Machine Learning Algorithms for Important Feature Evaluation and Prediction of Severe Hand-Foot-Mouth Disease in Hunan Province, China

Research-in-Progress

Xiaochi Liu **Yilan Liao**

Zhiyu Zhu

Abstract

Hand, foot, and mouth disease(HFMD) is an infectious disease of the intestines that damages people's health, severe cases could lead to cardiorespiratory failure or death. Therefore, the evaluation of important features and prediction for severe HFMD is critical for early prevention and control of the disease. With this goal in mind, 658,689 cases which include 6,579 severe cases were assessed. In this research-in-progress paper, we are trying to establish an easy, automatic and efficient server HFMD prediction system based on hospital case data and meteorological data, and Random Forests and Adaboost algorithm were utilized in this paper for feature importance evaluation. Preliminary experimental result shows that our model can evaluate the importance of features but parameters still need further adjustment for predictions of severe HFMD.

Keywords: Machine Learning, Random Forests, Adaboost, Feature Importance Evaluation, Sever HFMD Prediction

1 Introduction

Hand-foot-mouth disease (HFMD) is a common infectious disease caused by a group of enteroviruses, with enterovirus 71 (EV-71) and Coxsackie virus A16 (CA-V16) being the most prevalent in China (Takahashi et al. 2016; Xing et al. 2014). Over the last decade, outbreaks of HFMD that were associated with EV71 have been reported in countries in the Western Pacific Region, including Japan, Malaysia,

Singapore, and China (Xu et al. 2012; Yang et al. 2012). The cumulative total of the reported cases in China has reached approximately 1.7 million, 1.9 million, and 2.7 million in 2010, 2013, and 2014, respectively (Liu et al. 2015; Xu et al. 2011). The clinical manifestations of most HFMD cases were mild and limited to fever, rash, or herpes on hand, foot, and mouth (Secrest & Shah 2013). In general, mild infections are self-limited and not life-threatening, while severe HFMD are often associated with neurological and systemic complications, such as aseptic meningitis, brainstem encephalitis, acute flaccid paralysis, myocarditis and pulmonary edemas that requires hospitalization, or even causing death (Li et al. 2014; Xing et al. 2014). Unfortunately, the incidence of severe HFMD in mainland China is high.

Outbreaks of HFMD have been associated with two viruses belonging to the species Enterovirus A: Enterovirus 71 (EV71) and Coxsackie virus A16 (CoxA16). While the prevalence of EV71-induced HFMD has increased in recent years, a vaccine or effective treatment method for EV71 or CoxA16 remains elusive. Thus, epidemiological surveillance of HFMD is important for the development of public-health interventions that prevent outbreaks. Meteorological factors may also play a key role in epidemic outbreaks and seasonal HFMD activity. Several reports have suggested an association between meteorological factors and HFMD, but the findings have been inconsistent. Temperature and relative humidity were significantly associated with HFMD infection in children. In southern China, each 1 ℃ increase in temperature and relative humidity led to a 1.42 % and 1.86 % increase, respectively, in the weekly number of hospitalized cases reported (Huang et al. 2013). In Singapore, each 1 ℃ increase in maximum temperature above 32 ℃ increased the risk of HFMD incidence by 36 %, while a 1-mm increase in weekly cumulative rainfall up to 75 mm increased the risk of HFMD by 0.3 % (Hii et al. 2011). In contrast, an average temperature above 25 ℃ was negatively associated with the incidence of HFMD in Tokyo, Japan (Urashima et al. 2003). In northern China, only atmospheric temperatures were positively associated with HFMD activity using partial correlations (Feng et al. 2014). Beside temperature and humidity, high wind velocity was also associated with HFMD consultation rates in Hong Kong (Ma et al. 2010). Consequently, in different countries/geographic regions, meteorological factors may have varying impacts on HFMD activity.

Evidences from global reports on HFMD epidemics have substantiated that the incidence of severe HFMD is elevating gradually, along with mortality rate (Solomon et al. 2010). Thus, identifying potential early indicators for severe HFMD is essential, which enable early medical interventions and alleviating the disease severity, subsequently reducing the mortality rate. Hence, the aim of this study was to establish prediction system for the occurrence of severe HFMD based on hospital cases data and metrological data, evaluating the important features using machine learning algorithms.

