Observational Study

Medicine®

OPEN

# A Bayesian classification model for discriminating common infectious diseases in Zhejiang province, China

Fudong Li, MM[a], Yi Shen, BM[b], Duo Lv, MM[c], Junfen Lin, MPH[a], Biyao Liu, MM[a], Fan He, MM[a], Zhen Wang, BM[a,*]

**Abstract**

To develop a classification model for accurately discriminating common infectious diseases in Zhejiang province, China.

Symptoms and signs, abnormal lab test results, epidemiological features, as well as the incidence rates were treated as predictors, and were collected from the published literature and a national surveillance system of infectious disease. A classification model was established using naïve Bayesian classifier. Dataset from historical outbreaks was applied for model validation, while sensitivity, specificity, accuracy, area under the receiver operating characteristic curve (AUC) and M-index were presented.

A total of 146 predictors were included in the classification model, for discriminating 25 common infectious diseases. The sensitivity ranged from 44.44% for hepatitis E to 96.67% for measles. The specificity varied from 96.36% for dengue fever to 100% for 5 diseases. The median of total accuracy was 97.41% (range: 93.85%–99.04%). The AUCs exceeded 0.98 in 11 of 12 diseases, except in dengue fever (0.613). The M-index was 0.960 (95%CI 0.941–0.978).

A novel classification model was constructed based on Bayesian approach to discriminate common infectious diseases in Zhejiang province, China. After entering symptoms and signs, abnormal lab test results, epidemiological features and city of disease origin, an output list of possible diseases ranked according to the calculated probabilities can be provided. The discrimination performance was reasonably good, making it useful in epidemiological applications.

**Abbreviations:** AI = artificial intelligence, AUC = area under the receiver operating characteristic curve, CI = confidence interval, CISDCP = China Information System for Disease Control and Prevention, FN = false negative, FP = false positive, GIDEON = Global Infectious Disease and Epidemiology Network, ROC = receiver operating characteristic, TN = true negative, TP = true positive.

**Keywords:** Bayes, classification, diagnosis, discrimination, infectious diseases

## 1. Introduction

Zhejiang province is located on the southeast coast of China, with high incidence rates of many infectious diseases.[1] In recent years, outbreaks of infectious diseases were still common in many counties. Therefore, the prevention and control of infectious diseases is a public health priority in Zhejiang. The early detection and efficient control are very important for prevention of further spreading of the disease. As a consequence, epidemiological field investigators must determine the cause of outbreaks as soon as possible. However, it can be quite challenging because the available information is always limited at the early phase of an outbreak.[2] In many cases, epidemiological field investigators make a diagnosis based on personal experience and understanding of various diseases, with possibility of misdiagnosis. Timely and accurate diagnoses are vital to ensuring that the proper measures for disease control will be administered and new cases will be minimized.[3]

Recently, data mining techniques, including Bayesian classifiers, decision tree classifiers and neural network classifiers, have been the widely utilized for discriminating or diagnosing diseases.[4] The infectious disease diagnosis module within the well-known Global Infectious Disease and Epidemiology Network (GIDEON) was developed based on Bayesian formula.[5,6] Decision tree was applied for psychiatric diagnosis.[7] Cualing et al[8] and Shaw[9] also used this technique to assist with diagnosing various diseases. Moreover, neural network technique was used to establish a system for diagnosing acute myocardial infarction.[10] These studies supported the utility of

*Editor: Leyi Wang.*

*FL and YS contributed equally to this work and should be considered as co-first authors.*

*This study was supported by the Program for Zhejiang Leading Team of Science and Technology Innovation (2011R50021), the Medical Research Program of Zhejiang Province (2017KY286), and Zhejiang Provincial Natural Science Foundation of China (LQ19H260001). The funding sources had no involvement in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The researchers confirm their independence from funders and sponsors.*

*The authors declare that they have no conflicts of interests.*

*Supplemental Digital Content is available for this article.*

*[a] Zhejiang Provincial Center for Disease Control and Prevention, [b] Department of Epidemiology and Health Statistics, School of Public Health, Zhejiang University, [c] The First Affiliated Hospital of Zhejiang University, Hangzhou, Zhejiang Province, People's Republic of China.*

*[*] Correspondence: Zhen Wang, Zhejiang Provincial Center for Disease Control and Prevention, 3399 Binsheng Road, Binjiang District, Hangzhou, Zhejiang 310051, People's Republic of China (e-mail: zjwangzhen@126.com)*



*How to cite this article: Li F, Shen Y, Lv D, Lin J, Liu B, He F, Wang Z. A bayesian classification model for discriminating common infectious diseases in zhejiang province, China. Medicine 2020;99:8(e19218).*

*Received: 5 October 2019 / Received in final form: 30 December 2019 / Accepted: 20 January 2020*

*http://dx.doi.org/10.1097/MD.0000000000019218*

disease classification. However, many of them were based on clinical data of individual cases, and the models in these studies are typically designed to assist medical professionals for clinical diagnoses. Limited research is available within the public health field to assist in determining the causative disease or pathogen of an epidemic or outbreak. Last, although the incidence rates of infectious diseases are relatively high in Zhejiang, no such study has been performed for this region.

