CHINA CDC WEEKLY

Vol. 2 No. 52 Dec. 25, 2020

Preplanned Studies

Estimating the Incidence of Tuberculosis — Shanghai, China, 2025−2050 995

Predicting the Incident Cases of Emerging Infectious Disease Using a Bayesian Probability Model — China, February 2020 999

Methods and Applications

Zero-Shot Medical Image Retrieval for Emerging Infectious Diseases Based on Meta-Transfer Learning — Worldwide, 2020 1004

Commentary

Internet Data for Improving Prevention and Control of Global Infectious Diseases 1009

Dr. Yi Zeng (1929−2020) A Pioneer on HIV and Tumor Virology in China 1013

Jibin Tan (1972–2020) Director of China CDC Center for Logistics Operation and Management 1014
This week’s issue was organized by Guest Editor Zhongwei Jia.
Estimating the Incidence of Tuberculosis — Shanghai, China, 2025–2050

Shegen Yu; Jiaqi Ma; Zhongwei Jia

Summary

What is already known on this topic?

Despite the impressive achievements in eliminating tuberculosis (TB), the TB burden is still heavy in China. By 2010, China halved the prevalence and mortality reported in 1990, but China is still one of 30 high-TB burden countries in the world.

What is added by this report?

A dynamic transmission model including both rifampin resistant TB (RR-TB) and relapse of pulmonary TB was created. The TB incidence of Shanghai in 2025 and 2035 was predicted, and sensitively analysis of reducing transmission, treating latent TB infection (LTBI), and reducing the recurrence rate was conducted.

What are the implications for public health practice?

Screening for latent TB infections should be carried out regularly in high-risk groups and areas using tuberculin skin testing and/or interferon gamma release assays.

Tuberculosis (TB) is a global public health problem. The World Health Organization (WHO) proposed that the incidence rate of TB should be reduced to less than 55/100,000 population by 2025, less than 10/100,000 by 2035, and to eliminate TB by 2050 (incidence rate <1/100,000) (1). Based on the directly observed treatment, short-course (DOTS) strategy nationwide, China halved the prevalence and mortality of TB in 2010 as compared to 1990 (2), and the cure rate of TB has been reached 92.9% in 2013 (3). The TB incidence rate fell 43.1% from 1990 (130/100,000) to 2010 (74/100,000) (4). Despite impressive achievements, China is still one of the 30 high-TB burden countries in the world. In 2019, there were about 833,000 new TB cases in China with a TB incidence rate of 58/100,000 (4). China also has the highest latent TB infection (LTBI) burden globally with approximately 350 million infections that are at risk for active TB disease (5).

Shanghai is one of the areas with the best implementation of TB control measures in China, but the incidence rate was still above 25/100,000 in 2019 (6). Shanghai failing to reach the target by 2035 would indicate a high likelihood of failure for other areas in China. We established a dynamic TB model to estimate the predicted incidence in Shanghai and the impact of different prevention and control measures.

According to the natural progressive history of pulmonary TB, the overall population was divided into 7 classes: S referred to people who are not infected with Mycobacterium tuberculosis (M. tb); E referred to people infected with M. tb but were not yet infectious; I referred to patients with infectious pulmonary TB; I\[1\]\[1\] referred to the undetected and drug sensitive population; I\[1\]\[2\] referred to the undetected and rifampicin resistant population; I\[2\]\[1\] referred to the detected and drug sensitive population; I\[2\]\[2\] refers to the detected and rifampicin resistant population; and R referred to TB patients who have been successfully treated. In this model, we made the following assumptions: 1) the population was evenly mixed, and contact between all individuals was equally likely; 2) patients were likely to infect susceptible population and the recovered after contact with them; 3) all detected pulmonary TB cases were reported to National Notifiable Disease Reporting System; and 4) the total population of the system was relatively stable. The population supplement was due to births in each year, and population loss was due to natural deaths from each group and deaths due to pulmonary TB from patients.

The equations of the model are as follows:

$$\begin{align*}
\frac{dS}{dt} &= \lambda - \beta S(I_{11} + I_{12})S - \delta S - \rho_0 S \\
\frac{dE}{dt} &= \beta S(I_{11} + I_{12})S + \beta I_{11}S - \alpha E - \rho E \\
\frac{dI_{11}}{dt} &= \alpha (1-\gamma_1)(1-\sigma)E + \omega (1-\gamma_2)(1-\sigma)R - \rho I_{11} - \sigma I_{11} - (\rho_0 + \mu_1) I_{11} \\
\frac{dI_{12}}{dt} &= \kappa \gamma_1 I_{11}E + \omega \gamma_1 I_{11}R + \rho I_{11} - \alpha I_{12} - (\rho_0 + \mu_2) I_{12} \\
\frac{dI_{21}}{dt} &= \kappa \gamma_2 I_{12}E + \omega \gamma_2 I_{12}R + \rho I_{12} - \alpha I_{21} - (\rho_0 + \mu_1) I_{21} \\
\frac{dI_{22}}{dt} &= \kappa \gamma_2 I_{12}E + \omega \gamma_2 I_{12}R + \rho I_{12} - \alpha I_{22} - (\rho_0 + \mu_2) I_{22} \\
\frac{dR}{dt} &= \gamma I_{21} + \gamma I_{22} - \omega R - \rho R
\end{align*}$$
The model involves 7 classes and 14 parameters. Each equation represents the change rate of the number of people in each class in unit time, and the right side includes the moving in and out of items that lead to the change of class population. The unit time of this model is one year.

Λ is the constant recruitment in the system. β₁ and β₂ are the transmission rates of infectious drug-sensitive TB cases and rifampin resistant TB cases (RR-TB). κ is the progressive rate from the exposed to the infectious; ρ is the progressive rate from drug-sensitive TB to RR-TB; σ is the detection rate of the infectious; γ₁ and γ₂ are the successful treatment rates of detected patients with infectious drug-sensitive TB and RR-TB, respectively; and μ₀ is the natural mortality rate, while μ₁ and μ₂ are the fatality rates of TB in infectious drug-sensitive TB cases and RR-TB cases, respectively. The transmission diagram is shown in Figure 1. We collected the reported incidence of pulmonary TB in Shanghai from 2004 to 2017 provided by the Public Health Science Data Center, of which 2004–2012 were used as training data and 2013–2017 years were used as test data. The values of the parameters were determined by the reports of earlier studies and adjusted according to TB data, then the incidence of pulmonary TB in Shanghai was estimated for the near future. The incidence of pulmonary TB was numerically defined as the number of new reported cases of pulmonary TB within each year as a proportion of the number of average annual population.

The parameters in the model were adjusted to simulate the effect of three different TB prevention and control strategies. We reduced the values of parameters β₁, β₂, and ω to simulate reducing the probability of infection or reinfection of susceptible and recovered patients (assuming 60% of recurrent patients are due to reinfection) to simulate strengthening personal protection and isolation of active cases during contagious period. The effect of preventive treatment on LTBI cases was evaluated by reducing the rate of progression (κ) of the exposed group to the infectious groups. We reduced the recurrence rate (ω) of the recovered group to study the

FIGURE 1. The flow diagram of tuberculosis model considering rifampicin resistance and recurrence. S referred to people who are not infected with Mycobacterium tuberculosis (M. tb); E referred to people infected with M. tb but were not yet infectious; I₁ referred to patients with infectious pulmonary tuberculosis (TB); I₁, referred to the undetected and drug sensitive population; I₂, referred to the detected and rifampicin resistant population; and R referred to TB patients who have been successfully treated. β₁ and β₂ are the transmission rates of infectious drug-sensitive TB cases and rifampin resistant TB (RR-TB) cases. κ is the progressive rate from the exposed to the infectious; ρ is the progressive rate from drug-sensitive TB to RR-TB; σ is the detection rate of the infectious; γ₁ and γ₂ are the successful treatment rates of detected patients with infectious drug-sensitive TB and RR-TB, respectively; ω is the disease recurrence rate from the recovered population; τ₁ and τ₂ are the drug resistance rates of new patients and recurrent patients, respectively; and μ₀ is the natural mortality rate, while μ₁ and μ₂ are the fatality rates of TB in infectious drug-sensitive TB cases and RR-TB cases, respectively.
impact of recurrence rate on the TB epidemic.

