The Journal of Veterinary Medical Science

FULL PAPER

Public Health

Evaluation of multinomial logistic regression models for predicting causative pathogens of food poisoning cases

Hideya INOUE[1,2], Tomoyuki SUZUKI[2], Masashi HYODO[3] and Masami MIYAKE[1]*

[1]Department of Veterinary Science, Graduate School of Life and Environmental Sciences, Osaka Prefecture University, Izumisano, Osaka 598-8531, Japan
[2]Shiga Prefectural Institute of Public Health, 13-45 Gotenhama, Otsu, Shiga 520-0834, Japan
[3]Department of Mathematical Sciences, Graduate School of Engineering, Osaka Prefecture University, Sakai, Osaka 599-8531, Japan

ABSTRACT. In cases of food poisoning, it is important for food sanitation inspectors to determine the causative pathogen as early as possible and take necessary measures to minimize outbreaks. Interviews are usually conducted to obtain epidemiological information to aid in the rapid determination of the cause. However, the current method of determining the causative pathogen has the disadvantage of being reliant upon the experience and knowledge of food sanitation inspectors. Here, we analyzed 529 infectious food poisoning incidents reported in five municipalities in the Kinki region to develop a tool for evaluation using a multinomial logistic regression model, which can predict the causative pathogen based on the patients' epidemiological information. This tool predicts the most probable cause of the incident by generating a list of pathogens with the highest probability. As a result of leave-one-out cross validation, the agreement ratio with the actual pathogen was 86.4%, and this ratio increased to 97.5% when the agreement was judged by including the true pathogen within the top three pathogens with the highest probability. In cases where the difference of probability between the first and second candidate pathogen was ≥50%, the agreement ratio increased to 94.2%. Using this tool, it is possible to accurately estimate the causative pathogen at an early stage based on patient information, and this will further help narrow the target of investigations to identify causative agent, thereby leading to a prompt identification, which can prevent the spread of food poisoning.

KEY WORDS: causative pathogen, food poisoning, leave-one-out cross validation, multinomial logistic regression

J. Vet. Med. Sci.
80(8): 1223–1227, 2018
doi: 10.1292/jvms.17-0653

Received: 4 December 2017
Accepted: 27 May 2018
Published online in J-STAGE: 11 June 2018

Foodborne pathogens have been reported to cause detrimental human diseases all over the world [3]. Even in the developed countries with good hygienic conditions, foodborne illnesses still result in a significant economic burden due to employee absenteeism, resources needed for treatment, and hospitalization and can sometimes be life-threatening [7, 16]. In Japan, an estimated 20,000 domestically acquired infectious foodborne illnesses associated with over 20 known pathogens are reported each year [19]. Upon the occurrence of a foodborne illness incident, food sanitation inspectors conduct an investigation to determine the causative foodborne pathogen. Investigation usually begins with an in-person interview with patients to collect epidemiological information, such as symptoms and their frequency, dietary history, and clinical outcomes.

Although the clinical features of most foodborne illnesses are non-specific for individual cases, outbreaks frequently have characteristic features with respect to the length of incubation period, duration of symptoms, and percentage of cases wherein patients experience specific signs and symptoms depending on their etiology [6, 14]. For example, the incubation period of *Staphylococcus aureus* (Sa)-induced food poisoning is relatively shorter compared with that of other pathogens (between 30 min and 4 hr), with the major clinical symptoms being vomiting and diarrhea. *Norovirus* (noro) causes similar symptoms, but its incubation time is relatively longer compared with that of other pathogens (between 24 and 48 hr). Campylobacteriosis shows different symptoms such as abdominal pain, cramping, and high fever in different cases [1].