2 Data Preparation

2.1 Study Area

In this study, Hunan, a province in south central China, was selected as the study area. Hunan is located south of the middle course of the Yangtze River and south of Lake Dongting, situated between the 108°47′–114°16′ east longitudes and the 24°38′–30°08′ north latitudes. It covers an area of 211,800 square kilometers and is divided into 14 cities. Mountains and hills occupy more than 80% of the area, with plains comprising less than 20% of the whole province. Hunan's climate is subtropical, where sunshine and rainfall are both abundant concentrated. In spring, the temperature would fluctuate in a wide range, whereas in summer and autumn, it is pretty hot and humid, and in winter it's usually cool and damp. According to the meteorological data, the annual average temperature of Hunan is around 16–19 ℃ and the average annual precipitation is 1200–1700 mm and has an uneven distribution

In terms of the socioeconomic condition, according to the statistical yearbook of Hunan province, during the study period, there were around 66,380,000 people living in Hunan with a GDP of 32903.71

RMB per capita, and the people living in cities accounted for 46.65% of the total population. Since its geographical environment advantages and rich natural resources, the development of Hunan's economy is good.

During 2010–2014, 658,689 HFMD cases were reported in Hunan. The 5-year prevalence reached 167.83 per 100,000 persons. There were 6,579 severe cases (1.03% of all cases). There were 296 cases of death, a mortality rate of 0.05%. The annual HFMD prevalence varied significantly over these 5 years, ranging from 287.32 (in 2012) to 54.31 per 100,000 persons (in 2009).

2.2 Data Collection

The study unit was each HFMD patient in Hunan province, China, during the period of January 1 2010, to December 31. The HFMD case data were provided by the Hunan Center for Disease Control and Prevention (Hunan CDC), which include each patient's gender, age, residential address, job, the date of illness, the date of admitting to the hospital, and whether it is a sever case (ICU), totally 658,689 cases with 6,579 sever cases.

In accordance with previous studies (Gao et al. 2014; Yang et al. 2015), the meteorological data in this study mainly included 6 variables, which was daily precipitation, daily average air pressure, daily average relative humidity, daily sunshine hours, daily average temperature, and daily average wind speed. The meteorological data of all cities were obtained from the China Meteorological Data Sharing Service System and were interpolated by the inverse distance weighted interpolation method, based on the data of the weather stations in Hunan province.

2.2 Data Preprocessing

Geocoding is the computational process of transforming a physical address description to a location on the Earth's surface (spatial representation in numerical coordinates). Since each HFMD patient's residential address information were included in the hospital case data, we can geocode the address into latitude and longitude format under a specific geographic coordinate. Then, based on each patient's location, we matched the meteorological data of seven days prior to the date of each patient getting HFMD. In order to get better classification results, we discretized the continuous meteorological data utilizing Entropy-MDL method, invented by Fayyad and Irani, which uses mutual information to recursively define the best bins, and we also completed the feature engineering process, which implements what is called one-of-K or 'one-hot' coding for categorical (aka nominal, discrete) features. Next, we define the variable of ICU (whether the patient was in intensive care unit, which means a sever HFMD case) as the response variable, and utilized Random Forests model and Adaboost Random Forests model as a comparison to evaluate the importance of all explanatory variables (features). Based on these important features, we can establish a more sensitive and accurate prediction system for HFMD sever cases. The whole data processing was shown in Figure 1.

