot method combined with expert judgment was used for identification; a small number of extreme values clearly deviating from actual conditions were winsorized to reduce impact on model training. To assess the potential impact of data preprocessing on the model, mean and standard deviation of main feature variables were compared before and after imputation, with no significant shifts observed. These steps aimed to ensure the integrity and reliability of data used for model training while minimizing bias introduced by missing and outlier values.

1.5 Risk Assessment Model Construction

Venue risk indicator determination: Based on previous research and literature review [13-15], two primary indicators and corresponding secondary indicators were determined (Table) to quantify and assess transmission risk of involved venues.

Descriptive analysis of involved venues: Epidemic prevention and control experts annotated 201 involved venues; descriptive statistical analysis was performed before data imputation. "Whether close contacts converted to positive" was the most direct measure of involved venues and was used as the dependent variable. Grouped by this dependent variable, 158 venues were "negative" and 43 were "positive," totaling 201 venue records.

Machine learning algorithm and model selection: The dependent variable was the risk of secondary cases occurring after case visits to venues, calculated using the number of close contacts developing disease associated with case activity venues. Field epidemiological investigations and case specimen gene sequencing were used to determine generational transmission chains between cases and associated close contacts to improve accuracy of determinations. Combined with actual epidemic prevention and control practices in China, the number of

close contacts developing disease at venues was converted to "secondary attack present" and "secondary attack absent," transforming venue risk modeling into a binary classification problem.

For this binary classification problem, logistic regression (LR) served as the classical linear model baseline to evaluate gains from non-linear models. Random Forest (RForest), Gradient Boosting Decision Tree (GBDT), and Extreme Gradient Boosting (XGBoost) are tree-based ensemble learning methods capable of capturing non-linear feature relationships and performing stably in small-sample and multi-feature scenarios [11], and were therefore included for comparison. Model parameters were set primarily based on cross-validation. For example, in the GBDT model, the learning rate was set to 0.1, maximum tree depth to 5, and subsample ratio to 0.8 to balance model complexity and generalization capability. In the XGBoost model, early stopping was employed to avoid overfitting. Overall, the priority goal of parameter tuning was to improve sensitivity to minimize false negatives for high-risk venues. The dataset was randomly divided into training and testing sets at a 7:3 ratio; training set data were used for model training, and the testing set was used to validate the trained model using k-fold cross-validation.

1.6 Statistical Methods

SPSS 25.0 statistical software was used for data analysis. After normality testing, continuous variables were found to follow non-normal distributions and are expressed as median (P25, P75), with comparisons between groups using the Kruskal-Wallis test. Categorical data are expressed as relative frequencies and analyzed using the χ^2 test. To evaluate classification model performance, receiver operating characteristic (ROC) curves and the area under the ROC curve (AUC) were used. $P < 0.05$ was considered statistically significant.

Results

Among venues where close contacts converted to positive, the most common types were Chinese restaurants (20.9%), fast-food bento boxes (9.3%), and bars (7.0%); ventilation was predominantly indoor (over 90%); and mask-wearing by the first case was predominantly "no mask worn throughout." There were no statistically significant differences in venue area, Health Code scan records, environmental ratings, per capita consumption, interval from first case onset, first case initial Ct value, dwell time of first case, venue type, or whether the first case was symptomatic between venues with and without close-contact conversion ($P > 0.05$). However, there were statistically significant differences in review counts, ventilation grade, and mask-wearing status of the first case between venues with and without close-contact conversion ($P < 0.05$), as shown in Table .

Analysis of LR, GBDT, XGBoost, and RForest model performance for predicting close-contact conversion showed that all four classification models achieved

AUC > 0.5 (better than random classification), as shown in Figure [Figure 1: see original paper]. GBDT achieved AUC = 0.78, RForest AUC = 0.71, and LR AUC = 0.581. Specific sensitivity and specificity indices are shown in Table .

For the GBDT model, feature importance rankings are shown in Table . The first case' s initial Ct value (minimum), Health Code scan records, mask-wearing status, dwell time (hours), venue area, and interval from onset (hours) were variables consistently ranking high in feature importance across models. Stratified by venue type, high-traffic areas such as Chinese restaurants, fast-food bento boxes, and bars showed higher relative importance. Ventilation grade showed limited classification effect because original data only distinguished indoor versus outdoor environments, with most locations being indoor.