Given the fact that Bayesian classifier possesses high predictive accuracy and is suitable for large database,[11] this study utilized Bayesian method to construct a classification model for discriminating common infectious diseases in Zhejiang province. The model was designed to provide epidemiological field investigators with an artificial intelligence (AI)-based, efficient method for discriminating various infectious diseases at the early phases of an epidemic or outbreak. It was expected to promote timely implementation of appropriate control measures, and provide potential clues to narrow the range of laboratory pathogens screening.

## 2. Methods

### 2.1. Classification algorithm

A naïve Bayes algorithm was used in this study.[12] There is a group of diseases that contains j types of disease: $\mathbf{D_{total}} = (D_1, D_2 \ldots D_j)$. The prior probabilities of these diseases are $P(D_1)$, $P(D_2) \ldots P(D_j)$. Furthermore, there are k attributes or predictors including symptoms and signs, abnormal lab test results and epidemiological features, that form a set of attributes: $\mathbf{S_{total}} = \{S_1, S_2 \ldots S_k\}$. The conditional probabilities of these attributes when certain diseases exist are $P(S_1|D_j)$, $P(S_2|D_j) \ldots P(S_k|D_j)$.

When a patient presents n attributes, which form a set of presence attributes: $\mathbf{S} = \{S_1, S_2 \ldots S_n\}$, the posterior probability of a disease for this patient, according to the Bayesian formula, would be:

$$P(D_f|\ \mathbf{S}) = \frac{P(D_f) \times P(\ \mathbf{S}|D_f)}{\sum_{i=1}^{j} P(D_i) \times P(\ \mathbf{S}|D_i)} = \frac{P(D_f) \times \prod_{m=1}^{n} P(S_m|D_f)}{\sum_{i=1}^{j} P(D_i) \times \prod_{m=1}^{n} P(S_m|D_i)}$$

$f = 1, 2 \ldots j$.

$P(D_f|\mathbf{S})$ is the probability of the $f^{th}$ type of disease being accompanied by the presence of attribute set **S**. The probability of the disease depends on the value of $P(D_f|\mathbf{S})$. That is, if the value of $P(D_g|\mathbf{S})$ is the highest of all j posterior probabilities, then the likelihood of the $g^{th}$ type of disease is the highest with the presence of attribute set **S**, and $D_g$ is the maximum likelihood diagnosis. Finally, all possible diseases with posterior probabilities are ranked from highest probability to lowest probability, and presented on the output list.

A flow chart of the algorithm is shown in Figure 1. The model was performed using SAS (V.9.3, SAS institute) software.

### 2.2. Data collection

The prior probability, $P(D_j)$, was estimated according to the incidence rates of all included infectious diseases in every cities of Zhejiang province. The incidence data was collected from the China Information System for Disease Control and Prevention (CISDCP),[13] a national surveillance system of infectious disease reported by medical institutions in real time.

The conditional probability, $P(S_k|D_j)$, was estimated based on the frequency of the corresponding attribute prevalent in individuals with each disease. Epidemiological features were collected from CISDCP. The symptoms and signs, as well as abnormal lab test results within each specific disease were derived from the epidemiology literature. The data including total number of patients, numbers of patients with each symptoms and signs, and numbers of patients with each abnormal lab test result, were abstracted from each literature to calculate weighted frequencies.

We systematically searched the following Chinese databases: the China Knowledge Resource Integrated Database (www.cnki.net), Wanfang Data (wanfangdata.com.cn), VIP Journal Integration Platform (www.cqvip.com) and China Biology Medicine disc (www.sinomed.ac.cn). The search terms included "epidemic investigations" OR "outbreak investigation" AND names of each infectious disease. The references in published articles were also searched. Initially, titles and abstracts were screened to exclude ineligible studies. Then the full texts were reviewed for all the remaining studies. The literature screening procedures are presented in Figure 2.

### 2.3. Model validation

Dataset from historical outbreaks was utilized to validate the model. The data was collected from several outbreak investigations of infectious disease in Zhejiang province. During each investigation, epidemiological and clinical data has been collected based on a standard questionnaire by trained investigators for each patient. Most symptoms and signs were recorded in a dichotomized way (yes/no). General laboratory results have been documented as continuous variables with threshold of normality if available. Infectious etiology was determined according to strict case definitions from the Chinese Guideline of Diagnosis and Treatment for corresponding diseases issued by the National Health Commission of the People's Republic of China.

The sensitivity, specificity, total accuracy, and area under the receiver operating characteristic curve (AUC) have been widely used as criteria for evaluating a diagnosis model.[14] The following terms are fundamental to understanding the utility of them:

a) True positive (TP): the patient has a disease and the prediction is positive.
b) False positive (FP): the patient does not have a disease but the prediction is positive.
c) True negative (TN): the patient does not have a disease and the prediction is negative.
d) False negative (FN): the patient has a disease but the prediction is negative.