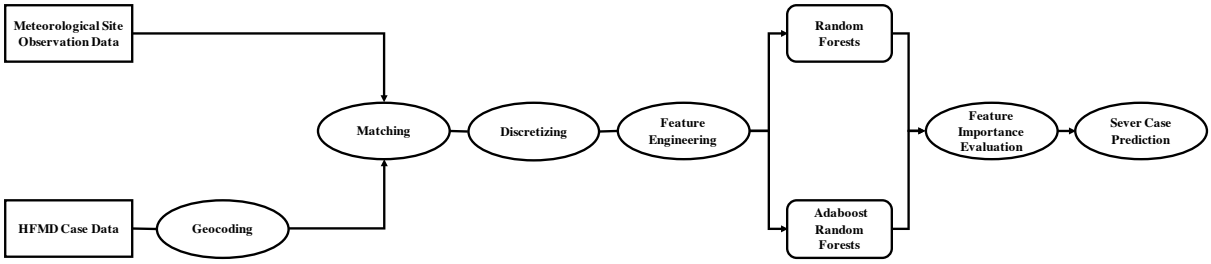

Figure 1. Data Preprocessing Method

3 Methodology

3.1 Methodology

The goal of ensemble methods is to combine the predictions of several base estimators built with a given learning algorithm in order to improve generalizability / robustness over a single estimator. There are two families usually distinguished: averaging methods and boosting methods. In averaging methods, the driving principle is to build several estimators independently and then to average their predictions. On average, the combined estimator is usually better than any of the single base estimator because its variance is reduced. For example: Random Forests algorithm. By contrast, in boosting methods, base estimators are built sequentially and one tries to reduce the bias of the combined estimator. The motivation is to combine several weak models to produce a powerful ensemble, for example: AdaBoost Random Forests algorithm. Both algorithms are based on randomized decision trees.

3.1 Decision Trees

Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

Given training vectors $x_i \in R^n, i = 1, \ldots, l$ and a label vector $y \in R^l$, a decision tree recursively partitions the space such that the samples with the same labels are grouped together. Let the data at node m be represented by Q. For each candidate split $\theta = (j, t_m)$ consisting of a feature j and threshold t_m, partition the data into $Q_{left}(\theta)$ and $Q_{right}(\theta)$ subsets

$$Q_{left}(\theta) = (x, y)|x_j \leq t_m$$

$$Q_{right}(\theta) = Q \backslash Q_{left}(\theta)$$

The impurity at m is computed using an impurity function $H()$, the choice of which depends on the task being solved (classification or regression)

$$G(Q, \theta) = \frac{n_{left}}{N_m} H\left(Q_{left}(\theta)\right) + \frac{n_{right}}{N_m} H(Q_{right}(\theta))$$

Select the parameters that minimizes the impurity

$$\theta^* = argmin_\theta G(Q, \theta)$$

Recurse for subsets $Q_{left}(\theta^*)$ and $Q_{right}(\theta^*)$ until the maximum allowable depth is reached, $N_m < min_{samples}$ or $N_m = 1$.

If a target is a classification outcome taking on values $0,1, \ldots K-1$, for node m, representing a region R_m with N_m observations, let

$$p_{mk} = 1/N_m \sum_{x_i \in R_m} I(y_i = k)$$

be the proportion of class k observations in node m

Common measures of impurity are Gini

$$H(X_m) = \sum_k p_{mk}(1 - p_{mk})$$

Entropy

$$H(X_m) = -\sum_k p_{mk} \log (p_{mk})$$

and Misclassification

$$H(X_m) = 1 - \max (p_{mk})$$

where X_m is the training data in node m.

3.2 Random Forests

In random forests, each tree in the ensemble is built from a sample drawn with replacement (i.e., a bootstrap sample) from the training set. In addition, when splitting a node during the construction of the tree, the split that is chosen is no longer the best split among all features. Instead, the split that is picked is the best split among a random subset of the features. As a result of this randomness, the bias of the forest usually slightly increases (with respect to the bias of a single non-random tree) but, due to averaging, its variance also decreases, usually more than compensating for the increase in bias, hence yielding an overall better model.

3.3 AdaBoost Algorithm

The core principle of AdaBoost is to fit a sequence of weak learners (i.e., models that are only slightly better than random guessing, such as small decision trees) on repeatedly modified versions of the data. The predictions from all of them are then combined through a weighted majority vote (or sum) to produce the final prediction. The data modifications at each so-called boosting iteration consist of applying weights, $w_1, w_2, \ldots w_N$ to each of the training samples. Initially, those weights are all set to $w_i = 1/N$, so that the first step simply trains a weak learner on the original data. For each successive iteration, the sample weights are individually modified and the learning algorithm is reapplied to the reweighted data. At a given step, those training examples that were incorrectly predicted by the boosted model induced at the previous step have their weights increased, whereas the weights are decreased for those that were predicted correctly. As iterations proceed, examples that are difficult to predict receive ever-increasing influence. Each subsequent weak learner is thereby forced to concentrate on the examples that are missed by the previous ones in the sequence.

3.4 Feature Importance Evaluation

The relative rank of a feature used as a decision node in a tree can be used to assess the relative importance of that feature with respect to the predictability of the target variable. Features used at the top of the tree contribute to the final prediction decision of a larger fraction of the input samples. The expected fraction of the samples they contribute to can thus be used as an estimate of the relative importance of the features. In scikit-learn, the fraction of samples a feature contributes to is combined with the decrease in impurity from splitting them to create a normalized estimate of the predictive power of that feature. Meanwhile, by averaging the estimates of predictive ability over several randomized trees one can reduce the variance of such an estimate and use it for feature selection.

