Liu *et al. BMC Public Health* (2025) 25:975
https://doi.org/10.1186/s12889-025-21740-5

The symbiotic e□ect of online searches and vaccine administration—a nonlinear correlation analysis of baidu index and vaccine administration data

Yixin Liu[1], Lingshi Ran[1], Yang Wang[1] and Yixue Xia[1*]

Abstract

This study primarily addresses the analytical problem of the mathematical mechanism underlying the associative impact between online searches and vaccine uptake, a relationship that has become increasingly relevant in the context of public health management. As internet search behaviors reflect public interest and sentiment, understanding their impact on vaccination trends is crucial for real-time health decision-making. A Logistic model is constructed to observe the fundamental evolutionary patterns between online searches and vaccine uptake. To explore their mutual influence, an impact function is defined, and the common structural factors with the highest fitness are determined through data fitting. Subsequently, a dynamic detection model of the associative impact between online data and societal objects, based on the mathematical mechanism, is established. Using this model, dynamic predictions are conducted to verify its predictive capability at certain stages. Through research, a symbiotic e□ect between online searches and vaccine uptake is identified, revealing a nonlinear correlation between the two. The model demonstrates the ability to predict vaccine uptake trends based on online search data, with certain prediction windows showing high accuracy. This research not only clarifies the mathematical mechanism underlying this relationship but also demonstrates the advantage of integrated analysis and prediction. It provides a new method for predicting online searches and vaccine uptake, o□ering theoretical and empirical support for public health and social science research.

Keywords COVID-19 vaccination, Web search, Mathematical mechanism, Symbiotic e□ect, Logistic model

Introduction

In today's societal systems, the entities that are interconnected can be roughly divided into two categories: network data and social objects. Network data refers to all digital information generated online that can be quantitatively analyzed. Social objects, on the other hand, constitute the tangible, foundational, and highly crystallized portion of social reality o ine. ey are the fundamental constituents of society, forming an integrated whole with specific quantities, qualities, modes of combination, and degrees of consolidation. Social objects possess tangible power to instigate change.

With the construction of an interactional world of virtuality and reality, the interaction between network data and social entities has become more complex. Many studies have explored the correlation between the two, such as predicting trends in social entities using network data, uncovering influencing factors, and measuring the

*Correspondence:
Yixue Xia
rcofnpog@126.com
[1] Research Center of Network Public Opinion Governance, China People's Police University, Langfang 065000, China

degree of influence. Among these, the analysis of the mathematical mechanism behind their correlation is still worth exploring. Clarifying the mathematical mechanisms, especially constructing continuous models evolving over time, will increase the interpretability of the correlation, endowing it with a philosophical foundation such as social evolution mechanisms, and thus achieving a combination of data-driven and theory-driven appro4.1aches. Particularly in enhancing interpretability, correlation detection based on mathematical mechanisms o ers the advantage of integrated analysis and prediction.

In recent years, with the rapid development and widespread adoption of Internet technology, there has been a multitude of ways in which the public accesses information. Among these, internet search engines stand out as one of the primary channels through which the public

indicated that among 70 studies conducted between 2009 and 2013, 70% utilized time trend analysis (comparing across time periods), 11% conducted cross-sectional analysis (comparing across di erent locations within a single time period), and 19% of the studies employed both methods simultaneously [11]. ese three approaches have been applied in studies related to vaccination. For instance, Pullan and Dey observed the popularity of search terms related to COVID-19 vaccines in Google Trends, suggesting that internet searches can help monitor public attitudes towards vaccines during rapidly

for increasing vaccination rates. Additionally, in the internet era, people tend to prefer actively searching for information online and proactively getting vaccinated.

In general, this study has made new advances compared to existing research in several aspects: (1) Proposing a method for detecting nonlinear correlations between online data and o ine entities. (2) Building a new mathematical model based on the Logistic model to explicitly describe and quantify the interaction between internet searches and vaccine administration. (3) rough predictive research, this study provides a new method for forecasting the number of COVID-19 vaccine doses administered. is not only provides theoretical support for adjusting public health policies related to COVID-19 vaccine administration but also o ers valuable insights for the administration of other vaccines.

