2010

Comparison of three time-series models for predicting campylobacteriosis risk

Jennifer Weisent
University of Tennessee-Knoxville

William Seaver
University of Tennessee-Knoxville

Agricola Odoi
University of Tennessee-Knoxville

Barton Rohrbach
University of Tennessee-Knoxville

Epidemiol. Infect. (2010), **138**, 898–906. © Cambridge University Press 2010
doi:10.1017/S0950268810000154

Comparison of three time-series models for predicting campylobacteriosis risk

J. WEISENT[1]*, W. SEAVER[2], A. ODOI[1] AND B. ROHRBACH[1]

[1] *Departments of Comparative Medicine and* [2] *Statistics, Operations and Management Science, The University of Tennessee, Knoxville, TN, USA*

(*Accepted 31 December 2009; first published online 22 January 2010*)

SUMMARY

Three time-series models (regression, decomposition, and Box–Jenkins autoregressive integrated moving averages) were applied to national surveillance data for campylobacteriosis with the goal of disease forecasting in three US states. Datasets spanned 1998–2007 for Minnesota and Oregon, and 1999–2007 for Georgia. Year 2008 was used to validate model results. Mean absolute percent error, mean square error and coefficient of determination (R^2) were the main evaluation fit statistics. Results showed that decomposition best captured the temporal patterns in disease risk. Training dataset R^2 values were 72·2%, 76·3% and 89·9% and validation year R^2 values were 66·2%, 52·6% and 79·9% respectively for Georgia, Oregon and Minnesota. All three techniques could be utilized to predict monthly risk of infection for *Campylobacter* sp. However, the decomposition model provided the fastest, most accurate, user-friendly method. Use of this model can assist public health personnel in predicting epidemics and developing disease intervention strategies.

Key words: *Campylobacter*, epidemiology, forecasting, time series, zoonoses.

INTRODUCTION

Campylobacter sp. bacteria are motile, spiral-shaped, Gram-negative organisms found ubiquitously in the environment [1, 2]. They have been identified as a leading cause of human gastroenteritis in developed nations, surpassing pathogens such as *Salmonella* sp. and *E. coli* [3, 4]. An estimated 1% of the US population (2 400 000 persons) are infected annually resulting in 13 000 hospitalizations and 124 deaths [5]. *Campylobacter* sp. can be found in the gastrointestinal tracts of a wide variety of domestic and wild animals and birds [6–8]. As a result, establishment of causative associations between human infection and contaminated food or water, animal contact and other environmental sources is a formidable task. Geographic region, climate patterns, drinking and recreational water, land use and human behaviour comprise some of the complex set of determinants which have been shown to affect the rate of gastrointestinal disease [9–15]. The incidence of campylobacteriosis varies seasonally and geographically, and tends to be highest in summer months, specifically in temperate climate zones [3, 14, 16, 17]. While the seasonality of the disease has been well documented worldwide, extensive studies have not been performed to predict the future risk of disease in different geographic regions in the USA. Comparing seasonal patterns in regions with different environmental

* Author for correspondence: Dr J. Weisent, Department of Comparative Medicine, The University of Tennessee, College of Veterinary Medicine, 205A River Drive, Knoxville, TN 37996, USA.
(Email: jweisent@utk.edu)

characteristics may help identify transmission routes making reliable time-series forecasting of great benefit to epidemiologists and public health officials [10, 18–20].

A variety of modelling approaches have been applied to surveillance data over the past 20 years in an attempt to accurately predict patterns of infectious diseases [14, 15, 18, 19, 21–29]. Statistical time-series modelling is appropriate since *Campylobacter* sp. disease surveillance data can be aggregated into equally spaced time intervals, exhibits autocorrelation, trend and seasonality [27, 30]. The potential for emerging infectious disease patterns to change in response to anthropogenic climate and land-use changes warrants the continual improvement and updating of current forecasting systems. Technological advances in forecasting software and program capability produce systematic review of methods and their applicability in the realm of public health. Recent interest in automated, real-time detection techniques have met with varying levels of success [28]. Our study incorporates a univariate methodological approach to forecast monthly disease risk using campylobacteriosis incidence from three US states.

Finding the most accurate time-series disease risk model at the state level holds numerous practical implications. Systematic analyses of multiple modelling techniques aims to create an optimal model to be used by public health officials with a state-specific, accurate and user-friendly method for predicting disease risk. The best model could potentially be implemented by trained public health professionals. Risk forecasting could provide public health officials with an early indication of irregularity in disease incidence and act as an epidemic alert system [18, 24, 27, 31, 32]. Model application could subsequently result in more efficient and cost-effective control strategies [33].