4 Preliminary Result

4.1 Parameters and Results of Random Forests Algorithm

The main parameters to adjust when using these methods is the number of trees in the forest (n_estimators) and the number of features to consider when looking for the best split of a node (max_features). The larger of the number of trees are better, but also the longer it will take to compute. In addition, note that results will stop getting significantly better beyond a critical number of trees. Usually, n_estimators=1000 is sufficient. The other one, max_features is the size of the random subsets of features to consider when splitting a node. The lower the greater the reduction of variance, but also the greater the increase in bias. Empirical good values for classification tasks are max_features =

$\sqrt{n\_features}$, where n_features is the number of features in the data. Besides, good results are often achieved when the max depth of the trees is set to None, in combination with the minimum samples to split a node is 2 (i.e., when fully developing the trees).

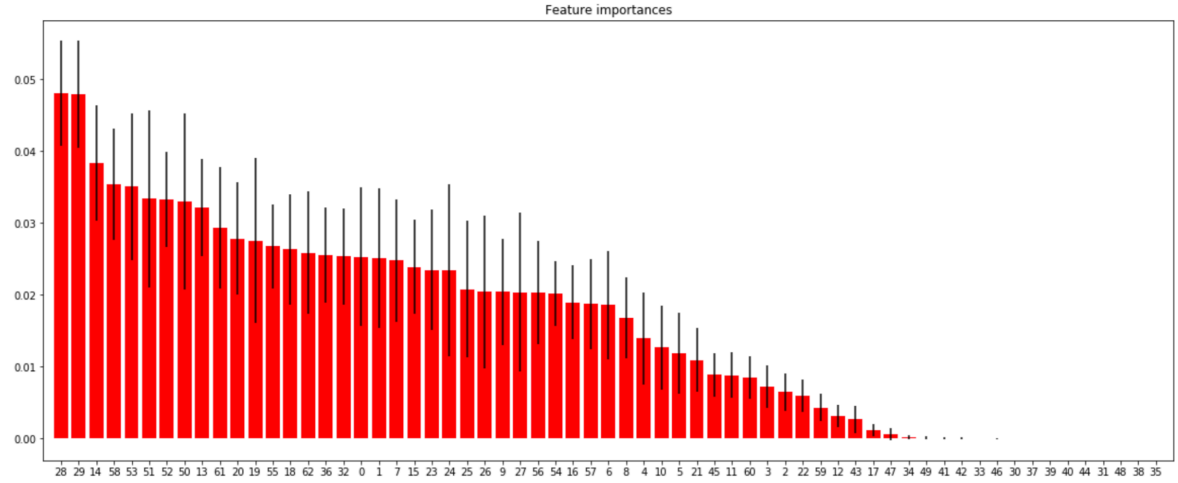

Figure 2. Feature Importance Evaluation by Random Forests Algorithm

Figure 2 shows each feature's importance evaluated by Random Forests Algorithm, and from left to right, the importance is reduced in turn. The feature 28, standing for "Gender = Female', was ranked as the most important feature for the response variable ICU (server HFMD cases), which means that it was more likely to get sever HFMD if a patient's gender is female. Similarly, we can extract top 5, top 10, or other numbers of important features for predicting sever HFMD. All of features are categorical and falling in the specific interval.

4.2 Parameters and Results of Adaboost Algorithm

The number of weak learners is controlled by the number of estimators. The learning rate parameter controls the contribution of the weak learners in the final combination. By default, weak learners are decision stumps. Different weak learners can be specified through the base estimator parameter. The main parameters to tune to obtain good results are the number of estimators and the complexity of the base estimators (e.g., its depth or minimum required number of samples to consider when splitting a node).