The sensitivity of a diagnosis model refers to the ability of the model to correctly identify those patients with the disease:

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

The specificity of a diagnosis model refers to the ability of the test to correctly identify those patients without the disease:

$$\text{Sensitivity} = \frac{TN}{FP + TN}$$

The accuracy of a diagnosis model refers to the ability of the model to correctly identify those patients with the disease and

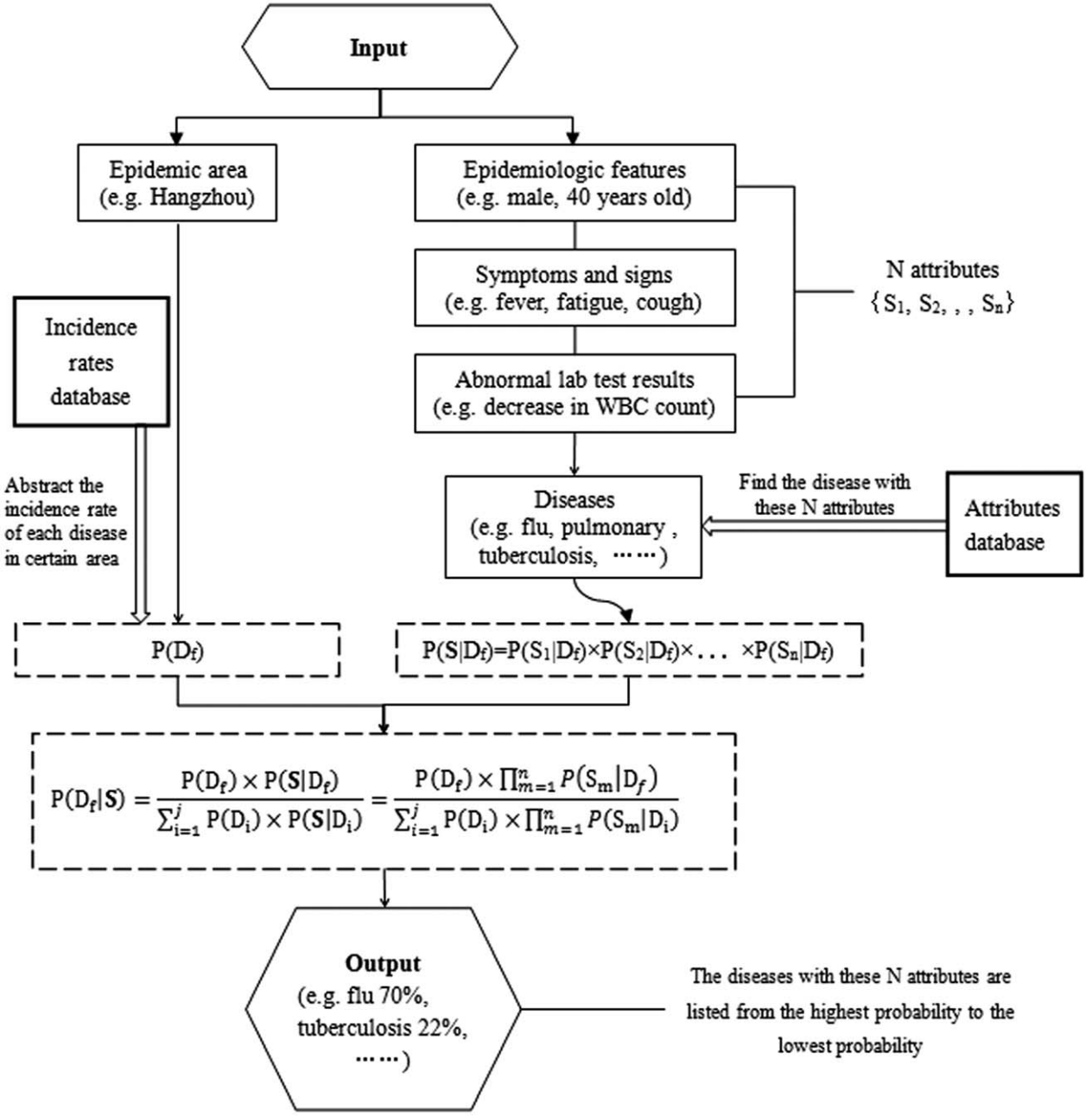


**Figure 1.** Flow chart of the algorithm.

without the disease:

$$\text{Total accuracy} = \frac{TP + TN}{TP + Fp + TN + FN}$$

The receiver operating characteristic (ROC) plot expresses relationship between sensitivity and 1-Specificity. The closer the ROC curve is located to upper-left hand corner, the better the model. The AUC can have any value between 0 and 1 and it is a good indicator of the goodness of the model.

We primarily employed these four parameters to assess discrimination performance of the model. The above-mentioned parameters were usually applied for binary outcomes. Considering the model designed for discriminating various diseases (polytomous outcomes), we obtained these parameters of category *i* by comparing category *i* with all other categories combined (1-vs-rest measure).[15] Additionally, the M-index,[16] a pairwise approach that averages all pairwise AUCs, was also evaluated, where the pairwise AUC measures the discrimination between any two categories. It is suggested independent of the category prevalence,[17] with 0.5 and 1 as the values represented for random and perfect discrimination. All results were presented as point estimation with 95% confidence intervals (CIs).