The purpose of this study is to evaluate three time-series models using data from three US states, Georgia, Oregon and Minnesota, to forecast the monthly risk of campylobacteriosis one year in advance. We also aim to determine if current software is capable of accurately simplifying time-series methods for practical use in the public health arena.

METHODS

Data source and study area description

The data utilized for this project were obtained from FoodNet, an active surveillance system implemented in 1996 by the Centers for Disease Control and Prevention (CDC) [34]. To meet the operational case definition of campylobacteriosis, samples of either stool or blood must be laboratory-confirmed as positive for *Campylobacter* sp.

Data from Georgia, Oregon and Minnesota were chosen for completeness and climatic diversity. Both direct and indirect disease transmission may be affected by weather conditions, therefore, it is important to predict disease risk for geographically diverse regions [14]. Oregon experiences temperate climatic conditions characterized by 9 months of consistent cloud cover and rain [35]. Regional variation in annual precipitation (50–500 cm) occurs. During the summer months, July–September, there are about 50 days of clear sky with average daily temperatures between 30 and 38 °C. Georgia is characterized by a humid subtropical climate and receives about 114 cm of annual rain in the middle of the state and 180 cm in the northeast mountains [36]. Summers are hot and humid with an average daily temperature of 32 °C. Minnesota climate is the most extreme, with average daily temperatures ranging in January between −14 and −11 °C, and between 19 and 23 °C in July [37]. Average annual precipitation is 48 cm in Minnesota's northwest region and 86 cm in the southeast. We hypothesize that climatic differences between states may affect the characteristics of the campylobacteriosis risk curve over the course of the year. Subsequently, this may influence statistical forecasting methods, as well as prevention and control strategies.

Data preparation

FoodNet surveillance data was aggregated into counts by month for each state over the study period resulting in 108 data points in Georgia and 120 data points in Oregon and Minnesota, equally spaced over time. The series lengths are statistically appropriate for the three time-series methods [38]. To ensure the regional integrity of the risk estimates, cases identified as travel related were eliminated from the dataset. The years 1998 (1999 for Georgia) to 2007 were used to model each time series and the year 2008 was held out of the dataset for model validation. Data manipulation was performed in SAS version 9.2 [39]. Risks were determined using annual population estimates as denominators obtained from the U.S. Census Bureau [40]. The risk estimates were presented as number of cases/100 000 persons. The statistical analyses were performed in NCSS-2007 [41]. The forecasting methods

used were time-series regression, decomposition, and Box–Jenkins autoregressive integrated moving averages (ARIMA). Fit statistics and holdout R^2 values were calculated manually. Separate model forecasts were assessed for each state.

Pattern analysis and outlier identification

Pattern analysis was performed on monthly risk data using autocorrelation (ACF) and partial autocorrelation (PACF) plots. Kruskal–Wallis ANOVA was performed on monthly medians to verify seasonality ($P<0{\cdot}05$). A simplistic strategy of identifying outliers as data points 3 s.d. from the mean for time-series data are invalid since this ignores the autoregressive or moving average patterns in the data. Instead, the time-series outliers were identified by fitting a basic ARIMA model to the data series. The resulting residuals are saved and standardized by the root mean square error for the ARIMA (1,0,0)(0,1,1). These standardized residuals are then control charted. Observations outside 3 s.d. from the mean of zero are then flagged as outliers in the time-series data. This outlier identification procedure avoids over-identification of outliers in time series [42].

Outbreak information on individual cases is incomplete in this dataset. All cases were aggregated by month regardless of outbreak status. Outliers identified by control charting were individually checked for potential outbreak status. No association between outlier months and reported outbreak cases was found.

Time-series modelling techniques

The models were quantitatively evaluated based on their predictive ability using mean square error (MSE), MAPE (mean absolute percentage error), R^2 on the training data, and a holdout R^2 based on 2008 data for all three modelling techniques. Outside of time-series analysis, most people associate R^2 only with multiple regression. However, there is a pseudo-R^2 that can be computed for any time-series model as follows:

$$R^2_{\text{pseudo}} = 1.0 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}. \quad (1)$$

This pseudo-R^2 is simply the sum of the residuals squared divided by the total sum of squares in the model. For a holdout R^2, the calculation is basically the same as in equation (1), except that the model from the training sample is applied to the holdout, the sample size is only over the holdout, and $\bar{y}$ is the mean for the holdout time period. In essence, all models can be evaluated in the same way. Henceforth, all further comparisons will be addressed simply as R^2.