Discussion

This study compared the performance of four classification models—LR, RForest, GBDT, and XGBoost—for assessing respiratory infectious disease transmission risk levels. ROC curves and AUC values were used to evaluate these models' ability to distinguish high-risk from low-risk venues. Results showed all models achieved AUC > 0.5, indicating certain predictive capability. Among them, the GBDT model achieved the best performance with AUC = 0.78, followed by RForest (AUC = 0.71), while the LR model performed worst with AUC = 0.581. Sensitivity analysis further confirmed the GBDT model' s advantage, showing the best true positive rate, followed by XGBoost, while LR performed poorly in sensitivity. These results indicate that tree-based models, particularly GBDT, perform better in this risk assessment task due to their ability to capture non-linear relationships and complex interactions in the data.

Further analysis of experimental results revealed that the first case' s initial Ct value, Health Code scan records, mask-wearing status, and dwell time were variables ranking high in feature importance. The significance of these features aligns with epidemiological principles, emphasizing the critical role of early identification and intervention in controlling epidemic transmission. The Ct value, as an indicator of viral load, with lower values indicating higher viral loads and potentially increased transmission risk. Health Code scan records provide case activity trajectories and contact history, which are crucial for epidemic source tracing and contact tracing. The significance of mask-wearing status further confirms the role of personal protective measures in blocking respiratory infectious disease transmission, providing data support for public health policy. Additionally, longer dwell times of the first case in venues imply higher potential contact and transmission risk; the significance of this factor in the model suggests that temporal factors should be considered in epidemic management.

In stratified analysis by venue type, we found that high-traffic areas such as Chinese restaurants, fast-food bento boxes, and bars showed higher relative importance in epidemic transmission. This finding emphasizes the necessity of implementing stricter prevention and control measures in high-traffic venues.

These venues are characterized by frequent person-to-person contact and long dwell times, providing more opportunities for virus transmission.

Although ventilation grade showed limited classification effect in this study due to data constraints (only indoor versus outdoor distinction), its epidemiological importance cannot be ignored. Good ventilation is considered an effective means of reducing indoor viral aerosol transmission. Future research should consider more detailed ventilation condition data to more accurately assess its impact on epidemic transmission.

The innovation of this study lies in the comprehensive utilization of multiple data sources and machine learning technologies to identify key factors affecting epidemic transmission and provide targeted insights for epidemic prevention and control in different venue types. This approach not only improves the precision of epidemic risk assessment but also provides a scientific basis for public health decision-making, particularly in guiding prioritization and resource allocation under limited resources. Because respiratory infectious diseases share many similar transmission mechanisms, the findings of this study also have potential generalizability. The methodology of this study can provide reference for risk assessment of other respiratory infectious diseases such as influenza, SARS, and MERS. The transmission of these diseases is similarly affected by case characteristics, venue types, and environmental factors. Therefore, the model and methods of this study can be trained and adjusted using data from different infectious diseases to adapt to risk assessment of different respiratory infectious diseases. The high sensitivity of the model means it can more effectively identify high-risk venues, enabling timely prevention and control measures to prevent further spread.

Despite the high accuracy demonstrated by our model in validation, certain limitations exist. First, the model' s predictive performance depends heavily on data quality. Epidemiological investigation data may have incomplete information or subjective bias, and Ct value testing may produce measurement errors due to experimental condition differences; these factors may lead to prediction bias. Second, this study trained and validated using data from only 201 involved venues in one district of Beijing; thus, generalizability across different regions, populations, and multi-type respiratory infectious disease scenarios has not been fully verified. Model application requires external validation using multi-center, multi-period data to ensure robustness and universality. For example, venue types, population density, and ventilation conditions in rural areas may differ significantly from urban areas, and model predictive performance may decline in these scenarios. Additionally, with the emergence of new variants, transmissibility and transmission routes may change, potentially affecting existing model validity. Therefore, future research should introduce multi-region, multi-period data for external validation and explore dynamic model updates to enhance generalizability and robustness in practice.

Future research can further optimize model algorithms to improve adaptability and accuracy in different scenarios. Furthermore, exploration of incorporating

additional data types such as meteorological data and socioeconomic data into model training may improve predictive capability. Researchers should also focus on model applicability across different regions and cultural backgrounds to achieve broader application.

Conclusion

This study successfully compared the performance of multiple machine learning models in respiratory infectious disease transmission risk assessment. The established assessment model provides an effective tool for risk assessment of respiratory infectious disease epidemic venues, helping policymakers prioritize and make decisions under limited resources. Given potential similarities in transmission mechanisms among respiratory infectious diseases, this model is expected to provide reference for risk assessment of other respiratory infectious diseases.