We set the initial values of the model classes as $S(0) = 14,453,131$, $E(0) = 3,834,319$, $I_{U1}(0) = 6,462$, $I_{U2}(0) = 333$, $I_{F1}(0) = 7,011$, $I_{F2}(0) = 361$, $R(0) = 48,194$, and the values of parameters are shown in Table 1. The first curve of each panel in Figure 2 shows our prediction of the incidence of pulmonary TB in Shanghai under current strategies. We predicted that the estimated incidence of pulmonary TB in Shanghai will continue to decline from 2004 to 2050. In 2025, the incidence of TB in Shanghai was estimated to be 24.27/100,000, which will achieve the WHO’s goal in 2025 (<55/100,000). However, the incidence was estimated to be 20.81/100,000 in 2035, still far from the goal set for 2035 (<10/100,000).

Figure 2 shows the impact of 3 different prevention and control strategies on pulmonary TB in Shanghai. The incidence will decrease slightly with the values of parameters $\beta_1$, $\beta_2$ and $\omega$ reduced (Figure 2A). The incidence of pulmonary TB in Shanghai in 2035 will be 19.69/100,000 when the parameters dropped by 50%. Reducing the progressing rate ($\kappa$) of the exposed group to the infectious groups, the incidence of pulmonary TB in Shanghai will decrease significantly (Figure 2B). In 2035, the incidence will be 1.15/100,000 with the parameters $\kappa$ dropped by 50%. The incidence of pulmonary TB in Shanghai will be slightly decreased by reducing their recurrence rate ($\omega$) (Figure 2C). With the recurrence rate reduced by 50%, the incidence of pulmonary TB in Shanghai will be 19.08/100,000 in 2035.

**DISCUSSION**

In this study, a dynamic transmission model concerning both RR-TB and a relapse of TB was established, and the RR-TB rate of recurrent patients was distinguished from that of new patients. Current prevention and control strategy for TB in Shanghai was estimated to be able to achieve the goal set forth by the WHO in the End TB Strategy in 2025 but were not sufficient to achieve the goal in 2035. The target was estimated to be unachievable due to many reasons including the large number of latent infections (7), coinfections with HIV / AIDS (8), and a large migrant population (9).

Among the three prevention and control strategies, strengthening preventive treatment for LTBI cases had the best effect on TB epidemic control. The incidence of pulmonary TB in Shanghai was estimated to decrease to 11.55/100,000 when the progressing rate dropped by 50%, which was close to the goal for 2035. The other two strategies were estimated to only reduce the incidence of pulmonary TB slightly. This indicated that the large number of LTBI cases was the reason why the incidence of TB was not decreasing as fast as expected. The WHO estimated that the global LTBI population was close to 2 billion, accounting for 1/3 of the global population (4).

Carrying out preventive treatment for latent infections is based on strengthening screening for latent infections. China has previously shown the ability to conduct large-scale screenings for an infectious disease in a large city (10), which indicates

---

**TABLE 1. Definitions and estimated values of parameters.**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Definition</th>
<th>Estimated value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Lambda$</td>
<td>Constant recruitment of the population</td>
<td>486,245</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>Transmission rate of infectious drug-sensitive TB cases</td>
<td>$8.69 \times 10^{-12}$</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>Transmission rate of infectious RR-TB cases</td>
<td>$2.52 \times 10^{-10}$</td>
</tr>
<tr>
<td>$\kappa$</td>
<td>Progressive rate from the exposed to the infectious</td>
<td>$1.65 \times 10^{-3}$</td>
</tr>
<tr>
<td>$\tau_1$</td>
<td>Drug resistance rate of new patients</td>
<td>$2.39 \times 10^{-2}$</td>
</tr>
<tr>
<td>$\sigma$</td>
<td>Detection rate of the infectious</td>
<td>$6.20 \times 10^{-1}$</td>
</tr>
<tr>
<td>$\rho$</td>
<td>Progressive rate from drug-sensitive TB to RR-TB</td>
<td>$8.85 \times 10^{-2}$</td>
</tr>
<tr>
<td>$\gamma_1$</td>
<td>Treatment successful rates of infectious drug-sensitive cases</td>
<td>$9.14 \times 10^{-1}$</td>
</tr>
<tr>
<td>$\gamma_2$</td>
<td>Treatment successful rates of infectious RR-TB cases</td>
<td>$8.99 \times 10^{-1}$</td>
</tr>
<tr>
<td>$\omega$</td>
<td>Recurrence rate from recovered</td>
<td>$5.94 \times 10^{-3}$</td>
</tr>
<tr>
<td>$\mu_0$</td>
<td>Drug resistance rate of recurrent patients</td>
<td>$1.25 \times 10^{-1}$</td>
</tr>
<tr>
<td>$\mu_1$</td>
<td>Natural mortality rate</td>
<td>$7.46 \times 10^{-3}$</td>
</tr>
<tr>
<td>$\mu_2$</td>
<td>Fatality rates of TB in infectious drug-sensitive TB cases</td>
<td>$9.32 \times 10^{-3}$</td>
</tr>
<tr>
<td>$\omega_1$</td>
<td>Fatality rates of TB in infectious RR-TB cases</td>
<td>$2.26 \times 10^{-2}$</td>
</tr>
</tbody>
</table>

Abbreviations: TB=tuberculosis; RR-TB=rifampin resistant TB.
that it is possible to strengthen screening. Considering the large scope of consumption in the process of the project, screening for latent TB infections using tuberculin skin testing and/or interferon gamma release assays can be carried out regularly in high-risk groups and areas.

This study was subject to some limitations. For example, the parameters used in model operation and prediction were fixed values, but the parameters in the model change dynamically with time in real life. The fixed parameter value could only predict long-term trends for TB but not short-term fluctuations. In addition, considering the difficulty of obtaining the parameters, the model only set 7 classes without further subdivisions, so there was still a gap with real circumstances of TB in Shanghai. Further studies could consider dividing the population of recovered individuals into subgroups with different recurrence rates according to their recurrence risk.

**Funding:** Key Joint Project for Data Center of the National Natural Science Foundation of China and Guangdong Provincial Government (U1611264); and the National Natural Science Foundation of China (91546203, 91846302).

doi: 10.46234/ccdcw2020.266

* Corresponding author: Zhongwei Jia, uchiinj@163.com.

1 School of Public Health, Peking University, Beijing, China; 2 Chinese Center for Disease Control and Prevention, Beijing, China; 3 Center for Intelligent Public Health, Institute for Artificial Intelligence, Peking University, Beijing, China; 4 Center for Drug Abuse Control and Prevention, National Institute of Health Data Science, Peking University, Beijing, China.