Allowing the food sanitation inspectors to focus on the investigation before results of cultures are available will enable the investigation to be completed more rapidly, with a greater chance of successfully identifying the underlying cause of the outbreak. For outbreaks in commercial establishments, such as restaurants, the prompt identification of causative agent will prevent

*Correspondence to: Miyake, M.: mami@vet.osakafu-u.ac.jp

Table 1. Summary of the incubation periods and the incidences of each symptoms observed in the foodborne incidents used in this study

Causatives	Cam	Noro	Vp	Sal	Sa	Cp	DEC	Ah	Bc
Number of cases	139	138	112	73	44	11	10	1	1
Incubation period (hr)[a)]	62.3 ± 13.5	36.9 ± 5.6	17.4 ± 2.9	31.6 ± 14.7	5.0 ± 2.9	12.5 ± 1.8	31.4 ± 17.4	13	3.63
Diarrhea (%)[a)]	95.4 ± 10.2	74.2 ± 15.7	96.7 ± 5.8	94.3 ± 7.8	71.1 ± 21.4	92.1 ± 15.4	87.0 ± 26.9	83	37
Fever (%)[a)]	63.2 ± 20.6	54.6 ± 18.2	38.5 ± 18.0	69.1 ± 22.3	26.3 ± 21.0	6.2 ± 12.3	36.9 ± 16.9	50	3
Vomiting (%)[a)]	9.2 ± 13.0	59.6 ± 17.3	44.2 ± 18.9	19.2 ± 14.5	70.8 ± 25.5	1.8 ± 1.6	19.7 ± 17.1	67	56
Abdominal pain (%)[a)]	81.8 ± 16.1	55.6 ± 18.2	84.3 ± 18.9	73.3 ± 17.5	57 ± 23.5	60.3 ± 24.2	77.6 ± 19.8	67	16

a) The incubation periods and the incidence of each symptoms are expressed as means ± standard deviation. Vp, *Vibrio parahaemolyticus*; Noro, Norovirus; Sal, non-typhoidial *Salmonella*; Sa, *Staphylococcus aureus*; Cam, *Campylobacter*; DEC, Diarrheagenic *Escherichia coli*; Cp, *Clostridium perfringens*; Ah, *Aeromonas hydrophila*; Bc, *Bacillus cereus*.

additional incidents of illness and help develop better prevention strategies [5, 12, 13]. For example, noro has a human reservoir and does not replicate in food or the environment. Thus, rapidly identifying noro as a probable causative pathogen would allow the food sanitation inspectors to focus on issues pertaining to the health and hygiene of food workers and the potential ongoing transmission through these workers [8]. However, food sanitation inspectors currently infer the cause of food poisoning based on experience and the patient interview; therefore, inexperienced inspectors may not be able to rapidly suspect the cause of an outbreak, leading to dither and delay in the prompt initial action.

Although the usefulness of the clinical data for the prediction of causatives has been proposed [8], to the best of our knowledge, there is no report describing the mathematical evaluation of this interrelation and the development of a tool by multinomial logistic regression models. Here, we describe an algorithm to link the relation and the evaluation of the algorithm by leave-one-out cross validation (LOOCV). The purpose of this research was to predict causative pathogens primarily from clinical information collected from patient interviews at the early stage during the course of foodborne diseases.

MATERIALS AND METHODS

The data in this study included food poisoning incidents that occurred from 1950 to 2016 in Shiga prefecture and cases published between 2005 and 2016 in Osaka prefecture, Osaka City, Kyoto prefecture, and Nara prefecture (data are available from the National Epidemiological Surveillance of Foodborne Disease at the following URL only to public authorities of Japanese central and local governments: http://www.mhlw.go.jp/topics/bukyoku/iyaku/syoku-anzen/nesfd/index.html, accessed on June 2, 2018). Among the data, 529 cases of food poisoning incidents caused by pathogens, with five or more patients in each case, for which all epidemiological data used in this study were available, were selected [8]. The data used in this study included six variables according to each food poisoning case: the name of the causative pathogen; the average incubation period; and the proportion of patients showing any of the following symptoms: diarrhea, fever, abdominal pain, and vomiting). The causative pathogens that were confirmed by the examination were grouped into nine categories of foodborne pathogens: *Vibrio parahaemolyticus* (Vp), noro, non-typhoidal *Salmonella* (Sal), *Staphylococcus aureus* (Sa), *Campylobacter* (Cam), diarrheagenic *Escherichia coli* (DEC), *Clostridium perfringens* (Cp), *Aeromonas hydrophila* (Ah) and *Bacillus cereus* (Bc). These pathogens were identified by the official procedures as instructed by Japanese government [18]. Due to the concern that a dataset with too few cases may skew results, pathogens with fewer than 25 cases each (DEC, Cp, Ah and Bc) were merged into the category "others" on the basis of the criterion proposed by other investigators [10, 11]. Consequently, six causative pathogens were considered in this study: Vp, noro, Sal, Sa, Cam and others.