## 3. Results

The initial search identified 2963 potentially relevant articles of 25 infectious disease. After screening duplicate records, titles,

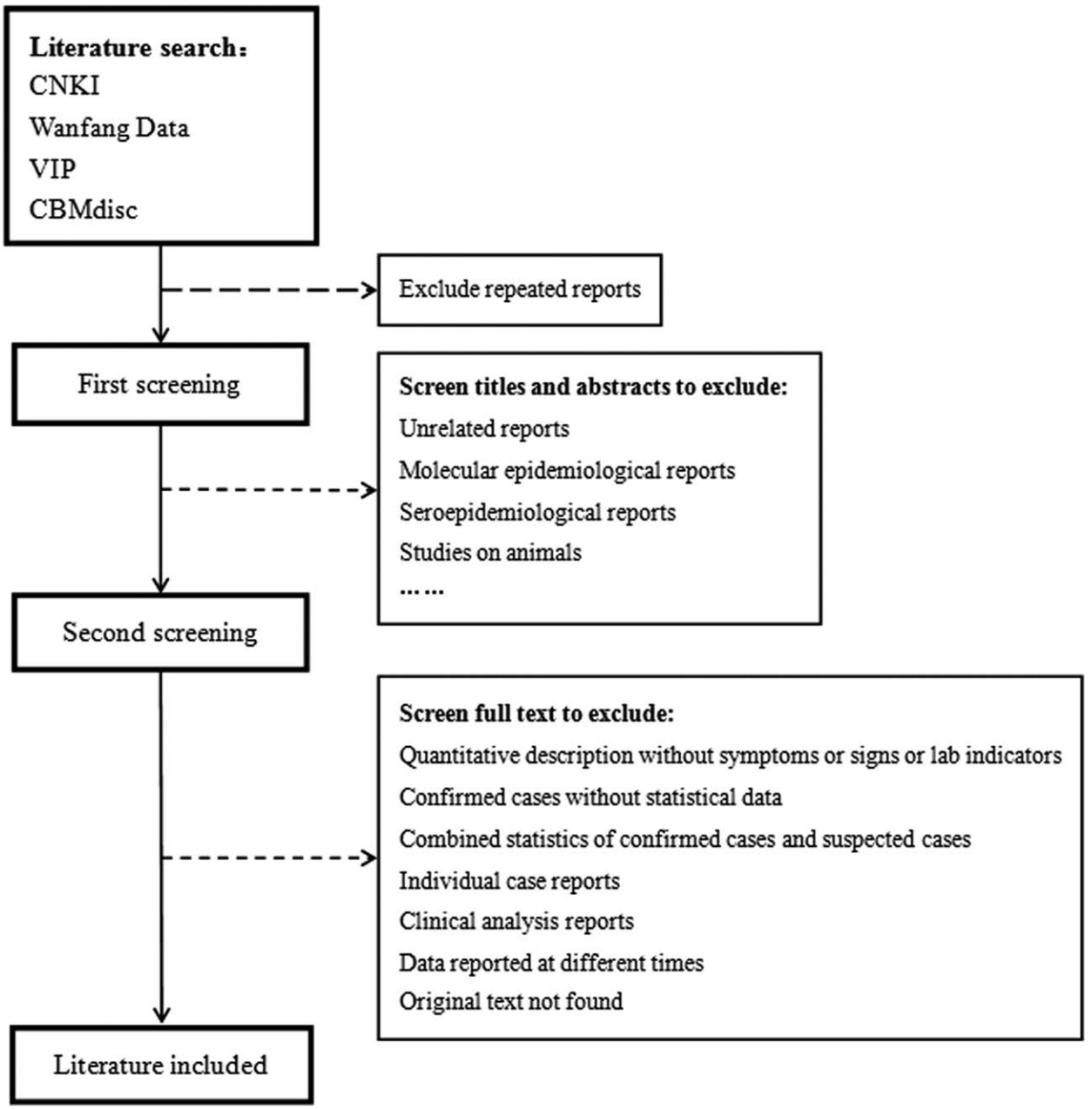


**Figure 2.** Flow diagram of literature screening procedure. CNKI, China Knowledge Resource Integrated Database; VIP = VIP Journal Integration Platform; CBMdisc = China Biology Medicine disc.

abstracts and full texts, 2400 articles were exclude. Finally, 563 articles were included for further data extraction. Table 1 showed the included 25 infectious diseases with the information on included literature. Frequencies of symptoms and signs, and abnormal lab test results were derived for estimation of conditional probabilities (see supplementary material, http://links.lww.com/MD/D825). Meanwhile, the data of incidence rates as well as epidemiological features (age and gender) within each disease was collected to establish the database for model construction.