Submitted: December 03, 2020; Accepted: December 23, 2020

**REFERENCES**

**Summary**

**What is already known about this topic?**
The exact number of incident cases of emerging infectious diseases on a daily basis is of great importance to the disease control and prevention, but it is not directly available from the current surveillance system in time.

**What is added by this report?**
In this study, a Bayesian statistical method was proposed to estimate the posterior parameters of the gamma probability distribution of the lag time between the onset date and the reporting time based on the surveillance data. And then the posterior parameters and corresponding cumulative gamma probability distribution were used to predict the actual number of new incident cases and the number of unreported cases per day. The proposed method was used for predicting COVID-19 incident cases from February 5 to February 26, 2020. The final results show that Bayesian probability model predictions based on data reported by February 28, 2020 are very close to those actually reported a month later.

**What are the implications for public health practice?**
This research provides a Bayesian statistical approach for early estimation of the actual number of cases of incidence based on surveillance data, which is of great value in the prevention and control practice of epidemics.

On January 20, 2020, the Chinese State Council added the latest coronavirus disease (COVID-19) to the Category B list of nationally notifiable diseases under Category A management (1–2). This means that if a case is diagnosed with COVID-19, it must be reported by the physician to the National Notifiable Disease Reporting System (NNDRS) within two hours. However, new cases reported every day from the surveillance system often contain cases that had onset on or before the reporting date, indicating a lag between onset and diagnosis. A more accurate assessment of incidence will allow public health professionals to better assess ongoing outbreaks, the pattern and scale of further epidemics, and the effectiveness of current prevention and control strategies, etc. (3). However, in the case of an emerging infectious diseases, such as COVID-19, the precise distribution of lag time between the dates of onset and the reporting times at the early stage of transmission was not known due to lack of historical data. Furthermore, all statistical incidence counts of each day are censored or truncated, i.e. up to the last reporting date, making it more difficult to estimate the precise distribution of delayed onset-reporting times.

In this study, a Bayesian statistical method was proposed to estimate the exact probability distribution of the lag time between the onset date and the reporting time, and then to predict the actual number of new incident cases and the number of unreported cases per day.

All data for this study were obtained from NNDRS. The dataset for all suspected and confirmed cases of COVID-19 was downloaded from NNDRS around 24:00 on March 26, 2020, and the difference between date of onset and report time was used to calculate the lag time for each case. The lag time was assumed to follow the same probability distribution over a certain timeframe if the case diagnostic criteria, diagnostic methods, and other factors related to case reporting remained relatively stable. For this reason, new cases with onset dates between February 5 and February 28 were chosen for this analysis, and a training dataset (data reported as February 28 at 24:00) and a validation dataset (data reported as March 25 at 24:00) were built. Based on the distributions of time delays for other infectious diseases, particularly influenza cases in the last few years and the empirical distribution of time delays for all COVID-19 cases in NNDRS, the gamma probability distribution was selected to be validated with high priority in this study.
At first, the lag times were transformed using the base-2 logarithm. A random variable $X$ that is gamma-distributed with parameters $\alpha$ and $\beta$ is denoted as follows:

$$X \sim \Gamma(\alpha, \beta) \equiv \text{Gamma}(\alpha, \beta).$$  \hfill (1)

Where $\alpha$ is a shape parameter and $\beta$ is a scale parameter, also called a rate parameter. Its corresponding probability density function is as follows:

$$f(x; \alpha, \beta) = \frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)} \quad \text{for } x > 0, \alpha, \beta > 0 \quad (2)$$

In our study, the logarithms of lag times were assumed to follow the truncated gamma distribution.

$$\log_2(\text{lagtime}[i]) \sim \Gamma(\alpha, \beta) \, T(0, \text{index}[i]) \equiv \text{Gamma}(\alpha, \beta) \, T(0, \text{index}[i]) \quad (3)$$

$$\alpha \sim \text{uniform}(9.0, 15) \quad (4)$$

$$\beta \sim \text{uniform}(3.5, 6.5) \quad (5)$$

Where the $\alpha$ and $\beta$ are the parameters to be estimated for the gamma distribution, the prior distributions of both parameters were set to follow the uniform distributions, lagtime[i] is the logarithm of the lag time of the i-th case, $T$ stands for the truncated distribution, and index[i] is the logarithm of the time interval from the start date of the i-th case to the end of the study time frame (24:00 on February 28) of the training dataset.

The cumulative reported rate was calculated using the posterior parameters of gamma distribution estimated from the above-mentioned steps and the gamma cumulative distribution as follows.

$$\text{rate}[n] = F(n; \alpha, \beta) = \int_0^n f(u; \alpha, \beta)du \quad (6)$$

Where $F$ and $f$ are the gamma cumulative probability function and gamma probability density function, respectively. The number of unreported incident cases was assumed to follow the negative binomial distribution. The number of cases of incidence for each day were estimated as follows.

$$\text{NR}[n] \sim \text{dnegbin} \left(\text{rate}[n], \text{reported}[n]\right) \quad (7)$$

$$\text{NN}[n] = \text{NR}[n] + \text{reported}[n] \quad (8)$$

Where $n$ is the number of days from date of onset to the 24:00 on February 28, the dnegbin is the negative binomial distribution, the $\text{rate}[n]$ is the cumulative probability of reporting of $n$ days, the $\text{reported}[n]$ is the numbers of reported cases of $n$ days calculated from the training data samples, and the $\text{NR}[n]$ is the number of unreported cases, and the $\text{NN}[n]$ is the predicted total number of cases. Since the number of new cases reported within two days was almost zero, we estimated the real incident cases from February 5 to February 26, 2020.

The validation data included all the cases with onset dates between February 5 and February 28, which were reported until 24:00 on March 26. The logarithm of the lag times in the validation dataset were fitted with gamma and other probability distribution models. The parameters of gamma probability distribution were estimated in both the training dataset using Bayesian Markov chain Monte Carlo (MCMC) method with truncated distribution and in the validation dataset using maximum likelihood method, respectively. The parameters of gamma probability distribution from training were used to estimate unreported and total number of incident cases every day. The estimated number of unreported and total incident cases were compared with the actual reported ones in the validation dataset. All statistical analyses were performed in R statistical software (version 4.0.3; The R Foundation for Statistical Computing, Vienna, Austria) (4), where Bayesian statistics were performed using the rjags (5) and runjags (6) packages, and general probability distribution fitting was performed using the fitdistrplus (7) package.

As of 24:00 on March 26, there were 24,551 cases in the validation dataset with onset dates between February 5 and February 28. The median lag time was 4.92 days, with 25 and 75 percentile values of 3.41 and 8.43 days, respectively. Gamma distribution was found to be the best fitted model compared to other probabilistic distribution models based on either the Akaike information criterion (AIC) or the Bayesian information criterion (BIC) values. The shape and rate parameters of the gamma distribution of the validation dataset were 8.56 and 3.54, respectively (Table 1).

The gamma probability distribution curves obtained for the two different methods as mentioned above are shown in Figure 1.

For the number of cases with an onset date between February 5 and February 26, the model predicted that there would be 2,112 unreported cases, while the actual reporting resulted in 1,665 newly reported cases as of March 26. The number of unreported cases predicted by the Bayesian model was 26.84% higher than the actual number of reported cases, with a ratio of 1:0.7881.