The mean incubation period and the proportion of symptom appearance were calculated for each pathogen. A tool for predicting causative pathogens were established by applying multiple logistic regressions [4, 9]. We used foodborne pathogens as the dependent variable with using the "others" category as the reference category. The mean incubation periods and proportions of symptom appearance were used as the independent variables. Independent variables included in the final model were selected using a forward–backward selection method with Akaike Information Criterion (AIC) [2]. All analyzes were performed using R version 3.1.0 with the package VGLM [15, 17].

The evaluation of the classification rule was conducted using LOOCV. LOOCV is often used to estimate the generalization ability of a statistical classifier (i.e., the performance of previously unseen data). At each LOOCV iteration, one sample was selected as testing set and the rest of the samples were used as the training set. For the calculation of coefficients, training steps were only performed with the data in the training set, without using the data of the testing set. Since the data in this study comprised 529 incidents, a multinomial logistic regression analysis was performed 529 times and the agreement ratio was calculated. In this evaluation, three criteria for judgement were used; i) whether the true pathogen perfectly matched with the first rank prediction, ii) whether the true pathogen matched with the ones within the top three pathogens with highest probabilities, iii) whether the agreement ratios become increased when the case shows that the difference between the first and second predicted probabilities was ≥50%.

Table 2. Coefficient of dependent variables with food poisoning pathogens by multinomial logistic regression analysis[a)]

The casative agent (k)	Cam (β_2)	Noro (β_3)	Vp (β_4)	Sal (β_5)	Sa (β_6)
Incubation period	0.244[b)]	0.165[b)]	−0.114[b)]	0.101[b)]	−0.706[b)]
Diarrhea	0.03	−0.028	0.085[b)]	0.026	0.03
Fever	0.114[b)]	0.065[b)]	0.026	0.123[b)]	0.031
Vomiting	−0.015	0.126[b)]	0.078[b)]	−0.008	0.073[b)]
Abdominal pain	0.006	−0.033	0.017	−0.009	0.009
Intercept	−15.763[b)]	−5.933[c)]	−8.441[b)]	−8.955[b)]	2.803
AIC[d)]			490.96		
Log-likelihood			−215.48		
Residual deviance			430.96		

a) Category "others" (Cp, DEC, Ah and Bc) was considered as reference. b) $P<0.01$. c) $P<0.05$. d) Akaike's information criterion.

RESULTS

Regarding the proportion of symptoms observed among patients with foodborne poisoning caused by each pathogen, Vp showed the highest diarrhea and abdominal pain rates (Table 1). Cam showed the longest incubation period among the six categories. Sal showed high fever rate. Sa showed the shortest incubation period and the highest vomiting rate. In most cases, diarrhea was a common symptom, with the lowest incidence of 71.1% for Sa. These clinical features align with those described in the literature [1].

On the basis of the result of forward–backward method, incubation period, diarrhea, fever, vomiting, and abdominal pain were selected as independent variables in the final model. Using the vector of coefficients (β_κ), the conditional probability for causative pathogen "k" is given by the following equations:

$$\pi_1 = \frac{1}{1+\sum_{\kappa=2}^{\kappa}\exp(\eta_\kappa)} \quad (1)$$

$$\pi_\kappa = \frac{\exp(\eta_\kappa)}{1+\sum_{\kappa=2}^{\kappa}\exp(\eta_\kappa)} \quad (2)$$

Where π is the probability of predictive pathogens, k is the number of causative pathogens, and η is the risk ratio between pathogen k and pathogen 1 ("others" as the reference rank), which is given by the following formula:

$\eta_{ik}=\beta_{0k}+\beta_{1k}\times incubation\ period+\beta_{2k}\times diarrhea+\beta_{3k}\times fever+\beta_{4k}\times vomiting+\beta_{5k}\times abdominal\ pain$

Where, β_{0k} is intercept and β_{ik} is the coefficient of variable i for pathogen k. Intercepts and coefficients estimated by the multinomial logistic model for each pathogen are shown in the Table 2.