Time-series regression

Ordinary least squares multiple regression models were evaluated using additive (untransformed) and multiplicative (logarithm transformation) risks. Predictors included trend, month, year and trend × month interactions. Variables were retained if they improved predictive value (R^2), produced globally significant ($P<0{\cdot}05$) models with significant regression coefficients, and lacked collinearity (variance inflation index <5). The basic time-series regression model used was additive and shown in equation (2):

$$Y_t = \beta_0 + \beta_1 x_{\text{trend}} + \sum_{i=2}^{p=12} \beta_i d_i + \varepsilon_t . \quad (2)$$

This model assumes linear trend and seasonality but no interaction between the two.

Residual time-series plots were examined for all models and checked for normality using the Shapiro–Wilk's goodness-of-fit for normality. In addition, one wants to find white noise (no pattern) in the residuals after fitting a time-series model. Therefore, the Portmanteau test was used to assess white noise, with degrees of freedom adjusted according to the number of predictor variables [38]. This test ensures that the pattern has been fully extracted from the series and that the residuals are randomly scattered.

Automatic decomposition

A decomposition macro available in NCSS and other software was applied [41]. The series was decomposed into trend, seasonal, cyclic and error components. The decomposition model that worked best on this data was multiplicative as shown in equation (3):

$$Y_t = T_t \cdot S_t \cdot C_t \cdot E_t. \quad (3)$$

Residual analysis for white noise and normality was performed as described for time-series regression.

Box–Jenkins ARIMA

The ARIMA modelling was based on the techniques described by Box & Jenkins in 1976 and further explained by DeLurgio [38, 43]. The ACF and PACF plots were used to identify starting orders. Exhaustive

combinations of autoregressive (AR), moving average (MA) and differencing parameters were fitted up to the third order. Orders above three were not attempted due to the high likelihood of model overspecification. First-order seasonal differencing resulted in the best models for all three states and compensated for non-stationarity in the mean [43]. The best models were selected after various fit statistics were evaluated. The best ARIMA model was ARIMA (1,0,0)(0,1,1), which is captured using backshift operators in equation (4):

$$(1-\varphi_1 B)(1-B^{12})Y_t=(1-\theta_1 B^{12})\,. \qquad (4)$$

Significant ($P<0{\cdot}05$) coefficients were retained with correlations $<0{\cdot}8$ between parameter estimates. Residual analysis, as to normality and white noise, was performed as described for time-series regression.

RESULTS

Pattern analysis

Monthly risks ranged from 0·236–1·191/100 000 persons (mean 0·593) in Georgia, 0·635–2·895 (mean 1·443) in Oregon and 0·333–4·655 (mean 1·435) in Minnesota. All three series demonstrate seasonality (Fig. 1*a–c*). The vacillating seasonal pattern in the ACF plots dominates and potentially masks AR and MA components. The ACF and PACF plots for Georgia are shown in Figure 2(*a*, *b*). The exponential decay in of the seasonality in the ACF along with the singular PACF first-order spike is indicative of AR(1). Further looks at regular and seasonal differencing hinted at a possible MA(1) for the seasonal component. The patterns were not clean, implying other model possibilities or potential outliers or both.

Outlier identification

Outliers were not identified in Georgia using the ARIMA control process techniques, therefore, no further pre-processing or smoothing methods were applied to the Georgia time series. For the Oregon series, June 1998 was flagged as an outlier in both the raw and residual control chart analysis. The mean risk in June was 2·11/100 000 persons. For the June observations a running median of five consecutive June values was chosen to preserve the seasonal effect. The original outlier value of 2·864/100 000 persons was replaced with 2·017. The models performed consistently worse with smoothed data. As a result, all Oregon forecasting was applied to the original unsmoothed data.