The model’s prediction of the total number of incident cases per day and its trend from February 5 to February 26 were generally consistent with the actual incidence data reported as of March 26. However, the
model’s predictions after February 21 were significantly higher than the actual data reported as shown in Figure 2 and Table 2.

**TABLE 1. The results of probability distribution fittings of the logarithm of the lagtimes based on validation dataset.**

<table>
<thead>
<tr>
<th>Distribution</th>
<th>LogLikelihood</th>
<th>Akaike information criterion (AIC)</th>
<th>Bayesian information criterion (BIC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gamma</td>
<td>–29,206.3</td>
<td>58,416.6</td>
<td>58,432.8</td>
</tr>
<tr>
<td>Burr</td>
<td>–29,705.1</td>
<td>59,416.3</td>
<td>59,440.6</td>
</tr>
<tr>
<td>Weibull</td>
<td>–29,757.8</td>
<td>59,519.6</td>
<td>59,535.8</td>
</tr>
<tr>
<td>Normal</td>
<td>–30,102.4</td>
<td>60,208.7</td>
<td>60,224.9</td>
</tr>
<tr>
<td>Pareto</td>
<td>–46,263.8</td>
<td>92,531.6</td>
<td>92,547.8</td>
</tr>
</tbody>
</table>

**DISCUSSION**

We attempted to estimate the actual number of incident cases during an outbreak of a new infectious disease. We assumed that the probability distribution of the lag time between the start date and the reporting time was relatively stable. Based on data from reported cases, the gamma probability distribution, and the Bayesian statistical method for truncated data, the parameters for the gamma probability distribution of lag times were inferred. Using the cumulative reporting rates calculated from the gamma probability distribution of lag time, the number of reported cases, the negative binomial distribution, the numbers of both unreported and total incident cases, and their 95% CI, were predicted. The results showed that the Bayesian probability model predictions based on data reported by February 28, 2020 were similar to those actually reported 1 month later.

The parameters of distributions for the lag reporting time were obtained using two different methods and datasets. The first was inferred from the truncated training dataset using the Bayesian MCMC-based parameter estimation method, and the second was estimated from the actual data reported 1 month later using the maximum likelihood estimation (MLE) method. It was found that both the parameters and the curve patterns of the two distribution models were consistent with the Kullback–Leibler divergence 0.048.

As for the number of unreported cases, the Bayesian model prediction was 26.84% higher than the actual reported number, but the absolute difference was only 1.82% of the total number of cases reported, i.e. \((2,112–1,665) / 24,551 \times 100\%\). The total number of incident cases reported by March 26 were within the 95% CI of total number of incident cases predicted by the model.

The number of actual daily reported cases was highly congruous with the predicted number of cases and associated trends between February 5 and 20. However, the model predicted an increase from February 20 to 23, and subsequent reports verified an anomalous increase in the number of cases on February 21 that could have changed the direction and magnitude of the model’s predictions and partly explain why the model’s predictions were higher than the actual reported number. In addition, the model forecast trend for the period of February 23–26 was consistent with the actual report, although the predictions were higher than the subsequent report.

Limitations in the application of this model may
TABLE 2. Comparison of the predicted number of COVID-19 incident cases by the Bayesian probability model and the number actually reported.

<table>
<thead>
<tr>
<th>Date of onset</th>
<th>Number of incident cases reported by February 28</th>
<th>Number of unreported cases and 95% CI predicted by model</th>
<th>Total number of incident cases and the 95% CI predicted by model</th>
<th>Number of incident cases reported by March 26</th>
<th>Cases reported from February 29 to March 26</th>
</tr>
</thead>
<tbody>
<tr>
<td>2020/2/5</td>
<td>2,529</td>
<td>24(15–34)</td>
<td>2,553(2,544–2,563)</td>
<td>2,559</td>
<td>30</td>
</tr>
<tr>
<td>2020/2/6</td>
<td>2,072</td>
<td>22(13–32)</td>
<td>2,094(2,085–2,104)</td>
<td>2,087</td>
<td>15</td>
</tr>
<tr>
<td>2020/2/7</td>
<td>2,007</td>
<td>25(16–36)</td>
<td>2,032(2,023–2,043)</td>
<td>2,036</td>
<td>29</td>
</tr>
<tr>
<td>2020/2/8</td>
<td>1,814</td>
<td>26(17–37)</td>
<td>1,840(1,831–1,851)</td>
<td>1,833</td>
<td>19</td>
</tr>
<tr>
<td>2020/2/9</td>
<td>1,438</td>
<td>24(15–35)</td>
<td>1,462(1,453–1,473)</td>
<td>1,453</td>
<td>15</td>
</tr>
<tr>
<td>2020/2/10</td>
<td>1,830</td>
<td>36(25–49)</td>
<td>1,866(1,855–1,879)</td>
<td>1,875</td>
<td>45</td>
</tr>
<tr>
<td>2020/2/11</td>
<td>1,343</td>
<td>31(21–43)</td>
<td>1,374(1,364–1,386)</td>
<td>1,366</td>
<td>23</td>
</tr>
<tr>
<td>2020/2/12</td>
<td>1,465</td>
<td>41(28–54)</td>
<td>1,506(1,493–1,519)</td>
<td>1,495</td>
<td>30</td>
</tr>
<tr>
<td>2020/2/13</td>
<td>1,265</td>
<td>42(30–56)</td>
<td>1,307(1,295–1,321)</td>
<td>1,313</td>
<td>48</td>
</tr>
<tr>
<td>2020/2/14</td>
<td>1,027</td>
<td>42(30–55)</td>
<td>1,069(1,057–1,082)</td>
<td>1,111</td>
<td>84</td>
</tr>
<tr>
<td>2020/2/15</td>
<td>816</td>
<td>41(29–54)</td>
<td>857(845–870)</td>
<td>868</td>
<td>52</td>
</tr>
<tr>
<td>2020/2/16</td>
<td>739</td>
<td>46(33–61)</td>
<td>785(772–800)</td>
<td>787</td>
<td>48</td>
</tr>
<tr>
<td>2020/2/17</td>
<td>642</td>
<td>50(37–65)</td>
<td>692(679–707)</td>
<td>893</td>
<td>251</td>
</tr>
<tr>
<td>2020/2/18</td>
<td>813</td>
<td>81(64–101)</td>
<td>894(877–914)</td>
<td>851</td>
<td>38</td>
</tr>
<tr>
<td>2020/2/19</td>
<td>616</td>
<td>81(63–100)</td>
<td>697(679–716)</td>
<td>684</td>
<td>68</td>
</tr>
<tr>
<td>2020/2/20</td>
<td>384</td>
<td>68(51–86)</td>
<td>452(435–470)</td>
<td>461</td>
<td>77</td>
</tr>
<tr>
<td>2020/2/21</td>
<td>365</td>
<td>89(70–111)</td>
<td>454(435–476)</td>
<td>570</td>
<td>205</td>
</tr>
<tr>
<td>2020/2/22</td>
<td>379</td>
<td>134(109–161)</td>
<td>513(488–540)</td>
<td>452</td>
<td>73</td>
</tr>
<tr>
<td>2020/2/23</td>
<td>371</td>
<td>201(167–238)</td>
<td>572(538–609)</td>
<td>445</td>
<td>74</td>
</tr>
<tr>
<td>2020/2/24</td>
<td>282</td>
<td>259(217–305)</td>
<td>541(499–587)</td>
<td>350</td>
<td>68</td>
</tr>
<tr>
<td>2020/2/25</td>
<td>181</td>
<td>335(277–398)</td>
<td>516(458–579)</td>
<td>347</td>
<td>166</td>
</tr>
<tr>
<td>2020/2/26</td>
<td>74</td>
<td>408(314–517)</td>
<td>482(388–591)</td>
<td>281</td>
<td>207</td>
</tr>
<tr>
<td>Total</td>
<td>22,452</td>
<td>2,106(1,641–2,628)</td>
<td>24,558(24,093–25,080)</td>
<td>24,117</td>
<td>1,665</td>
</tr>
</tbody>
</table>