For example, supposing a scenario whereby the following food poisoning cases occurred: $\dot{x}$=(incubation period, diarrhea, fever, vomiting, and abdominal pain)=(7.6, 79.2, 25.0, 79.2, and 58.3), η is:

$\eta_{Cam}=\dot{x}\times\beta_{i,2}^{T}=-15.763+(0.244\times7.6)+(0.03\times79.2)+(0.114\times25.0)+(-0.015\times79.2)+(0.006\times58.3)=-9.521$

$\eta_{noro}=\dot{x}\times\beta_{i,3}^{T}=-5.933+(0.165\times7.6)+(-0.028\times79.2)+(0.065\times25.0)+(0.126\times79.2)+(-0.033\times58.3)=2.784$

$\eta_{Vp}=\dot{x}\times\beta_{i,4}^{T}=-8.441+(-0.114\times7.6)+(0.085\times79.2)+(0.026\times25.0)+(0.078\times79.2)+(0.017\times58.3)=5.243$

$\eta_{Sal}=\dot{x}\times\beta_{i,5}^{T}=-8.955+(0.101\times7.6)+(0.026\times79.2)+(0.123\times25.0)+(-0.008\times79.2)+(-0.009\times58.3)=-4.212$

$\eta_{Sa}=\dot{x}\times\beta_{i,6}^{T}=2.803+(-0.706\times7.6)+(0.03\times79.2)+(0.031\times25.0)+(0.073\times79.2)+(0.009\times58.3)=6.895$

Therefore, π of Sa is calculated by the equations (2):

π_{Sa}=exp (6.895)/(1 + exp (−9.521) + exp (2.784) + exp (5.243) + exp (−4.212) + exp (6.895))=0.827

Consequently, the probability of Sa is calculated as 82.7%.

As a result of LOOCV, the agreement ratio when the agreement was determined between the actual pathogen and that predicted by the model with the highest probability was 86.4% (Table 3). Among the pathogens, Noro showed the highest agreement ratio (94.2%). Noro, Vp, Sa and Cam showed agreement ratios over 90%, but Sal showed the lowest agreement ratio (68.5%) among five pathogens. When the agreement ratio was judged as the agreement between true pathogen and the top three pathogens predicted with the highest probability, the total agreement ratio increased to 97.5%. Of note, the agreement ratios for Vp and Sal by this criterion were 100 and 95.9% respectively. In addition, when looking at 433 cases where the difference in the predicted

Table 3. Estimated agreement ratio calculated by LOOCV

Causative pathogen	Number of cases	By 1st[a)] (%)	By 2nd[b)] (%)	By 3rd[c)] (%)	Number of cases when the differences between the first and second probabilities is over 50%	By 1st[d)] (%)
Cam	139	92.1	97.8	97.8	119	96.6
Noro	138	94.2	97.8	99.3	127	98.4
Vp	112	91.1	99.1	100.0	92	96.7
Sal	73	68.5	89.0	95.9	45	82.2
Sa	44	93.2	93.2	97.7	41	95.1
Others	23	26.1	60.9	78.3	9	33.3
Total	529	86.4	94.9	97.5	433	94.2

a) Estimated agreement ratio for the indicated pathogen that matched the first rank prediction. b) Estimated agreement ratio for the indicated pathogen that matched either the first or the second ranks of prediction. c) Estimated agreement ratio for the indicated pathogen that matched with the top three with highest probability. d) Estimated agreement ratio for the indicated pathogen that matched the first rank prediction, when the differences between the first and second probabilities is over 50%.

probability between the first and second pathogen was 50% or more, the agreement ratio for the most probable pathogen was 94.2%, and the agreement ratio for Sal also increased to 82.2%.