Data modeling environment

Network data reflects social entities online and serves as a "barometer" of societal activities. Without social entities, network data would cease to exist. Following every instance of online public opinion, there lies an attribution to a social entity issue. However, research on the regulatory mechanisms governing the correlation between network data and social entities within the social system is relatively scarce after sudden events occur. ere are two reasons for this: firstly, it is often challenging to quantify both network data and social entities simultaneously after a sudden event; secondly, the impact of some sudden events may be limited in time or space, rendering the quantified data less representative. So, how can we explore the regulatory mechanisms governing the correlation between network data and social entities within the social system?

From the perspective of sudden events, since the World Health Organization declared the outbreak of novel coronavirus pneumonia as an international public health emergency, the COVID-19 pandemic has persisted for nearly four years. Up to now, the novel coronavirus is still causing deaths and mutations. e COVID-19 pandemic, as a sudden public health emergency, has a long time span, a wide spatial scope, and is representative, making it suitable for studying the regulatory mechanisms governing the correlation between network data and social entities within the social system. erefore, in the face of the big data environment triggered by the COVID-19 pandemic, how can we select appropriate methods for quantifying network data and social entities?

Quantitative data

As the most severe global public health emergency in the past century, the COVID-19 pandemic has brought about significant changes in people's lifestyles and survival status both online and o ine. Online, social media has almost overnight become the primary means of communication and social interaction for people worldwide, with the quantity of public opinion information showing a high correlation with confirmed COVID-19 cases.

e real-time dynamics of the pandemic have inundated media platforms, with daily updates on confirmed cases, infected regions, trajectory tracking, and incidents of mask hoarding constantly entering the public eye. Concerns among the public have shifted from when the pandemic will end to whether there will be e ective vac-

against COVID-19, resulting in the approval of multiple COVID-19 vaccines for emergency use by December 2020. Despite ongoing global e orts, the pandemic persists, with various mutated viruses emerging, suggesting that humanity may have to coexist with the COVID-19 virus for an extended period. Vaccination, aimed at enhancing immunity, may become a long-term process. In response to the complex and severe impact of the COVID-19 pandemic, governments worldwide have prioritized the development and administration of COVID-19 vaccines as a key measure to curb the outbreak [28]. Given the continuous mutation of the virus, future vaccination e orts may be necessary to strengthen immunity against di erent strains, making vaccination an enduring process. erefore, the COVID-19 vaccine administration exhibits characteristics of long-term, widespread, and sustained impact, meeting the requirements for detecting the relational impact patterns in social entities.

In the era of the pandemic, COVID-19 vaccination has continuously transitioned towards public participation, consolidation, and coverage. Starting from March 23, 2021, the National Health Commission of China (http://www.nhc.gov.cn) began publishing the COVID-19 vaccination situation in China and updating the cumulative vaccination data daily (http://www.nhc.gov.cn/xcs/xxgzbd/gzbd_index.shtml). In view of this, in this study, we crawled daily COVID-19 vaccination data in China from March 23, 2021, to December 23, 2022, quantified as the number of vaccinations administered per day (unit: 10,000 doses), as the quantitative data for social entities.

Data correlation analysis

After obtaining the initial data, analysis reveals two corresponding relationships in terms of time and data volume between Baidu search index data (SI) and China's COVID-19 vaccine inoculation data (VI). Firstly, there is a variability relationship where each data point in the BI corresponds to the search index quantity for each day.

is variability relationship slices time and analyzes the data generated at each time point, corresponding to an instantaneous relationship. Secondly, there is an evolutionary relationship in the VI, where each data point corresponds to the cumulative number of vaccine inoculations up to that day. e evolutionary relationship accumulates time and analyzes all generated data, corresponding to an accumulated relationship. When exploring the rules of the association between network data and social entities in the social system, it is necessary to choose whether to use instantaneous data or cumulative data. Firstly, the data format needs to be standardized by segmenting VI into daily instantaneous data to obtain an instantaneous dataset; and SI is accumulated daily to obtain a cumulative dataset.