arise primarily from the assumption that the lag time distribution from onset to report is relatively stable. This is difficult to achieve in the process of preventing and controlling new infectious diseases. In the case of COVID-19, for example, changes in diagnostic or reporting criteria, improvements in diagnostic techniques, and increased prevention and control efforts may change the interval between onset and reporting, e.g. the use of square-cabin hospitals and the widespread availability of nucleic acid testing have significantly reduced the lag time interval. Second, a small number of cases may have an impact on the stability of the model parameter estimates. Therefore, it is recommended to use national or province-wide pooled data for model parameter estimates at the early stages of an epidemic for new infectious diseases.

In conclusion, this study provides an early prediction method for the actual number of incident cases based on data from the surveillance report, which is of great importance to epidemic prevention and control personnel in estimating the actual occurrence of the epidemic, predicting trends, and assessing the effectiveness of prevention and control measures.

**Funding**: This study was supported by grants from the Key Joint Project for Data Center of the National Natural Science Foundation of China and Guangdong Provincial Government (U1611264), The National Major Scientific and Technological Special Project for HIV/AIDS and Hepatitis B prevention (2013ZX10004218-006, 2017ZX1030301-005, 2018ZX10201002), and the National Key Research and Development Program of China (2016YFC1200703).

**doi**: 10.46234/ccdcw2020.267

* Corresponding author: Jiaqi Ma, majq@chinacdc.cn.

1 Center for Public Health Surveillance and Information Service, Chinese Center for Disease Control and Prevention, Beijing, China; 2 School of Public Health, Peking University, Beijing, China; 3 State Key Laboratory for Diagnosis and Treatment of Infectious Diseases, Collaborative Innovation Center for Diagnosis and Treatment of
REFERENCES


5. Plummer M. Bayesian graphical models using MCMC. https://cran.r-project.org/web/packages/rjags/rjags.pdf. [2020-12-1].


Methods and Applications

Zero-Shot Medical Image Retrieval for Emerging Infectious Diseases Based on Meta-Transfer Learning — Worldwide, 2020

Yuying Zhao; Hanjiang Lai; Jian Yin; Yewu Zhang; Shigui Yang; Zhongwei Jia; Jiaqi Ma

ABSTRACT

Introduction: Due to the increasing number of medical images, image retrieval has become an important technique for medical image analytics. Although many content-based image retrieval methods have been proposed, the retrieval of images in datasets related to emerging/new infectious diseases still remain a challenge—mostly due to the lack of historical data. As a result, the current retrieval models have limited functionality in helping doctors make accurate diagnoses of new diseases.

Methods: In this paper, we propose a zero-shot retrieval model based on meta-learning and ensemble learning, which can obtain a model with stronger generalizability without using any relevant training data, and thus performs well on new types of test data.

Results: The experimental results showed that the proposed method is 3% to 5% higher than the traditional method, which means that our model can retrieve relevant medical images more accurately for newly emerging data types and provide doctors with more effective assistance.

Discussion: When a new infectious disease occurs, doctors can use the proposed zero-shot retrieval model to retrieve all relevant cases, quickly find the common problems of patients, find the locations of the new infections, and determine its infectivity as soon as possible. The proposed method is a new computer-aided decision support technology for emerging infectious diseases.

INTRODUCTION

Recently, artificial intelligence (AI) technologies have been widely used in the medical industry. With the developments of digital imaging techniques, e.g., computed tomography (CT) and X-ray, millions or even billions of medical images have been generated. Image retrieval technology (1), which retrieves similar medical images from large-scale image datasets that contain patient physiological, pathological, and anatomical information, can be used as an important objective basis for assisting doctors in clinical diagnosis, disease tracking, and surgical research. With the help of image retrieval technology, it was possible to retrieve all similar cases in the database using the patient’s medical pictures and to assist doctors in making more accurate and universal diagnoses.

In parallel, the newly emerging diseases, e.g., emerging infectious diseases, is a challenging problem for public health control. For example, coronavirus disease 2019 (COVID-19) emerged in Wuhan at the beginning of 2020 and caused numerous casualties and social losses. When a new infectious disease appears, it is difficult for the doctor, who can only rely on previous experiences, to quickly find the common patterns of the new disease without any historical data of the disease. In order to assist doctors in making a correct diagnosis quickly, a possible solution is computer-aided decision support technology, such as medical image retrieval. To analyze the new disease, the retrieval model can find all visually similar images of the new disease, which can be used to explore the common patterns of the disease, the therapeutic plan, etc. However, due to the lack of training data in new cases such as COVID-19, the performance of the existing retrieval models is greatly reduced.

The zero-shot retrieval model has been proposed to solve this problem. In the absence of relevant training data, the zero-shot retrieval model tries to find similar images from unseen image datasets. The current mainstream zero-shot learning models, such as SitNet (2) and AgNet (3), use text information as an aid to train the model. They map the image modality and the text modality into the same semantic space. In this way, semantic information of the label of unknown images can be used to learn the association between the unknown category and the known category despite the image data being unobtainable. However, for emerging infectious diseases, a new term is often used to name the disease, which has never appeared in the corpus of the text model. This will
METHODS

The goal of zero-shot retrieval is to retrieve images of novel classes although there were no training samples of these categories in the training set. Let \( S_p = \{(x_i, y_i) | y_i \in Y_{train}\} \) be denoted as the training set, where \( x_i \) is the i-th image and \( y_i \) is the class label. We also denoted \( S_w = \{(x_j, y_j) | y_j \in Y_{test}\} \) as the test set. Please note that none of the test classes occur in the training set. In this paper, we aimed to learn a retrieval model \( X \rightarrow Y \) using only the training set, and the model can also perform well on the test set.

To establish good mapping, we needed to solve the domain shift problem. Our main idea is that the training process should be the same as the testing process. The main procedures are shown in Figure 2. To introduce the virtual test domain to simulate the real test process in the zero-shot task, we split the training data into two parts with non-coincident categories at the beginning of each round of training and get the meta-train domain and meta-test domain. For example, if we have one class as the test domain, and the last 9 classes as the training domain, we can in each round of training choose one class randomly from the training domain as the meta-test domain and the last 8 classes as the meta-train domain. Our training goal is to minimize the loss of the model on the virtual training domain, while also guaranteeing that the direction of the gradient update can reduce the loss on the virtual test domain. Our training goal is actually to train the model to generalize the unknown domain. The training process is divided into three steps.