Author Contributions: HU Jian was responsible for study design, investigation and experimentation, and manuscript writing. LIU Min was responsible for research proposition, implementation, and final version revision. JIAO Zengtao was responsible for study design, data cleaning, and investigation and experimentation. JING Wenzhan was responsible for study design, guidance provision, and manuscript review. LIANG Wannian was responsible for study design, guidance provision, and manuscript review and revision.

Conflict of Interest: The authors declare no conflict of interest.

(Received: 10 August 2025; Revised: 30 November 2025)

References

[1] WANG W J, HUANG J L, GUO S J. Construction of hospital operation decision support system based on big data [J]. China Digital Medicine, 2019, 14(2): 49-51.

[2] ZHAO S, MUSA S S, LIN Q Y, et al. Estimating the unreported number of novel coronavirus (2019-nCoV) cases in China in the first half of January 2020: a data-driven modelling analysis of the early outbreak [J]. J Clin Med, 2020, 9(2): 388. DOI: 10.3390/jcm9020388.

[3] BROWNSTEIN J S, RADER B, ASTLEY C M, et al. Advances in artificial intelligence for infectious-disease surveillance [J]. N Engl J Med, 2023, 388(17): 1597-1607. DOI: 10.1056/NEJMra2119215.

[4] BASTANI H, DRAKOPOULOS K, GUPTA V, et al. Efficient and targeted COVID-19 border testing via reinforcement learning [J]. Nature, 2021, 599(7883): 108-113. DOI: 10.1038/s41586-021-04014-z.

[5] WU J Y, XIE X W, YANG L, et al. Mobile health technology combats COVID-19 in China [J]. J Infect, 2021, 82(1): 159-198. DOI: 10.1016/j.jinf.2020.07.024.

[6] DONG J C, WU H Q, ZHOU D, et al. Application of big data and artificial intelligence in COVID-19 prevention, diagnosis, treatment and management decisions in China [J]. J Med Syst, 2021, 45(9): 84. DOI: 10.1007/s10916-021-01757-0.

[7] ZHANG X Y, ZHANG D, ZHANG X F, et al. Artificial intelligence applications in the diagnosis and treatment of bacterial infections [J]. Front Microbiol, 2024, 15: 1449844. DOI: 10.3389/fmicb.2024.1449844.

[8] ZHU Q, LIU J. A united model for diagnosing pulmonary tuberculosis with random forest and artificial neural network [J]. Front Genet, 2023, 14: 1094099. DOI: 10.3389/fgene.2023.1094099.

[9] LEE A R, LEE H, CHO Y, et al. Development of a machine learning model to diagnose pediatric lower respiratory tract infections [J]. Sci Rep, 2025, 15: 41710.

[10] YE Y, GAO Z, ZHANG Z, et al. A machine learning model for predicting severe mycoplasma pneumoniae pneumonia in school-aged children [J]. BMC Infect Dis, 2025, 25(1): 570.

[11] CHEN T Q, GUESTRIN C. XGBoost: a scalable tree boosting system [C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, CA, USA: ACM, 2016: 785-794.

[12] MA J Q. Impact of Internet development on infectious disease reporting system in China in the 21st century [J]. Chinese Journal of Public Health Management, 2003, 19(3): 245-247. DOI: 10.3969/j.issn.1001-9561.2003.03.044.

[13] BUITRAGO-GARCIA D, EGLI-GANY D, COUNOTTE M J, et al. Occurrence and transmission potential of asymptomatic and presymptomatic SARS-CoV-2 infections: a living systematic review and meta-analysis [J]. PLoS Med, 2020, 17(9): e1003346. DOI: 10.1371/journal.pmed.1003346.

[14] MEYEROWITZ E A, RICHTERMAN A, BOGOCH I I, et al. Towards an accurate and systematic characterisation of persistently asymptomatic infection with SARS-CoV-2 [J]. Lancet Infect Dis, 2021, 21(6): e163-e169. DOI: 10.1016/S1473-3099(20)30837-9.

[15] SALA E, SHAH I S, MANISSERO D, et al. Systematic review on the correlation between SARS-CoV-2 real-time PCR cycle threshold values and epidemiological trends [J]. Infect Dis Ther, 2023, 12(3): 749-775. DOI: 10.1007/s40121-023-00772-7.

Note: Figure translations are in progress. See original paper for figures.