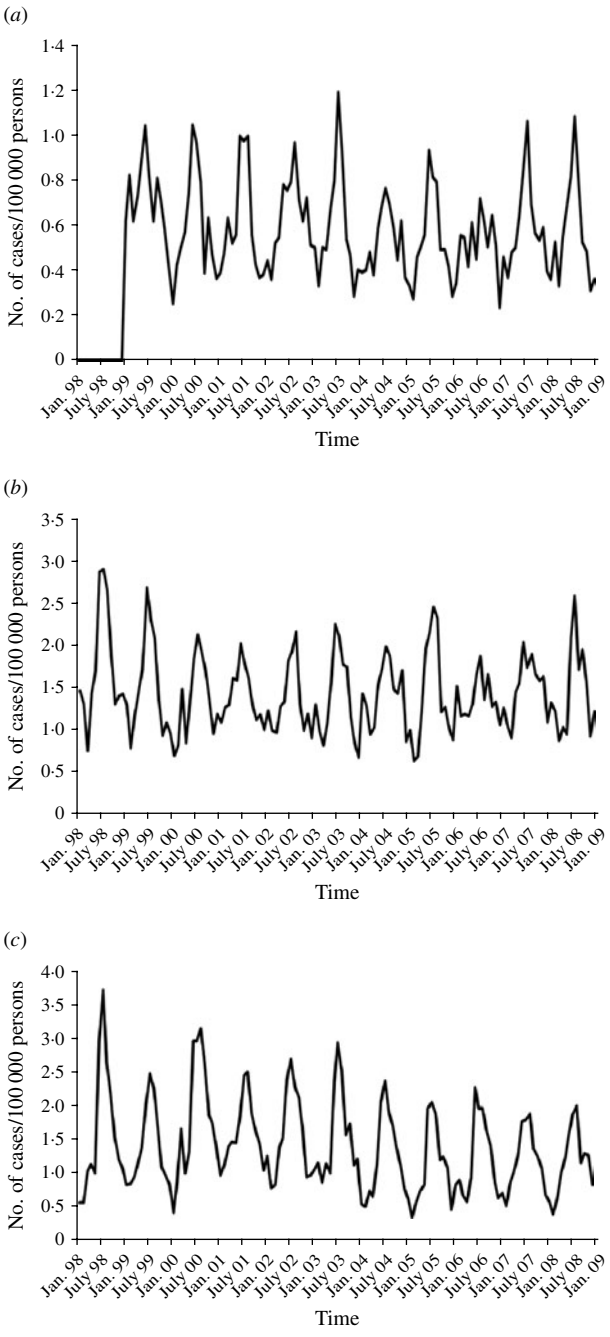

Fig. 1. Risk of campylobacteriosis per 100 000 persons in (*a*) Georgia (1999–2007), (*b*) Oregon (1998–2007) and (*c*) Minnesota (1998–2007).

Control charting of the Minnesota series indicated that June 1998 was out of range for both control charting techniques. The data point was above 3 S.D. from centre. To correct the outlier, a running median of 5 for consecutive June data was chosen for

Table 1. *Time-series model comparisons for campylobacteriosis risk per 100 000 persons in Georgia (1999–2007), Oregon and Minnesota (1998–2007)*

State	Model	R^2 PRED*	R^2 holdout†	MSE	Normality	WN (adequate lags)	MAPE
Georgia	Regression	0·693, 0·602	0·733	0·014	Yes	No (5–7)	0·162
	ARIMA (1,0,0)(0,1,1)	0·655	0·757	0·016	No – close	Yes	0·197
	Decomposition	0·727	0·662	0·011	Yes	No (1–2)	0·147
Oregon	Regression	0·710, 0·633	0·588	0·073	Yes	No (28 on)	0·154
	ARIMA (1,0,0)(0,1,1)	0·724	0·620	0·070	No – 2 off	Yes	0·177
	Decomposition	0·763	0·526	0·053	Yes	No	0·145
Minnesota	Regression	0·835, 0·792	0·682	0·089	Yes	No	0·194
	ARIMA (1,0,0)(0,1,1)	0·841	0·599	0·107	No – close	Yes – all except lag 1	0·219
	Decomposition	0·899	0·799	0·049	Yes	No (1, 4, 7–11)	0·156

MSE, Mean square error; WN, white noise; MAPE, mean absolute percent error.
* PRED is a prediction R^2 value (sometimes referred to as a Press R^2).This statistic is used to internally validate the regression model using jackknife techniques.
† R^2 of the 2008 validation sample.

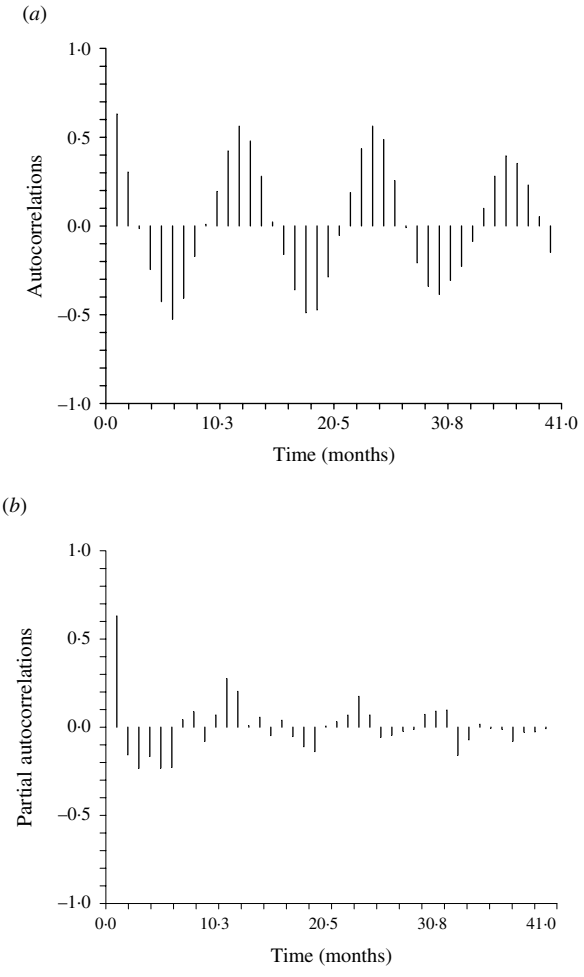

Fig. 2. (*a*) Autocorrelation and (*b*) partial autocorrelation plots for Georgia campylobacteriosis risk per 100 000 persons.

smoothing. The replacement median risk value of 2·372 (original value 4·65, June mean 2·420) was used in all further analyses.