The first step is virtual meta-train. We calculate the loss of model \( \theta \) on meta-train \( F(\theta) = \text{triplet loss} (\text{meta Train}, \theta) \) and backpropagate the obtained loss \( (\theta) \) to update the network parameters so that we can obtain new weights \( \theta_1 = \theta - \alpha F(\theta) \). It is worth mentioning that the loss function we use is triplet loss, which can shorten the distance between the image’s hash codes of the same category and increase the distance between the image’s hash codes of different categories. In order to calculate the triplet loss, we first need to construct a tuple \( < I, I_{pos}, I_{neg} > \) (where the origin \( I \) is a sample randomly selected from the training data, \( I_{pos} \) is the sample of the same category as \( I \), and \( I_{neg} \) is a sample of a different category from \( I \)). The calculation formula of triplet loss is as follows:

\[
\text{triplet loss} = \max(||I - I_{pos}||_2^2 - ||I - I_{neg}||_2^2 + \text{margin}, 0) 
\]

The hyperparameter margin in the formula represents the minimum difference between \( \text{dis}(I, I_{neg}) \) and \( \text{dis}(I, I_{pos}) \).

The second step is virtual meta-test. Because our ultimate goal is not only to make the trained model perform well on the training domain, but also hope that the \( \theta_1 \) model will also perform well on the test domain. We calculate the loss of model \( \theta_1 \) on meta-test \( G(\theta_1) = \text{triplet loss(meta Test}, \theta_1) \).
For each episode $E$:

Get meta-train domain and meta-test domain by split the training data into two parts with non-coincident categories.

For each batch $i$:

- Calculate loss $F(\theta) = \text{triplet loss(meta\_train, } \theta)$;
- Update the weights of network $\theta_i = \theta - \alpha \theta_{\text{meta\_train}} = \theta - F^i(\theta)$;

- Calculate loss $G(\theta_i) = \text{triplet loss(meta\_test, } \theta)$;

Calculate final loss $P(\theta) = F(\theta) = F(\theta) + \beta G(\theta_i) = F(\theta) + \beta G(\theta - \alpha F^i(\theta))$;

Calculate learning rate for the batch $\gamma = \gamma(\theta)$;

Update the weights of network $\theta = \theta - \gamma P(\theta)$;

if mod($i$, $c$) = 0 then (c is cycle length)

  - Calculate the number of models $n_{\text{models}} = i/c$;
  - Calculate the final weights $\theta_{\text{final}} = \frac{\theta_{\text{meta\_train}} + \theta}{n_{\text{models}} + 1}$

virtual meta-train step  
virtual meta-test step  
moving average method  
meta-optimization step

FIGURE 2. Main procedures of the zero-shot medical image retrieval.

The third step is meta-optimization. We use the weighted sum of $(\theta)$ and $G(\theta)$, which is $P(\theta) = F(\theta) + \beta G(\theta_i) = F(\theta) + \beta G(\theta - \alpha F^i(\theta))$, as the final loss to update the model $\theta$ with gradient backhaul. Performing a first-order Taylor transformation on the second term $(x)$, we can get

$$\text{final loss } = F(\theta) + \beta \times G(\theta) - \beta \times \alpha \times F^i(\theta) \times G'(\theta).$$

From this formula, we can see that the final loss has two functions: 1) minimize the loss of the model in the two domains of virtual meta-train and virtual meta-test; 2) maximize the product of the loss gradient of the model in the virtual meta-train and virtual meta-test domains. The smaller the angle, the larger the vector product. Therefore, the gradient directions of these two fields can be made to be consistent. Because each round of training will re-divide meta-train and meta-test, the entire training process will make the gradient directions of any two domains to tend to be the same. Finding the direction in which the loss of two sub-problems decreases simultaneously each time to update the parameters can reduce overfitting to a single domain.

In addition, we were inspired by the idea that the random weight average can find a wider optimal range compared to SGD, so we used cyclic learning rate to train the model and used the moving average method to calculate the average of the multiple SGD trajectory as the final model.

To verify the validity of this method, we do experiments on a widely-used medical dataset to evaluate the proposed method. We randomly sampled 5% of images from the NIH Chest X-Ray Dataset (9) and created a smaller dataset (10), which contains 5,606 images that were classified into 15 classes, including 14 common chest lesions (such as atelectasis, consolidation, infiltration, pneumothorax, edema, etc.) and one for “No findings.” To simulate the situation of new diseases, such as an emerging infectious disease, we randomly selected in our experiment one disease (e.g., infiltration) as the new disease, and the other 14 types of diseases as the training set. All the samples were used as the database for retrieval. In our experiment, we trained a retrieval model on the 14 diseases and aimed to achieve a good retrieval performance on the new disease without using any data from the new disease.

In the experimental setting, all images were resized to 224×224 resolution, and we used the pretrained model Alexnet (11) to extract image features with 4,096 dimensions. The learning rate was set to $10^{-4}$ and the momentum was set to 0.9. The weight decay parameter was 0.0005. The mini batch was set to 64. We chose the conventional training method and network update method as the baseline, and mean Average Precision (mAP) based on Hamming ranking as the evaluation metric.

All the experimental settings of the traditional method we used for comparison experiments were the same as the above settings. The only difference was that meta-learning was not used in the training process, and ensemble learning was not used in the network update process. The traditional method only used Alexnet (11) to extract image features, and then
obtained the hash code of the image and finally used triplet loss as the loss function and SGD as the network update method to train the network.

**RESULTS**

The comparison results were shown in Table 1. From Table 1, we can see that, in terms of retrieval index mAP, the proposed method is 3% to 5% higher than the traditional method, indicating that our method is effective. In addition, we can also see that we have tried 4 different lengths of hash codes, which are 8 bits, 16 bits, 32 bits, and 48 bits, which increase to 5.342%, 3.148%, 3.769%, and 4.527%, respectively. In the case of all hash code lengths, the proposed method has higher retrieval accuracy than traditional methods, which demonstrates the effectiveness of our proposed method.

**DISCUSSION**

In our study, we pointed out the importance of popularizing artificial intelligence applications in the diagnosis of emerging infectious diseases and analyzed the limitations of existing image retrieval models on this issue. In response to the lack of training samples and inaccurate cognition of new terms, we proceeded from the perspective of improving the model's generalizability for new categories, and finally proposed a zero-shot hashing model that can achieve good retrieval results without using additional text tags. We verified the effectiveness and feasibility of the proposed method on a widely used medical dataset and found that the simulation test process can indeed make the model accustomed to identifying new categories. The application flowchart of our image retrieval model was shown in Figure 3. Experimental results showed that our model could retrieve relevant pictures more accurately, so the proposed model could be used to assist doctors in making the correct diagnosis quickly when an emerging infectious disease occurs and improve public health.

This study was subject to some limitations. First, the loss function used in this method was no different from the ordinary retrieval model. We can further explore better loss functions that can improve the generalizability of the model, such as adding regularization items or considering the relationship between different classes. Second, if supplementary information of the sample is added to this method, such as image attributes, the retrieval effect can be further improved. Next, we will consider and practice these ideas in more detail.

**Fundings:** This study was supported by grants from the Key Joint Project for Data Center of the National Natural Science Foundation of China and Guangdong Provincial Government (U1611264), and the Pearl River Nova Program of Guangzhou (201906010080).

**TABLE 1. mAP of baseline and proposed method on Chest X-ray Dataset.**

<table>
<thead>
<tr>
<th>Method</th>
<th>8 bits</th>
<th>16 bits</th>
<th>32 bits</th>
<th>48 bits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>27.523</td>
<td>31.231</td>
<td>31.472</td>
<td>32.825</td>
</tr>
<tr>
<td>Ours</td>
<td>32.865</td>
<td>34.379</td>
<td>35.241</td>
<td>37.352</td>
</tr>
</tbody>
</table>

*Corresponding author: Hanjiang Lai, laihanj3@mail.sysu.edu.cn.*

1 School of Computer Science and Engineering, Sun Yat-Sen University, Guangzhou, Guangdong, China; 2 Center for Public Health Surveillance and Information Service, Chinese Center for Disease Control and Prevention, Beijing, China; 3 State Key Laboratory for Diagnosis and Treatment of Infectious Diseases, Collaborative Innovation Center for Diagnosis and Treatment of Infectious Diseases, The First Affiliated Hospital, College of Medicine, Zhejiang University, Hangzhou, Zhejiang, China; 4 School of Public Health, Peking University, Beijing, China.