DISCUSSION

We performed a multinomial logistic regression to provide a tool to predict causative pathogens of food poisoning cases. This tool requires clinical information regarding cases, obtained from personal interviews during early stage of an incident, as datasets for prediction. The use of equation (1) and (2) immediately returns the prediction of causative pathogens with probability ranking. In order to test the accuracy of the model prediction, we conducted LOOCV with different criterion for the judgement of agreement. Estimated agreement ratio for the predicted pathogen that matched the first rank prediction was 86.4%. These results suggest that reliable prediction is possible using the incubation period and the proportion of patient symptoms, as previously reported [8]. Since there was an agreement ratio of 94.2% when the difference in probability between the first and second pathogen was 50% or more, in such a case, food sanitation inspectors should mainly investigate predicted pathogen of the first probability. In case where the difference between the first and second probabilities was less than 50%, food sanitation inspectors should at least investigate the predicted pathogens by the third probability. This prediction would help food sanitation inspectors to narrow the target of investigations to identify causative pathogen, thereby leading to a prompt identification, which can prevent the spread of food poisoning.

In this study, DEC, Cp, Ah and Bc were merged into the category “others”. This approach was considered reasonable according to the literatures [10, 11]. On the other hand, this was considered necessary to avoid the exclusions a case whose clinical parameters did not match any established categories.

Although the tool we describe in this study proposes a potential solution to enable a rapid identification of causative pathogens in foodborne diseases, there is a limitation that the agreement ratio did not reach 100% in the current study. Considering this limitation, it is necessary for food inspectors not to overly commit on this tool, but just to utilize it as supportive information source, and to promptly act according to the evidence-based information obtained through microbiological examinations. Since we only used variables commonly documented in the National Epidemiological Surveillance of Foodborne Disease dataset in this study, this was considered an additional limitation of this study. The development of a comprehensive database of patients’ information including age, gender, and health and dietary histories, and the analysis using such metadata could be of significant value for increasing the agreement ratio of the similar prediction.

Changes in technologies and social environments during the long period for data collection (from 1950 to 2016) may affect the result of this study. To test the effect of dispersion of the reported years, we divided the whole cases into three groups, those reported between 1954 and 1979 (P1), 1980 and 1999 (P2), and 2000 and 2016 (P3). When the clinical information extracted from each group were separately subjected to the prediction by our tool, obtained agreement ratios of the first-ranked prediction were 82.4% for P1, 86.9% for P2 and 90% for P3. The results suggest that, although some degrees of influence cannot be negligible, changes in social environments during this period do not affect our conclusion. This influence, however, remains to be examined by further analysis.

ACKNOWLEDGMENTS. We thank Shiga prefecture’s food sanitation inspectors for providing data on food poisoning incidents and we would like to thank Dr. Eiji Inoshita and Hideyuki Sawa who gave permission to use the data. We would also like to thank all the food sanitation inspectors in Osaka prefecture, Kyoto prefecture, Nara prefecture, and Osaka city, where we obtained data from.

REFERENCES

1. Addis, M. and Sisay, D. 2015. A review on major food borne bacterial illnesses. *J. Trop. Dis.* **3**: 176.
2. Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. pp. 267–281. *In*: Proceedings of the 2nd International