Dataset from historical outbreaks in Zhejiang province involving 12 diseases were used for model validation. Patient's characteristics of the validation dataset were summarized in Table 2. A total of 520 cases were included in validation dataset. The sample sizes of 12 diseases ranged from 13 to 93. The mean age of all patients was 22.37 years, with a highest mean age (71.06 years) in those diagnosed with Hepatitis E. 66.92% (348/520) were male, with a male-to-female ratio of 2.02:1. The calendar year of the outbreaks varied from 2005 to 2012. The majority of diseases possessed data from one outbreak.

The validation results were presented in Table 3. The highest sensitivity of the model was achieved for measles (96.67%), and the lowest for hepatitis E (44.44%). The specificity varied from 96.36% for dengue fever to 100% for 5 diseases including leptospirosis, acute hemorrhagic conjunctivitis, epidemic cerebrospinal meningitis, hepatitis E, and epidemic hemorrhagic fever. The median of total accuracy was 97.41% (range: 93.85% for dengue fever to 99.04% for bacillary dysentery). The AUCs

**Table 1**

**Infectious diseases included in the model.**

| Disease | No. of literature* | No. of patients# | Disease | No. of literature* | No. of patients# |
|---|---|---|---|---|---|
| Amebic dysentery | 3 | 192 | Japanese encephalitis | 12 | 1079 |
| Acute hemorrhagic conjunctivitis | 10 | 5437 | Leptospirosis | 23 | 1996 |
| Bacillary dysentery | 33 | 2851 | Malaria | 17 | 360 |
| Brucellosis | 15 | 1080 | Measles | 33 | 2275 |
| Chickenpox | 64 | 4505 | Mumps | 24 | 1500 |
| Cholera | 22 | 985 | Paratyphoid fever | 43 | 4991 |
| Concurrent outbreak of typhoid and paratyphoid fever | 7 | 1567 | Pertussis | 9 | 562 |
| Dengue fever | 21 | 2493 | Pulmonary tuberculosis. | 26 | 432 |
| Epidemic cerebrospinal meningitis | 21 | 391 | Rubella | 8 | 1519 |
| Epidemic hemorrhagic fever | 14 | 1655 | Scarlet fever | 15 | 1787 |
| H1N1 Influenza | 27 | 1154 | Typhoid fever | 45 | 2861 |
| Hand Foot and Mouth Disease | 59 | 13,475 | Typhus | 8 | 785 |
| Hepatitis E | 7 | 149 | | | |

* Number of literature included finally.

# Total number of patients reported in included literature.

**Table 2**

**Patient's characteristics of the validation dataset.**

| Disease | Sample size | Gender (N, %) | | Age (years, mean, SD) | Year | Region |
|---|---|---|---|---|---|---|
| | | Male | Female | | | |
| Leptospirosis | 55 | 27 (49.09) | 28 (50.91) | 49.13 (11.65) | 2007 | Pan'an |
| Chickenpox | 48 | 25 (52.08) | 23 (47.92) | 8.00 (1.07) | 2006 | Jindong |
| Acute hemorrhagic conjunctivitis | 74 | 69 (93.24) | 5 (6.76) | 14.42 (6.37) | 2007 | Jindong |
| Bacillary dysentery | 20 | 19 (95.00) | 1 (5.00) | 36.80 (9.89) | 2006 | Pan'an |
| Epidemic cerebrospinal meningitis | 13 | 7 (53.85) | 6 (46.15) | 17.46 (13.04) | 2005 | Pujiang, Wuyi, Yiwu |
| Hepatitis E | 18 | 17 (94.44) | 1 (5.56) | 71.06 (10.98) | 2006 | Wuyi |
| Mumps | 71 | 44 (61.97) | 27 (38.03) | 9.14 (2.84) | 2005 | Jiande |
| Epidemic hemorrhagic fever | 33 | 25 (75.76) | 8 (24.24) | 51.09 (13.66) | 2010, 2011, 2012 | Shangyu |
| Hand, foot and mouth disease | 93 | 52 (55.91) | 41 (44.09) | 2.91 (1.89) | 2011, 2012 | Shangyu |
| Dengue fever | 27 | 14 (51.85) | 13 (48.15) | 41.33 (19.39) | 2009 | Yiwu |
| Measles | 30 | 15 (50.00) | 15 (50.00) | Unrecorded | 2013 | Jiande |
| H1N1 influenza | 38 | 34 (89.47) | 4 (10.53) | 22.24 (4.80) | 2009 | Dongyang, Yiwu |
| Overall | 520 | 348 (66.92) | 172 (33.08) | 22.37 (21.18) | | |

exceeded 0.98 in 11 of 12 diseases, except in one disease (0.613 for dengue fever). The M-index (0.960, 95%CI 0.941–0.978) appeared very close to 1, which also indicated high discrimination performance of the model.