**FIGURE 3.** Schematic diagram of image retrieval for new type of lung disease image.
REFERENCES


Commentary

Internet Data for Improving Prevention and Control of Global Infectious Diseases

Zhongwei Jia1,2,3,4,; Xiangyu Yan1; Yongjie Li1; Jiaqi Ma4

By the end of 2018, there were about 5.1 billion internet users worldwide who accounted for 67% of the total population (1). Health problems have emerged following widespread use of the internet as a new model of social activity, such as anonymous personal attacks or abuse and information leakages. A more potent threat is the unverified information being provided by the internet or electronic media that can lead to unsubstantiated judgments and improper responses to information regarding an individual’s health status. A real example was the Infodemic that occurred at the beginning of the coronavirus disease 2019 (COVID-19) pandemic (2).

In 2002, Gunther Eysenbach brought up a creative concept named “infodemiology” to investigate the impact of various information, especially that stemming from the web or electronic media, on people’s health (3). The information sources of infodemiology cover almost all forms of electronic media, such as browser search histories, social media, mobile phone apps, and some reprocessed internet data (e.g. Google Trends) (4). Infectious diseases, such as influenza and Zika virus were the most concerned topics in infodemiology. However, chronic diseases, such as diabetes, cancers, and health-related behaviors, such as drug use, suicide, smoking, and diet were also followed and covered (4). Infodemiology is primarily used for monitoring “trends of information” through distribution and change of health-related information rising from web or electronic media in order to provide alarm or prediction. How to identify reasons related with information, further implement interventions, evaluate the effect of interventions, and optimize the scheme have not yet been solved in infodemiology and are still open. Therefore, the theoretical research system of infodemiology needs to be expanded.

Because information is an outcome of internet users and each internet user is related with an actual individual in physical space, it is possible to detect the reasons and motivations of people who produce a kind of internet information. Based on different reasons, health interventions can be devised based on users’ characteristics and behaviors, which will be beneficial for those who regard internet and other electronic media as an important resource for health information. These terms and objectives are all covered by epidemiology, but by only focusing on internet users, new informatics methods are needed to analyze data of internet users.

To strengthen the usage and analysis of data rising from the internet, it is necessary to study the characteristics, distribution, and impact factors of internet events related with health and diseases (including the behavior, distribution, and influencing factors of internet users, laws associated with epidemics, and trends and redundancy benefits of internet information); explore the correlation between events on the internet and in real life; optimize the prevention and control strategies of internet information; and serve public health and social governance (5). Integration of traditional data and real-time internet data of individuals and subpopulations is the object of further internet studies, and interdisciplinary research methods, such as natural language processing, knowledge graph and machine learning techniques have become the basic methods in this discipline. The most challenging question for further studies is linking an internet event and a physical event. In addition, protecting personal privacy of internet users’ health and behavior information is also a rising concern, and methods such as anonymous processing of heterogeneous data from multiple sources must be further studied. Great impact and reform will be also confronted by traditional ethics in internet era.

An example of optimizing use of internet data is the ability to improve HIV/AIDS prevention and control for men who have sex with men (MSM). MSM accounted for 23.0% of new HIV infections in China, and homosexual transmission was the second major HIV transmission route after heterosexual transmission in 2019 (6). The internet socializing platforms, such as some geosocial networking applications (GSN apps), have been blamed for this problem. Recent studies indicated that about 41%–63% of Chinese MSM had
experienced looking for the casual sexual partners thought GSN apps, among which HIV incidence was about 4 times higher than that of non-users (8.5/100 person-years \textit{vs.} 2.0/100 person-years) (7). The anonymity and convenience of dating online was the main reason for GSN becoming popular among MSM, but it was also the primary culprit of increasing HIV infection rates among MSM because these casual sexual partners dating online did not know each other before, let alone their HIV statuses (8). In order to avoid these risk behaviors, HIV-related knowledge was recommended to be publicized and delivered through the internet, but the study showed that only about 50% of MSM in China were willing to browse HIV-related knowledge through the internet and electronic media (9).

Knowledge graph-based frameworks can be effectively used to provide this knowledge and will be the base for interventions and control efforts online. Knowledge graphs are constructed using graph model based on internet data sources that include two parts, one dataset is the user data from GSN apps, which is used to construct label-based user persona, and another dataset is the data from a website [e.g. World Health Organization (WHO) or Joint United Nations Programme on HIV and AIDS] and literature (e.g. PubMed or Embase), which provides high-quality and evidence-based health information that can be used to build health knowledge graphs and give targeted and timely interventions for users of GSN apps.

Data from GSN apps are composed of structured and semi-structured data, which include demographic information of MSM, self-introduction information, and dating requirements. In demographic information, there are age, height, and weight. The self-introduction information covers MSM users reporting their sex role, physical, and personality characteristics. Dating requirements include their criteria for whom they want to meet. Integrating these data, a personalized user persona for each MSM user will be created by two steps. First, 20% of the user data will be annotated from 5 dimensions by professional staff, which include demographic attributes, social attributes (social position, social relations), behavior habits (lifestyle, sexual behaviors, dating history), interest preferences (shopping, games), and disease statuses (HIV and other sexual transmitted diseases, chronic diseases, and psychological statuses). Second, a support vector machine model will be trained to classify the remaining users and add the other relevant labels (Figure 1).

A health knowledge graph of MSM can then be established by professional information extracted from the unconstructed text of WHO and the literature databases (e.g. PubMed), which includes common diseases and symptoms, professional intervention guidance materials of MSM, disease-specific medical guidelines, and treatment measures for infectious diseases such as HIV, syphilis, etc. The whole process was divided into 4 steps: data cleaning, information extraction, knowledge fusion, and knowledge reasoning (Figure 1) (10). First, data cleaning—the unstructured texts usually contain noise information, so we removed all the special characteristics (such as @, emoji) and stop words. Second, information extraction—for unstructured data, we performed entity extraction, relationship extraction, and attribute extraction on these text data. We used the Long Short–Term Memory Neural Network (LSTM) for named entity recognition (11). We converted the filtered text into word vectors and input them into LSTM, and the output was the named entities labeled in the sentences. We also introduced an attention mechanism to improve name recognition. For example, the extracted entities were sexual behavior, HIV, and condoms. For relation extraction, we chose the weakly supervised learning model to extract relationships for data lacking adequate labels. The OpenNRE was carried out to extract the relationship between diseases and symptoms, diseases and treatment measures, diseases and medical institutions, high-risk behavior and intervention guidance, infectious diseases and detection methods, and so on. We used a rule-based method to extract attributes, including symptoms of common diseases, intervention methods, and the outcome of taking antiviral drugs, etc. For example, an early symptom of HIV is herpes zoster, which has a high-risk sexual infection rate. Third, knowledge fusion—we merged different expressions of the same entity extracted from different literature or websites, including the merging of entities and the merging of entity relationships, such as hepatitis B virus and HBV. Hierarchical clustering was used to calculate the similarity between different entities and to partition the entities at different levels and finally form a tree-like clustering structure. Entities with higher similarity were more likely to be merged into the same one, and the fusion results will be reviewed by experts. The final constructed health knowledge graph was represented as an “entity-relationship-entity” triple; for example, anal sex-interventions-condoms, hepatitis B-clinical
Fourth, knowledge reasoning—based on the completed user persona of MSM, we used the health knowledge graph to push health intervention information for MSM and predict high-risk behaviors. For example, we can push information about the correct use of condoms for people with suspicious high-risk behavior labels. The entire reasoning process will be handled by the convolutional neural network (12). Finally, as the data continues to increase, our health knowledge graph will be updated to improve reasoning ability and efficiency.