To investigate the rationality of data selection, we conducted correlation analysis separately on instantaneous data and cumulative data to determine the correlation between BI and VI. We identified the dataset with the highest correlation coe cient to be used for subsequent modeling work.

Correlation analysis is divided into two types: global correlation analysis and dynamic correlation analysis.

e global correlation coe cient is used to measure the overall association between BI and VI over the entire study period from March 24, 2021, to December 23, 2022, comprising 640 data points. We calculate the global correlation coe cient using the following formula:

$$_{global} = \mathrm{cor}(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sigma X \sigma Y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2}\sqrt{\sum (y_i - \bar{y})^2}}$$

Where r_{global} represents the global correlation coe cient and , (, Y) represents the covariance between BI X and VI Y. σ X and σ Y respectively denote the standard deviations of the search index X and COVID-19 vaccination data Y. e linear correlation coe cient eliminates the influence of dimensions. Its values range from −1 to 1, where a value closer to 1 indicates a stronger linear correlation, while a value of 0 indicates no linear correlation.

e calculation results are as follows:

$$r_{global1} = 0.782171172$$

$$r_{global2} = 0.992609001$$

Where $r_{global1}$ represents the global correlation coe cient between instantaneous data, and $r_{global2}$ represents the global correlation coe cient between cumulative data. $0.5 < r_{global} < 1$, indicating a strong positive correlation. It is evident that both X and Y exhibit strong correlation in both instantaneous and cumulative data. $r_{global2} > r_{global1}$, indicating that the correlation in cumulative data is stronger compared to instantaneous data.

e dynamic correlation coe cient is used to describe the dynamic changes in the correlation between X and Y from the beginning to the t-th day. By calculating the evolution correlation coe cient $_{in}( )$, we can observe the dynamic changes in the correlation between X and Y during the evolutionary process. Its calculation formula is as follows:

Where $_{in}( )$ represents the dynamic correlation coe cient for the t-th day, $\mathrm{cov}(X_{1:t}, Y_{1:t})$ denotes the

covariance between X and Y from the beginning of the event to the t-th day, and $\sigma X_{1:t}$ and $\sigma Y_{1:t}$ respectively represent the standard deviations of X and Y during that time period. Since the Pearson correlation coe cient requires at least two data points to analyze the correlation, in the analysis of dynamic correlation coe cients, a total of 639 Pearson correlation coe cients were obtained. e dynamic trend of the correlation coe - cient over time is shown in Fig. 1.

It can be observed that the dynamic correlation coefficient $_{n1}$ of instantaneous data fluctuates more significantly compared to the average dynamic correlation coe cient $avg(r_{evolution1}(t)) = 0.703846124$. e dynamic correlation coe cient $r_{evolution2}$ of cumulative data exhibits smaller fluctuations, with the lowest value $r_{evolution2}(53)$ observed at 53 days, where $r_{evolution2}(53) = 0.975521704$. e average dynamic correlation coe cient $avg(r_{evolution2}(t)) = 0.99172064$ indicates that the overall dynamic correlation coe cient of cumulative data remains at a very high level.

From the results of the overall and dynamic correlation coe cients, it can be observed that the cumulative data of BI and VI exhibit higher correlation compared to the instantaneous data, indicating greater research value and significance. Furthermore, this also confirms the rationality of our data quantification and fundamental assumptions regarding network data and social entities, providing support for our subsequent establishment of basic models. In terms of the meaning of the data itself, cumulative data considers historical factors and past influences, while instantaneous data focuses more on the current state. We need to analyze historical evolutionary patterns and capture trends in data evolution over time.

erefore, we choose cumulative data as the final modeling data, with a time range from March 24, 2021, to December 23, 2022, totaling 640 data points.

Model establishment

Hypothesis

e evolution pattern of information data generated by sudden events has been extensively studied, although the models applied and the delineation of stages may vary. Nevertheless, fundamentally, they all acknowledge that information data evolve along an “S-curve”. In 1985, the renowned American information resource management expert, Horton, introduced the concept of information lifecycle, suggesting that information resources follow natural laws of motion and possess their own lifecycle. According to the information lifecycle theory, the evolution process of information data can be divided into stages. Currently, the academic community has di erent views on the division and naming of stages in the evolu-

s di xn and na(w)-5((ss t)6(heir(o anal[(a)-r)6(c)8.1(t)-6(age)-7p)-13(

online and o ine entities conform to an independently existing logistic model.