## 4. Discussion

A novel classification model was established for discriminating common infectious diseases in this study. The model can diagnose 25 common infectious diseases in Zhejiang province based on

**Table 3**

**Validation results of the classification model.**

| Disease | Sensitivity (%) | Specificity (%) | Total accuracy (%) | AUC | M-index |
|---|---|---|---|---|---|
| Leptospirosis | 74.55 (70.81–78.29) | 100.00 (100.00–100.00) | 97.31 (95.92–98.70) | 1.000 (0.999–1.000) | |
| Chickenpox | 93.75 (91.67–95.83) | 97.67 (96.37–98.97) | 97.31 (95.92–98.70) | 0.985 (0.972–0.999) | |
| Acute hemorrhagic conjunctivitis | 79.73 (76.27–83.19) | 100.00 (100.00–100.00) | 97.11 (95.67–98.55) | 1.000 (1.000–1.000) | |
| Bacillary dysentery | 95.00 (93.13–96.87) | 99.20 (98.43–99.97) | 99.04 (98.20–99.88) | 0.996 (0.988–1.000) | |
| Epidemic cerebrospinal meningitis | 61.54 (57.36–65.72) | 100.00 (100.00–100.00) | 99.03 (98.19–99.87) | 0.998 (0.996–1.000) | |
| Hepatitis E | 44.44 (40.17–48.71) | 100.00 (100.00–100.00) | 98.08 (96.90–99.26) | 0.983 (0.971–0.995) | 0.960 (0.941–0.978) |
| Mumps | 88.73 (86.01–91.45) | 99.78 (99.38–100.00) | 98.27 (97.15–99.39) | 0.997 (0.995–1.000) | |
| Epidemic hemorrhagic fever | 60.61 (56.41–64.81) | 100.00 (100.00–100.00) | 97.50 (96.16–98.84) | 0.997 (0.994–1.000) | |
| Hand, foot and mouth disease | 93.55 (91.44–95.66) | 98.36 (97.27–99.45) | 97.50 (96.16–98.84) | 0.987 (0.977–0.997) | |
| Dengue fever | 48.15 (43.86–52.44) | 96.35 (94.74–97.96) | 93.85 (91.79–95.91) | 0.613 (0.438–0.788) | |
| Measles | 96.67 (95.13–98.21) | 96.73 (95.20–98.26) | 96.73 (95.20–98.26) | 0.990 (0.983–0.997) | |
| H1N1 influenza | 65.79 (61.71–69.87) | 98.34 (97.24–99.44) | 95.96 (94.27–97.65) | 0.982 (0.969–0.995) | |

The 95% confidence intervals were given in parentheses.

AUC = area under the receiver operating characteristic curve.

symptoms and signs, abnormal lab test results, epidemiological features, and incidence rates. By using standard validation methods, we affirmed that the model had good discrimination performance.

Bayesian approach is widely adopted in epidemiology and clinical studies on developing discrimination or diagnostic models, due to its adequate capability in classifying multiple categories and suitability for large databases.[11] The infectious disease diagnosis module contained in the well-known Global Infectious Disease and Epidemiology Network (GIDEON) was developed based on Bayesian formula.[5,6] An evaluation study of GIDEON showed the accuracy was 64% of the 129 fevers with infectious etiology.[18] Another study indicated the correct diagnoses ranked first for 52% when diagnosing febrile illnesses in Japanese returning travelers.[19] Although the accuracy of GIDEON is acceptable, better predictive performance is needed. Furthermore, the data of symptoms and signs in GIDEON does not perfectly match those of the Chinese population, limiting its application in China. Therefore, some researchers tried to construct classification models especially for the Chinese population.[20–22] Unfortunately, the predictive accuracies were undesirable. Moreover, the models designed in those studies were not validated appropriately. Most of them did not conduct ROC-AUC analysis and provided poor statistical description (e.g., lack of confidence intervals). In our study, the model was established based on data from the Chinese population, which are permitted for use in field investigation of infectious disease outbreaks in China. By using standard validation methods, the model presented relatively excellent discrimination performance.

The major problem in developing an infectious disease diagnosis program is difficulty in obtaining reliable and accurate individual level training data. For the majority of statistical approaches, it is essential to acquire adequate sample size of individual level data within each category or disease. Nevertheless, it is unrealistic particularly when there are many outcome categories or diseases taken into consideration in modeling process. Fortunately, naïve Bayesian algorithm can overcome this challenge, in which the aggregated data instead of individual level data is sufficient for modeling. In our study, the conditional probabilities of symptoms and signs as well as abnormal lab test results, were estimated based on the frequencies of corresponding predictors derived from a certain amount of literature. Each literature of an epidemic or outbreak investigation possessed a certain amount of patients. Consequently, it could be assumed that a large enough sample size has been achieved within each disease (see Table 1).