**CONCLUSION**

It will be not comprehensive and objective to investigate a specific health problem without integrating the network and electronic media information today. Problems originating from the internet should be solved based on the internet, and it is suitable for the health problems. Using internet-based big data to improve disease prevention and control is still a novel subject which is aiming to narrow the rising gap between health issues caused by
internet and the intervention methods in the physical space. With the increasing expansion of network data, it is meaningful in the monitoring and intervention of specific populations on the internet. In this report, we tried to deliver a general idea of public health by a practical case of intervention and control online on MSM. More new approaches deserve to be explored and applied in health problems rising from internet users and enrich the research system of traditional epidemiology. New privacy protection mechanisms need further exploration.

**Funding:** Key Joint Project for Data Center of the National Natural Science Foundation of China and Guangdong Provincial Government (U1611264); and the National Natural Science Foundation of China (91546203, 91846302).

* doi: 10.46234/ccdcw2020.269

* Corresponding author: Zhongwei Jia, urchinjj@163.com.

1 School of Public Health, Peking University, Beijing, China; 2 Center for Intelligent Public Health, Institute for Artificial Intelligence, Peking University, Beijing, China; 3 Center for Drug Abuse Control and Prevention, National Institute of Health Data Science, Peking University, Beijing, China; 4 Chinese Center for Disease Control and Prevention, Beijing, China.

Submitted: December 03, 2020; Accepted: December 22, 2020

**REFERENCES**


Zhongwei Jia, PhD, Professor
School of Public Health, Institute for Artificial Intelligence, Peking University
Dr. Yi Zeng (1929–2020)
A Pioneer on HIV and Tumor Virology in China

Xiaoguang Zhang; Yongjun Zhang

Dr. Yi Zeng, a distinguished virologist, an academician of the Chinese Academy of Sciences, and a Chief Scientist of National Institute for Viral Disease Control and Prevention in China CDC, passed away at the age of 92 on July 13, 2020 in Beijing. He was Director of the Institute of Virology in Chinese Academy of Preventive Medicine (1983–1985), President of Chinese Academy of Preventive Medicine (1991–1996), President of Chinese Association of Preventive Medicine (1999–2005), and Principal of the College of Life Sciences and Bioengineering in Beijing University of Technology (2002–2014). He was a foreign academician of Russian Academy of Medical Sciences (since 1995), the French National Academy of Medicine (since 2003), and the Russian Academy of Sciences (since 2014).

Zeng was born in March 1929 in Wujingfu Village, Jiexi County, Guangdong Province and graduated from Shanghai Medical College in 1952. He worked for 4 years in Guangdong and moved to Beijing in 1956 where he joined the Virology Laboratory, Department of Microbiology, Chinese Academy of Medical Sciences [CAMS, later changed to be affiliated with Chinese Academy of Preventive Medicine in 1983 (now known as China CDC)], where he started his research in virology for more than 6 decades.

In the 1950s–1960s, Zeng and his colleagues initiated a nationwide epidemiological investigation in response to a large prevalence polio cases and conducted isolation and identification of polio virus. He also participated in studies on live-attenuated polio vaccines, which contributed the successful eradication of polio in China.

Since 1960, Zeng has focused on tumor viruses including polyomaviruses and avian leukosis virus (ALV). In his collaborations with colleagues, he helped report a positive rate of ALV of over 80% in hens in China and provided support for establishing ALV-free chickens. He was the first researcher to propose Epstein-Barr virus (EBV) as a direct cause of nasopharyngeal carcinoma (NPC), to develop diagnostic assays to identify antibodies against EBV, and to establish the first EBV-positive NPC cell line in the world from biopsy specimens of an NPC patient. Zeng also established the Coordination Center for NPC Prevention and Treatment in Wuzhou, Guangxi in 1970, which was the first field institution for NPC in the world and saved countless patients.

As an enthusiastic social activist, Zeng promoted health education on AIDS, gave lectures on AIDS prevention knowledges, and contributed greatly in improving AIDS awareness in the general population of China during the past decades. As a distinguished educator, he educated a large amount of elite talented researchers for China.

Zeng has dedicated his life to virology research and the cause of disease prevention, for which he organized and participated in its practice and policy-making in China and had made outstanding contribution to people’s health. He will continue to live on in our hearts, leading us to explore unsolved fields in modern virology.

1 National Institute for Viral Disease Control and Prevention, China CDC, Beijing, China; 2 Chongqing Medical and Pharmaceutical College, Chongqing, China.
Jibin Tan
(1972–2020)
Director of China CDC Center for Logistics Operation and Management

Jibin Tan, MPH, PhD, Director of China CDC Center for Logistics Operation and Management, and Vice Dean of the China CDC Graduate School, passed away at 02:40 on December 20, 2020 at the age of 48.

Born in Jiaozhou City, Shandong Province in November 1972, Tan graduated from Jining Medical College and then received his MPH degree from Peking University School of Public Health and a PhD degree from the School of Public Health at Jilin University. He joined the Qingdao City Epidemic Response Station in July 1994.

After joining China CDC, he became director and core expert for the Global Fund’s tuberculosis control project and had undertaken roles for top-level design, organization, implementation, project management, and evaluation for the project. He had made outstanding contributions to China’s tuberculosis prevention and control work.

Tan was responsible for the new site office of China CDC and the Center for Logistics Operation and Management in 2009. He successfully completed the smooth and orderly relocation of China CDC to the Northern (Changping) Campus and created a modern office environment for employees, promoted socialized and integrated transformation of managing logistics, and established a refined operation management system by continuously researching modern logistics management and intelligent disease control campus construction and development.

In addition, he served as Vice Dean of the China CDC Graduate School and actively promoted improvement of basic conditions for postgraduate education. Tan was honored with numerous awards/prizes including Outstanding Individual for Advancing the Eradication of Poliomyelitis in China.

In the face of the coronavirus disease 2019 (COVID-19) epidemic, Tan acted as the leader of the China CDC Safeguard Team and took initiative in overcoming extremely difficult challenges. When there were serious shortages of materials in the central emergency inventory, he extensively mobilized and actively coordinated the opening up of national emergency material safeguard channels. To ensure timely deployment of these supplies, he was tasked with coordinating the China CDC Safeguard Team response with three national ministries (National Health Commission, the National Development and Reform Commission, and the Ministry of Industry and Information Technology) to establish a point-to-point supply deployment mechanism. Tan made great contributions in our fight against COVID-19.

Tan was conscientious, sought the truth, pragmatic, selfless, and dedicated. He worked hard for the country until the day before his death with an emergent medical condition. He will be remembered dearly by his peers. His departure is a major loss for China CDC. Tan is survived by his wife, son, brother, and mother.

Chinese Center for Disease Control and Prevention.