Time-series model results and comparisons

The results for the best models identified for each technique are summarized in Table 1. All ARIMA and regression models were significant ($P<0·05$) both globally and for individual model coefficients. Decomposition models do not rely on overall model significance testing to assess fit.

Regression

The best regression model for all three states was additive and contained statistically significant ($P<0·05$) trend and monthly estimates. The R^2 value in Georgia was 69·3%, in Oregon 71·0% and in Minnesota 83·5%. In all three states, normality of the residuals was achieved but not white noise.

Automatic decomposition

The decomposition risk predictions for campylobacteriosis resulted in the highest fit statistics of the three methods. The R^2 value for Georgia was 72·7%, for Oregon 76·3% and for Minnesota 89·9%. Normality in the residuals was achieved for all three series. None of these models attained perfect white noise. The Georgia model was adequate for white noise only for lags 1 and 2. The Oregon model was

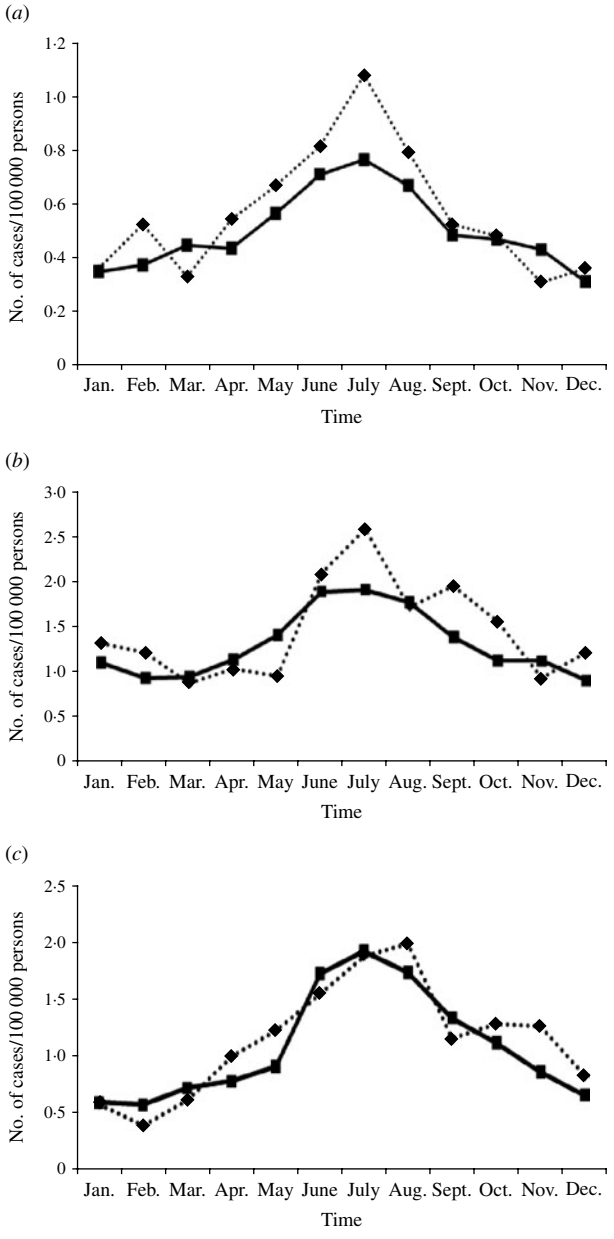

Fig. 3. Validation year (2008) actual (⋯◆⋯) *vs.* predicted (–■–) risk of campyloacteriosis per 100 000 persons in (*a*) Georgia, (*b*) Oregon and (*c*) Minnesota.

inadequate overall. The Minnesota residuals were adequate for white noise on lag periods 1, 4, and 7–11.

The actual and predicted risk values for the decomposition validation year (2008) are shown in Figure 3(*a–c*). The validation dataset R^2 values for Georgia, Oregon and Minnesota were 66·2%, 52·6% and 79·9%, respectively.