To select the most likely diagnoses among the multiplicity of possible diseases is another major challenge for modeling,[18] by the fact that some symptoms and signs are quite similar among diseases affected same organ systems. The majority of diseases could be correctly discriminated in our model. Whereas, the sensitivity is below 50% for hepatitis E and dengue fever, although the sensitivity cannot entirely reflect the discrimination performance. Patients with hepatitis E may present with few clinical features. In our validation dataset, more than one third of patients with hepatitis E (7/18) were asymptomatic, who were discriminated with pulmonary tuberculosis as 1st ranking by the model. It is worth noting that the correct diagnosis of hepatitis E retained in top 3 ranking for all these patients. Furthermore, symptoms and signs of dengue fever are partially nonspecific, resulting in other 6 diseases ranked 1st which were actually incorrect. The correct diagnosis of dengue fever appeared in top 3 ranking for 59.26% (15/28). According to our results, we think the output list of diseases ranking is helpful for users. Many previous studies[18,19,23] used the correct diagnoses appeared on the differential diagnosis lists or in the top 5 ranking as arbitrary indicators for evaluation. The validation results of 1st ranking performance in our model seem somewhat better than those using more tolerant indicators in previous studies. Besides the sensitivity, other parameters for validation demonstrated well discriminative capability of the model.

There are several advantages in this study.

(1) The quantitative results are provided on the output list of model. All possible diseases can be listed and ranked from the highest probability to the lowest probability. Existing medical decision-support programs are often inadequate in achieving a match to the most likely diagnosis.[24] It was suggested that a given disease was usually retained in the top 5 ranking when its probability exceeded 1%.[18] As a consequence, the list of predicted diagnoses is valuable in reminding users of alternative diseases that might otherwise have been ignored.[25] We have further assessed the top 3 ranking performance of our model, in which correct diagnosis in top 3 ranking was treated as correct discrimination. It was found that the sensitivity achieved 100% in 7 of 12 diseases, and also increased in rest 5 diseases than that using 1st ranking as correct discrimination before.
(2) Various types of information were utilized as predictors to discriminate the causative disease, including the incidence rates and hundreds of symptoms and signs, abnormal lab test results, and epidemiological features. Bayesian algorithm used in our study is suitable for such a large database with plenty of predictors. Meanwhile, by incorporating prior information on disease incidence, Bayesian classifiers have the potential to estimate disease probability better than other common machine-learning methods.[26]
(3) Data on incidence rates and epidemiological features was collected from a national surveillance system of infectious disease,[13] which guaranteed the data quality. In addition, the method for obtaining conditional probabilities ensured the enough sample size and adequacy for modelling.
(4) Standard statistical methods are utilized to validate the discrimination performance of the model, encompassing sensitivity, specificity, total accuracy, AUC and M-index. Seeing that 1-vs-rest measure for calculating former four parameters may be dominated by highly prevalent categories in the rest group,[17] M-index was calculated as an alternative measure in this study. Both of two measures demonstrated that our model gained a notably high level of discriminative ability across multiple infectious diseases.

Several methodological issues and limitations need to be mentioned. First, validation dataset involved the cases of 12 diseases only, due to the limited data resources in individual level we finally collected. Therefore, the discrimination performance was not able to be evaluated among other diseases, although the validation results were satisfactory in current 12 diseases. So we expect more validation data of other diseases. Second, the model is limited to discriminate infectious diseases already included in the database. Data on other diseases can be included to extend the application range of the model in future. Third, real-time updates of incidence rates should be carried out in future uses of the model. Meanwhile, data on conditional probabilities also needs to be updated regularly from information in latest literature.

Forth, the conditional probabilities of symptoms and signs as well as abnormal lab results of each disease differ between countries. Since the model was designed for application among Chinese population, only the data reported in Chinese literature was used. Moreover, data on incidence rates of Zhejiang province was used for modeling, and the application of model can be generalized nationwide if that of China was used instead. Fifth, the prior probabilities (incidence rates) may vary during different stages of an outbreak, while our model assumed that the prior probability was constant in a specific event. Last, the data of conditional probabilities is collected from different sources, and the quality of literature may not always be perfect for all diseases. Nevertheless, the publishing process of the included literature at least guarantee the data quality adequate for model construction.

## 5. Conclusion

In this study, we constructed a classification model based on Bayesian classifier to discriminate common infectious diseases within Zhejiang province. After entering symptoms and signs, abnormal lab test results, epidemiological features and city of disease origin, the probabilities of diseases can be calculated and an output list of possible diseases ranked from the highest to the lowest probability can be provided. This model offers excellent discrimination performance, which is expected to be beneficial to epidemiological field investigators in determining the cause of an outbreak and to provide clues for laboratory pathogen screening.

## 6. Ethical approval and consent to participate

The data, including the incidence rates and epidemiologic features, was collected from CISDCP. It was exempt from the requirement for ethical approval and informed consent according to the Law of the People's Republic of China on Prevention and Treatment of Infectious diseases.

The data, including symptoms and signs, and abnormal lab test results, was collected from previous published literature. Thus no ethical approval and informed consent is required.

The data for model validation was collected from the investigation in response to public health emergency. As such, it was exempt from the requirement for ethical approval and informed consent according to the Law of the People's Republic of China on Prevention and Treatment of Infectious diseases. Personal details of patients was anonymized and de-identified prior to analysis.

## Acknowledgments

We wish to thank the investigators from Hangzhou Municipal Center for Disease Control and Prevention, Shaoxing Municipal Center for Disease Control and Prevention, Jinhua Municipal Center for Disease Control and Prevention for their investigation and data collection.