Symposium on Information Theory (Petrov, B. N. and Caski, F., eds.), Akadimiai Kiado, Budapest.
3. European Food Safety Authority and European Centre for Disease Prevention and Control 2015. The European Union summary report on trends and sources of zoonoses, zoonotic agents and food-borne outbreaks in 2013. *EFSA J.* **13**: 3991. [CrossRef]
4. Foley, R. W., Shirazi, S., Maweni, R. M., Walsh, K., McConn Walsh, R., Javadpour, M. and Rawluk, D. 2017. Signs and symptoms of acoustic neuroma at initial presentation: an exploratory analysis. *Cureus* **9**: e1846. [Medline]
5. Giorgi Rossi, P., Faustini, A., Perucci C. A., Regional Foodborne Disease Surveillance Group 2003. Validation of guidelines for investigating foodborne disease outbreaks: the experience of the Lazio region, Italy. *J. Food Prot.* **66**: 2343–2348. [Medline] [CrossRef]
6. Hall, J. A., Goulding, J. S., Bean, N. H., Tauxe, R. V. and Hedberg, C. W. 2001. Epidemiologic profiling: evaluating foodborne outbreaks for which no pathogen was isolated by routine laboratory testing: United States, 1982-9. *Epidemiol. Infect.* **127**: 381–387. [Medline] [CrossRef]
7. Havelaar, A. H., Haagsma, J. A., Mangen, M. J., Kemmeren, J. M., Verhoef, L. P., Vijgen, S. M., Wilson, M., Friesema, I. H., Kortbeek, L. M., van Duynhoven, Y. T. and van Pelt, W. 2012. Disease burden of foodborne pathogens in the Netherlands, 2009. *Int. J. Food Microbiol.* **156**: 231–238. [Medline] [CrossRef]
8. Hedberg, C. W., Palazzi-Churas, K. L., Radke, V. J., Selman, C. A. and Tauxe, R. V. 2008. The use of clinical profiles in the investigation of foodborne outbreaks in restaurants: United States, 1982–1997. *Epidemiol. Infect.* **136**: 65–72. [Medline] [CrossRef]
9. Hosmer, D. W. and Lemeshow, S. 1989. Applied Logistic Regression. 2nd ed., A Wiley-Interscience Publications, New York (https://onlinelibrary.wiley.com/doi/book/10.1002/0471722146) [accessed on January 28, 2018].
10. Linacre, J. M. 2002. Optimizing rating scale category effectiveness. *J. Appl. Meas.* **3**: 85–106. [Medline]
11. McDonald, J. H. 2014. Handbook of Biological Statistics, 3rd ed. Sparky House Publishing, Baltimore, Maryland (http://www.biostathandbook.com) [accessed on January 15, 2018].
12. Naimi, T. S., Wicklund, J. H., Olsen, S. J., Krause, G., Wells, J. G., Bartkus, J. M., Boxrud, D. J., Sullivan, M., Kassenborg, H., Besser, J. M., Mintz, E. D., Osterholm, M. T. and Hedberg, C. W. 2003. Concurrent outbreaks of *Shigella sonnei* and enterotoxigenic *Escherichia coli* infections associated with parsley: implications for surveillance and outbreak control. *J. Food Prot.* **66**: 535–541. [Medline] [CrossRef]
13. O'Brien, S. J., Elson, R., Gillespie, I. A., Adak, G. K. and Cowden, J. M. 2002. Surveillance of foodborne outbreaks of infectious intestinal disease in England and Wales 1992-1999: contributing to evidence-based food policy? *Public Health* **116**: 75–80. [Medline]
14. Olsen, S. J., MacKinnon, L. C., Goulding, J. S., Bean, N. H. and Slutsker, L. 2000. Surveillance for foodborne-disease outbreaks—United States, 1993-1997. *MMWR CDC Surveill. Summ.* **49**: 1–62. [Medline]
15. R Core Team 2016. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. (https://www.R-project.org/) [accessed on December 10, 2017].
16. Scharff, R. L. 2012. Economic burden from health losses due to foodborne illness in the United States. *J. Food Prot.* **75**: 123–131. [Medline] [CrossRef]
17. Schneider, G., Chicken, E. and Becvarik, R. 2016. NSM3: Functions and Datasets to Accompany Hollander, Wolfe, and Chicken-Nonparametric Statistical Methods, 3rd ed. R package version 1.9. (https://CRAN.R-project.org/package=NSM3). [accessed on December 10, 2017].
18. The Ministry of Health, Labor, and Welfare, Japan (ed). 2015. Standard Methods of Analysis in Food Safety Regulation, Microbiological methods, Japan Food Hygiene Association, Tokyo.
19. Ushijima, H., Fujimoto, T., Müller, W. E. G. and Hayakawa, S. 2014. Norovirus and foodborne disease: a review. *Food Safety* **2**: 37–54. [CrossRef]