Specifically, it is assumed that the BI $x_1(t)$, corresponding to online data, is a monotonically increasing function of time t with an initial value of $x_1(0)$. e number of VI $x_2(t)$, corresponding to social entities, is also a monotonically increasing function of time t with an initial value of $x_2(0)$. Based on the above analysis, the foundational model for obtaining quantified data of online and o ine entities (model 1) is constructed as follows:

Where $r_1 > 0$ represents the inherent growth rate of the cumulative BI, and $r_2 > 0$ represents the inherent growth rate of cumulative VI. Due to the existence of a lifecycle, both the cumulative BI and cumulative VI have

a better fit. In terms of significance test, the P-values of parameters r_{12}, $-\frac{r_{12}}{K_{12}}$, and a_2 are smaller compared to the P-values of parameters r_{11}, $-^{r}$

social entities, resulting in an increase in COVID-19 vaccination volume. e degree to which a unit of BI promotes the VI $_2$ is represented by $\beta \frac{x_2}{K_2}$, while $\frac{x_1}{K_1}$ represents the degree to which $\beta \frac{x_1}{K_1}^{x}$

According to the theory: e local equilibrium point is stable when the determinant $DET(J) > 0$ and the trace $Tr(J) < 0$ of the Jacobian matrix. We could screen out the equilibrium points that meet the conditions by substituting each equilibrium point into the Jacobian matrix, as shown in Table 11.

According to the definition in this text, equilibrium points need to be greater than zero. P1 is not the equilibrium point sought in this paper. Only the case where P2 holds is discussed in this paper.

According to the model, the equilibrium point for the BI changes from K_1 to κ_1

capabilities can support intervention design by o ering insights into potential future vaccine uptake, enabling policymakers toT q /(sig)-.bling

Health Commission of China, covering the entire period of the evolution of China's COVID-19 vaccination campaign, ensuring the representativeness of the analysis. However, challenges arise when modeling long-term behavioral shifts, especially as public attention towards COVID-19 and vaccination has diminished over time. As public concern wanes, the dynamics of search behaviors may change, which could a ect the model's predictive accuracy in future scenarios. However, our team has noticed this situation and has already initiated further research [36].

(3)

Acknowledgements
Nothing to declare.

Disclaimer
The findings and conclusions of this report are those of the authors and do not necessarily represent the o□cial position of the Centers for Disease Control and Prevention.

Authors' contributions
YL: conceptualization, validation, investigation, formal analysis, and writing—original draft. LR: writing—editing and data curation. YW: data curation, investigation, and writing—editing. YX: conceptualization, funding acquisition, supervision, methodology, and writing—review. All authors contributed to the article and approved the submitted version.

Funding
This work was supported by the Ministry of Public Security Technology Research Program under Grant No. 2023JSYJC20.

Data availability
The dataset analyzed in this study is available from the corresponding author upon reasonable request.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 29 April 2024 Accepted: 31 January 2025
Published online: 12 March 2025

References

1. Shekhar R, Sheikh AB, Upadhyay S, Singh M, Kottewar S, Mir H, et al. COVID-19 vaccine acceptance among health care workers in the United States. Vaccines. 2021;9:119.
2. Marian AJ. Current state of vaccine development and targeted therapies for COVID-19: impact of basic science discoveries. Cardiovasc Pathol. 2021;50:107278.
3. Di Domenico G, Nunan D, Pitardi V. Marketplaces of misinformation: a study of how vaccine misinformation is legitimized on social media. J Public Policy Mark. 2022;41:319–35.
4. Nazli SB, Yigman F, Sevindik M, Ozturan DD. Psychological factors a□ecting COVID-19 vaccine hesitancy. Ir J Med Sci. 2022;191:71–80.
5.

International on Conference on Information and Knowledge Management. New York: Association for Computing Machinery; 2016. p. 1953–6.