ARIMA

For all three states the most parsimonious model with highest R^2 value (Georgia 65·5%, Oregon 72·4%, Minnesota 84·1%) was ARIMA (1,0,0)(0,1,1). Better R^2 values were possible using AR and MA components higher than two. However, complex models tend to over-fit, demonstrate multicollinearity and frequently do not pass the assumptions of normality or achieve white noise. The majority of the models had markedly lower R^2 values than decomposition models, but the ARIMA models had a higher holdout R^2, except for Minnesota.

The observed seasonal variation in campylobacteriosis identifies the months of June, July and August as the highest risk months of disease for all three states. However, the overall shape of the curve differs across series (Fig. 4). Minnesota's annual curvature has the sharpest, narrowest seasonal peak until 2004 at which time the shape coincides closely with Oregon. The seasonal peak in Georgia is more rounded and less severe.

DISCUSSION

The automatic decomposition procedure resulted in a 4–7% improvement in R^2 over the best regression and ARIMA models. Comparing the three methods, decomposition was the fastest and least technical, achieved normality in the residuals yet was uniformly unsuccessful at achieving white noise. Lack of white noise implies that there is a pattern in the residuals not accounted for by the model. Residual patterns may increase model uncertainty. However, good predictive performance can be achieved without perfect attainment of white noise [22].

The use of ARIMA modelling for disease risk data is well documented [18, 27, 31, 33, 44]. It was originally expected that ARIMA methods would be favoured based on previously published use with surveillance data, versatility and available prediction intervals [18]. These data show that ARIMA models were closer to achieving white noise in the residuals and improved holdout sample fit statistics in Georgia and Oregon. Compared with automatic decomposition, this method is technically challenging, requiring significant statistical background for appropriate and accurate implementation. Regression had the poorest model R^2 results. Advantages of regression include the ease of interpretation, computation of prediction intervals, robust and bootstrapping possibilities. Therefore, this technique should not be ruled out for risk forecasting of campylobacteriosis. For all three methods, MSE and MAPE were comparable and indicate accurate forecasting. However, fit statistics

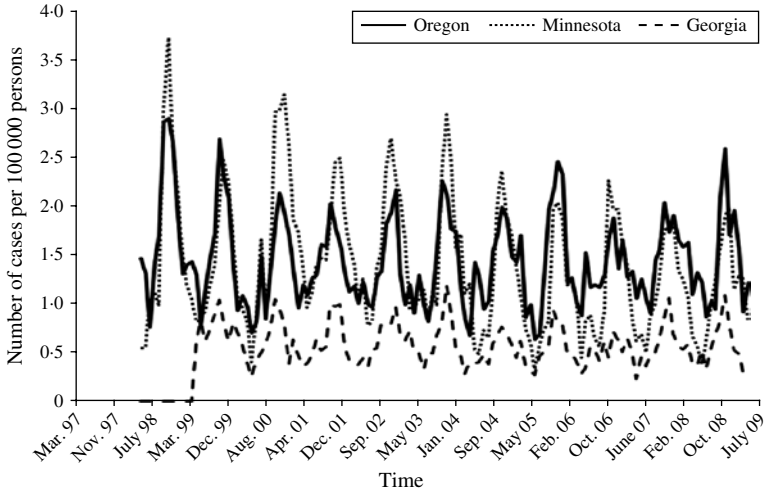

Fig. 4. Comparison of temporal patterns in risk of campylobacteriosis in Oregon, Minnesota and Georgia.

for decomposition were uniformly better than the other two methods across states. Furthermore, constant 95% prediction intervals can be manually calculated for decomposition to demonstrate a range in predicted risk values.

A specific strength of automatic decomposition is that it produced accurate monthly campylobacteriosis risk predictions for all three states. A unique characteristic of this technique is that it can be taught to public health officials with minimal statistical background. By combining accuracy with ease of use, improvements in epidemic preparation and timely intervention are attainable at state, regional and national levels [32].

The distinct seasonal pattern of campylobacteriosis may suggest climatic or environmental links to the risk of disease [13, 16, 17]. Climate affects the survival and reproduction of *Campylobacter* sp. in the environment and on food sources and previous studies have shown that climatic factors influence disease incidence over time [14, 25, 45, 46]. Hartnack *et al*. found significant cross-correlations between human incidence, monthly temperature and rainfall [10]. The study showed that peak prevalence in human campylobacteriosis preceded that in German broiler flocks, further implicating environmental *vs*. foodborne components to disease risk. Seasonal risk variation may also be due to human behavioural factors such as picnics, barbeques and other outdoor activities [9, 30]. Such behaviours may vary depending on the climatic and socioeconomic constraints of a geographic region. The timing of seasonal peaks in our study was comparable across states. This was in contrast to a recent study in Scotland which showed that the prominence of seasonal peak in incidence varied regionally [13]. However, both studies demonstrated differences in the shape of the seasonal curve by region or state. Future studies are needed to elucidate the impact of these factors on disease risk by dividing states into unique climatic zones for time-series analysis using environmental variables specific to different geographic regions.