## Author contributions

**Conceptualization:** Yi Shen, Fan He, Zhen Wang.
**Data curation:** Biyao Liu.
**Formal analysis:** Fudong Li, Duo Lv.
**Funding acquisition:** Fudong Li, Zhen Wang.
**Investigation:** Junfen Lin, Biyao Liu.
**Methodology:** Fudong Li, Yi Shen.
**Project administration:** Zhen Wang.
**Resources:** Zhen Wang.
**Software:** Duo Lv.
**Supervision:** Yi Shen, Fan He, Zhen Wang.
**Validation:** Junfen Lin.
**Visualization:** Biyao Liu.
**Writing – original draft:** Fudong Li.
**Writing – review & editing:** Yi Shen, Junfen Lin, Zhen Wang.

## References

[1] National Health Commission of the People's Repablic of China2018 China Health Statistical Yearbook. Beijing: Peking Union Medical College Press; 2018.
[2] Gregg MB. Field epidemiology. 3rd ed.New York: Oxford University Press; 2008.
[3] Goodman RA, Buehler JW, Koplan JP. The epidemiologic field investigation: science and judgment in public health practice. Am J Epidemiol 1990;132:9–16.
[4] Duda RO, Hart PE, Stork DG. Pattern classification. 2nd ed.New York: John Wiley & Sons; 2001.
[5] Berger SA. GIDEON: a comprehensive Web-based resource for geographic medicine. Int J Health Geogr 2005;4:10.
[6] Edberg SC. Global Infectious Diseases and Epidemiology Network (GIDEON): a world wide Web-based program for diagnosis and informatics in infectious diseases. Clin Infect Dis V 40 2005;123–6.
[7] Feldman S, Klein DF, Honigfeld G. The reliability of a decision tree technique applied to psychiatric diagnosis. Biometrics 1972;28:831–40.
[8] Cualing H, Kothari R, Balachander T. Immunophenotypic diagnosis of acute leukemia by using decision tree induction. Lab Invest 1999;79:205–12.
[9] Shaw J. A decision tree approach to psychodiagnosis. The diagnosis of abnormal behaviour. Aust Fam Physician 1985;14:284–5. 8, 90.
[10] Baxt WG. Use of an artificial neural network for the diagnosis of myocardial infarction. Ann Inter Med 1991;115:843–8.
[11] Michie D, Spiegelhalter DJ, Taylor CC. Machine learning, neural and statistical classification. Hertfordshire, U.K.: Ellis Horwood; 1994.
[12] Han J, Kamber M, Pei J. Data mining: Concepts and techniques. 3rd ed. Waltham, MA, USA: Morgan Kaufmann; 2012.
[13] Wang L, Wang Y, Jin S, et al. Emergence and control of infectious diseases in China. Lancet 2008;372:1598–605.
[14] Lalkhen AG, McCluskey A. Clinical tests: sensitivity and specificity. Cont Educ Anaesth Crit Care Pain 2008;8:221–3.
[15] Provost F, Domingos P. Tree induction for probability-based ranking. Mach Learn 2003;52:199–215.
[16] Hand D, Till R. A Simple generalisation of the area under the ROC curve for multiple class classification problems. Mach Learn 2001;45:171–86.
[17] Van Calster B, Vergouwe Y, Looman CW, et al. Assessing the discriminative ability of risk models for more than two outcome categories. Eur J Epidemiol 2012;27:761–70.
[18] Bottieau E, Moreira J, Clerinx J, et al. Evaluation of the GIDEON expert computer program for the diagnosis of imported febrile illnesses. Med Decis Making 2008;28:435–42.
[19] Kimura M, Sakamoto M, Adachi T, et al. Diagnosis of febrile illnesses in returned travelers using the PC software GIDEON. Travel Med Infect Dis 2005;3:157–60.
[20] Li B. Intelligent diagnosis system of disease based on Delphi (In Chinese). Comput Technol Dev 2010;20:250–2.
[21] Wang L-G, Dong S-C, Hao R-Z, et al. Auxiliary diagnosis system of infectious diseases and its application (In Chinese). Chin J Public Health 2010;26:1491–2.
[22] Hu B-S, Feng D, Cao W-C, et al. Mobile intelligent disease diagnosis system based on Bayesian analysis (In Chinese). J Comput Appl 2008;28 (B06):15–7.
[23] Luo RF, Bartlett JG. Use of the computer program GIDEON at an inpatient infectious diseases consultation service. Clin Infect Dis 2006;42:157–8.
[24] Berner ES. Diagnostic decision support systems: how to determine the gold standard? J Am Med Inform Assoc 2003;10:608–10.
[25] Kassirer JP. A report card on computer-assisted diagnosis–the grade: C. N Engl J Med 1994;330:1824–5.
[26] Gao X, Lin H, Dong Q. A dirichlet-multinomial bayes classifier for disease diagnosis with microbial compositions. mSphere 2017;2: e00536–617.