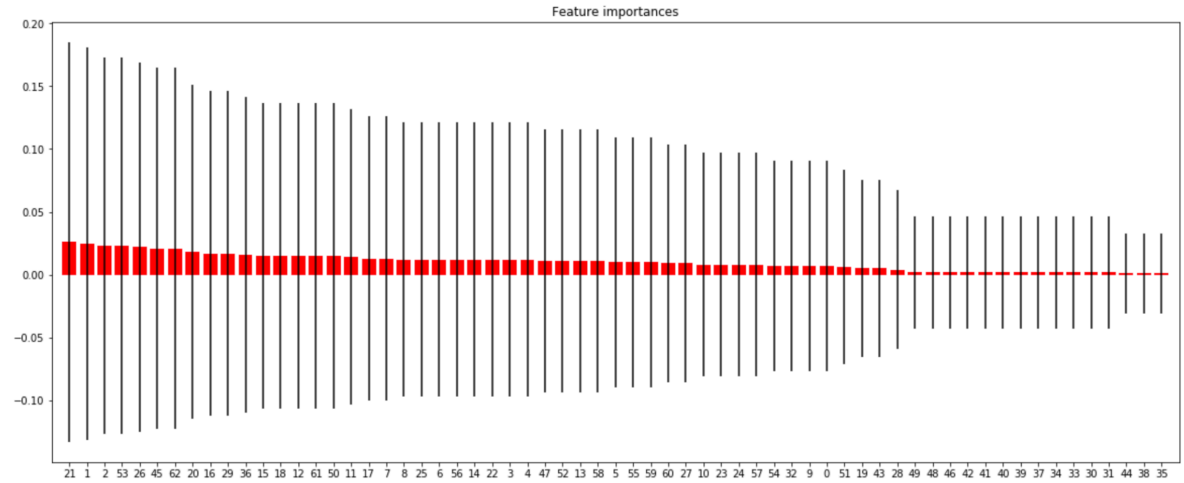

Figure 3. Feature Importance Evaluation by Adaboost Algorithm

Figure 3 shows each feature's importance evaluated by Adaboost Random Forests Algorithm, and from left to right, the importance is reduced in turn. The feature 21, standing for "daily average temperature ≤ 12 ℃" , was ranked as the most important feature for server HFMD cases, which means that it was more likely to get sever HFMD if the seven days' average temperature was less than 12 ℃. Similarly, we can also extract top 5, top 10 or other numbers of important features for predicting sever HFMD.

4.3 Parameters Adjustment

ROC_AUC score was utlized to evaluate classifier output quality. A receiver operating characteristic (ROC), or simply ROC curve, is a graphical plot which illustrates the performance of a binary classifier system as its discrimination threshold is varied. It is created by plotting the fraction of true positives out of the positives (TPR = true positive rate) vs. the fraction of false positives out of the negatives (FPR = false positive rate), at various threshold settings. TPR is also known as sensitivity, and FPR is one minus the specificity or true negative rate. ROC curves typically feature true positive rate on the Y axis, and false positive rate on the X axis. This means that the top left corner of the plot is the "ideal" point - a false positive rate of zero, and a true positive rate of one. This is not very realistic, but it does mean that a larger area under the curve (AUC) is usually better.