Considerable variation was observed in validation data R^2 results across models and states. This may be a reflection of the model's predictive accuracy, shifts in disease patterns or reflect irregular values or outliers in the dataset. In Oregon and Minnesota, aberrant risk values were evident in 2006 seasonal peaks (Fig. 4). These data were not flagged by control charting and were not smoothed prior to the analysis. The presence of outliers, change points or interventions can alter patterns and invalidate forecasts. We believe our results would have improved in these states had 2006 followed the typical seasonal curvature. Second, surveillance systems can underestimate actual disease risk, and reporting may vary between states. As a result, predictions based on surveillance data should be interpreted with caution.

Over the past 20 years active modern surveillance systems have been implemented in developed nations that offer more accurate statistical prediction capacity than was previously possible [29, 32]. Risk data from surveillance systems can be modelled as a means of assessing associations between disease risk and epidemiological factors over time [10, 32]. Detecting aberrant disease incidence can signal an impending epidemic [31]. Currently, advanced software offers forecasting methods that are applicable for use by

public health officials [18, 27, 28]. These statistical computing techniques allow interdependence of observations in both time and space to be incorporated into epidemiological models. As a result the temporal structure of risk data may assist epidemiologists in modelling biological, environmental and behavioural factors of disease with greater accuracy than the classical one-dimensional regression framework [47]. As demonstrated in this study, these techniques may provide health officials with practical, user-friendly and accurate predictive warning systems based solely on previous risk data [27]. The models can be implemented and validated monthly for the practical purpose of predicting the risk of campylobacteriosis. This information may be useful for public health professionals in early epidemic alert systems as well as adding to our knowledge of seasonal disease patterns over time.

ACKNOWLEDGEMENTS

We thank the Centers for Disease Control and Prevention FoodNet active surveillance for providing data for this study.

DECLARATION OF INTEREST

None.

REFERENCES

1. **Snelling W, *et al.*** Campylobacter jejuni. *Letters in Applied Microbiology* 2005; **41**: 297–302.
2. **Humphrey T, O'Brien S, Madsen M.** Campylobacters as zoonotic pathogens: a food production perspective. *International Journal of Food Microbiology* 2007; **117**: 237–257.
3. **Allos BM, Taylor DN.** *Campylobacter* infections. In: Evans AS, Brachman PS, eds. *Bacterial Infections of Humans, Epidemiology and Control*, 3rd edn. New York: Plenum Medical Book Company, 1998, pp. 169–190.
4. **Altekruse SF, Swerdlow DL.** *Campylobacter jejuni* and related organisms. In: Cliver DO, Riemann HP, eds. *Foodborne Diseases*, 2nd edn. Boston: Academic Press, 2002, pp. 103–112.
5. **Tauxe RV.** Incidence, trends and sources of campylobacteriosis in developed countries: an overview. In: WHO consultation on the increasing incidence of human campylobacteriosis. Report and Proceedings of a WHO Consultation of Experts, Copenhagen, Denmark, 21–25 November 2000. World Health Organization, 2001, pp. 42–43.
6. **Frost J.** Current epidemiological issues in human campylobacteriosis. *Symposium Series* (*Society for Applied Microbiology*) 2001; **30**: 85S–95S.
7. **Vandamme PAR.** Methods for identification of *Campylobacter*. In: WHO consultation on the increasing incidence of human campylobacteriosis. Report and Proceedings of a WHO Consultation of Experts, Copenhagen, Denmark, 21–25 November 2000. World Health Organization, 2001, pp. 94–99.
8. **Newell DG, Wassenaar TM.** Strengths and weaknesses of bacterial typing tools for the study of campylobacteriosis epidemiology. In: WHO consultation on the increasing incidence of human campylobacteriosis. Report and Proceedings of a WHO Consultation of Experts, Copenhagen, Denmark, 21–25 November 2000. World Health Organization, 2001, pp. 101–104.
9. **Jepsen MR, Simonsen J, Ethelberg S.** Spatio-temporal cluster analysis of the incidence of Campylobacter cases and patients with general diarrhea in a Danish county, 1995–2004. *International Journal of Health Geographics* 2009; **8**: 11.
10. **Hartnack S, *et al.*** Campylobacter monitoring in German broiler flocks: an explorative time series analysis. *Zoonoses Public Health* 2009; **56**: 117–128.
11. **Hearnden M, *et al.*** The regionality of campylobacteriosis seasonality in New Zealand. *International Journal of Environmental Health Research* 2003; **13**: 337–348.
12. **Kovats R, *et al.*** Climate variability and campylobacter infection: an international study. *International Journal of Biometeorology* 2005; **49**: 207–214.
13. **Miller G, *et al.*** Human campylobacteriosis in Scotland: seasonality, regional trends and bursts of infection. *Epidemiology and Infection* 2004; **132**: 585–593.
14. **Bi P, *et al.*** Weather and notified Campylobacter infections in temperate and sub-tropical regions of Australia: an ecological study. *Journal of Infection* 2008; **57**: 317–323.
15. **Zhang Y, Bi P, Hiller J.** Climate variations and salmonellosis transmission in Adelaide, South Australia: a comparison between regression models. *International Journal of Biometeorology* 2008; **52**: 179–187.
16. **Nylen G, *et al.*** The seasonal distribution of campylobacter infection in nine European countries and New Zealand. *Epidemiology and Infection* 2002; **128**: 383–390.
17. **Tam CC, *et al.*** Temperature dependence of reported Campylobacter infection in England, 1989–1999. *Epidemiology and Infection* 2006; **134**: 119–125.
18. **Nobre FF, *et al.*** Dynamic linear model and SARIMA: a comparison of their forecasting performance in epidemiology. *Statistics in Medicine* 2001; **20**: 3051–3069.
19. **De Greeff SC, *et al.*** Seasonal patterns in time series of pertussis. *Epidemiology and Infection*. Published online: 31 March 2009. doi: 10.1017/S0950268809002489.
20. **Altizer S, *et al.*** Seasonality and the dynamics of infectious diseases. *Ecology Letters* 2006; **9**: 467–484.
21. **Tokars JI, *et al.*** Enhancing time-series detection algorithms for automated biosurveillance. *Emerging Infectious Diseases* 2009; **15**: 533–539.