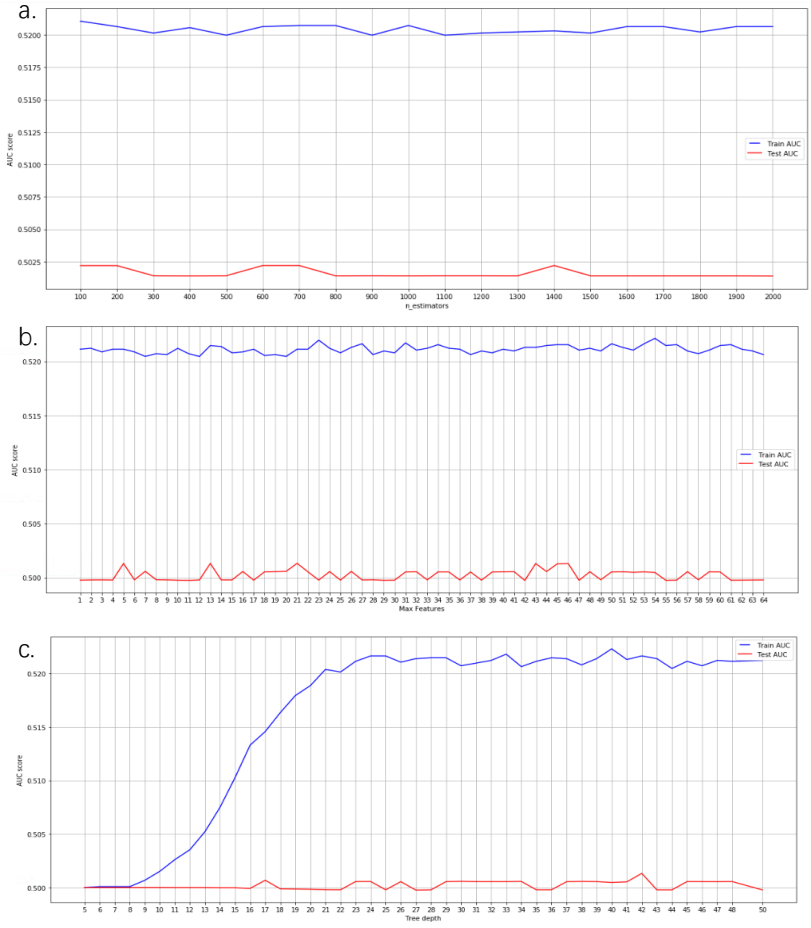

Figure 4. AUC Scores of Different Parameters Selection

According to Figure 4, which shows AUC scores of models set by different parameters, it can be found that there was small influence on models' AUC score by setting the number of estimators and the number of features when looking for the best split of a node, although some values could lead a relatively high AUC such as 100, 200, 600, 700, 1400 of estimators, and 5, 13, 21, 43 of features. However, setting different tree depth had a significant influence on model's performance. In this case, 24, 26, and 42 of tree depth had better predicting results.

Preliminary Conclusion

In this research-in-progress paper, we proposed utilizing ensemble methods, such as Random Forest algorithm and Adaboost algorithm, to established a real-time, automatic and efficient prediction tool for severe hand, foot, and mouth disease (HFMD) prediction. Our model extracted important features directly from HFMD case data, which is convenient to be wildly used in the real situation, and further parameters adjustment is very important for the prediction of severe HFMD. Once the important features have been extracted, a more sensitive and accurate warning system for sever HFMD prevalence can be established.

References

Feng H, Duan G, Zhang R et al. 2014. "Time series analysis of hand-foot-mouth disease hospitalization in Zhengzhou: establishment of forecasting models using climate variables as predictors," PLoS ONE 9:e87916.

Hii YL, Rocklo¨v J, Ng N. 2011. "Short term effects of weather on hand, foot and mouth disease," PLoS ONE 6:e16796.

Huang Y, Deng T, Yu S et al. 2013. "Effect of meteorological variables on the incidence of hand, foot, and mouth disease in children: a time-series analysis in Guangzhou, China," BMC Infect Dis 13:134.

Li, W. et al. 2014. "Study on risk factors for severe hand, foot and mouth disease in China," PLoS One 9(1), e87603.

Liu, S. L. et al. 2015. "Comparative epidemiology and virology of fatal and nonfatal cases of hand, foot and mouth disease in mainland China from 2008 to 2014," Rev Med Virol 25(2), 115.

Ma E, Lam T, Wong C et al. 2010. "Is hand, foot and mouth disease associated with meteorological parameters?" Epidemiol Infect 138:1779–1788

Solomon, T. et al. 2010. "Virology, epidemiology, pathogenesis, and control of enterovirus 71," Lancet Infect Dis 10(11), 778.

Secrest, A. M. & Shah, K. N. 2013. "Hand-foot-and-mouth disease," JAMA Pediatr 167(4), 387 (2013).

Takahashi, S. et al. 2016. "Hand, Foot, and Mouth Disease in China: Modeling Epidemic Dynamics of Enterovirus Serotypes and Implications for Vaccination," PLoS Med 13(2), e1001958.

Urashima M, Shindo N, Okabe N. 2003. "Seasonal models of herpangina and hand-foot-mouth disease to simulate annual fluctuations in urban warming in Tokyo," Jpn J Infect Dis 56:48–53.

Wang, Y. et al. 2011. "Hand, foot, and mouth disease in China: patterns of spread and transmissibility," Epidemiology 22(6), 781.

Xu, W. et al. 2012. "Distribution of enteroviruses in hospitalized children with hand, foot and mouth disease and relationship between pathogens and nervous system complications," Virol J 9, 8.

Xing, W. et al. 2014. "Hand, foot, and mouth disease in China, 2008-12: an epidemiological study," Lancet Infect Dis 14(4), 308.

Yang, T. et al. 2012. "A case-control study of risk factors for severe hand-foot-mouth disease among children in Ningbo, China, 2010–2011," Eur J Pediatr 171(9), 1359.