22. **Burkom HS, Murphy SP, Shmueli G.** Automated time series forecasting for biosurveillance. *Statistics in Medicine* 2007; **26**: 4202–4218.
23. **Paul M, Held L, Toschke AM.** Multivariate modelling of infectious disease surveillance data. *Statistics in Medicine* 2008; **27**: 6250–6267.
24. **Medina DC, *et al.*** Forecasting non-stationary diarrhea, acute respiratory infection, and malaria time-series in Niono, Mali. *PLoS ONE* 2007; **2**: e1181.
25. **Fleury M, *et al.*** A time series analysis of the relationship of ambient temperature and common bacterial enteric infections in two Canadian provinces. *International Journal of Biometeorology* 2006; **50**: 385–391.
26. **Rhodes CJ, Hollingsworth TD.** Variational data assimilation with epidemic models. *Journal of Theoretical Biology*. Published online: 10 March 2009. doi: 10.1016/j.jtbi.2009.02.017.
27. **Williamson GD, Weatherby Hudson G.** A monitoring system for detecting aberrations in public health surveillance reports. *Statistics in Medicine* 1999; **18**: 3283–3298.
28. **Rolfhamre P, Ekdahl K.** An evaluation and comparison of three commonly used statistical models for automatic detection of outbreaks in epidemiological data of communicable diseases. *Epidemiology and Infection* 2006; **134**: 863–871.
29. **Naumova EN, *et al.*** Use of passive surveillance data to study temporal and spatial variation in the incidence of giardiasis and cryptosporidiosis. *Public Health Reports* 2000; **115**: 436–447.
30. **Naumova EN, MacNeil IB.** Time-distributed effect of exposure and infectious outbreaks. *Econometrics* 2009; **20**: 235–348.
31. **Cardinal M, Roy R, Lambert J.** On the application of integer-valued time series models for the analysis of disease incidence. *Statistics in Medicine* 1999; **18**: 2025–2039.
32. **Myers MF, *et al.*** Forecasting disease risk for increased epidemic preparedness in public health. *Advances in Parasitology* 2000; **47**: 309–330.
33. **Benschop J, *et al.*** Temporal and longitudinal analysis of Danish Swine Salmonellosis Control Programme data: implications for surveillance. *Epidemiology and Infection* 2008; **136**: 1511–1520.
34. **CDC.** Preliminary FoodNet Data on the incidence of infection with pathogens transmitted commonly through food – 10 states, 2008. *Morbidity and Mortality Weekly Reports* 2009; **58**: 333–337.
35. **Oregon Climate.** (http://wwwcity-datacom/states/Oregon-Climatehtml). Accessed 8 September 2009.
36. **Net State Geography.** (http://wwwnetstatecom/states/geography/ga_geographyhtm). Accessed 15 October 2009.
37. **Minnesota Department of Natural Resources (DNR).** (http://wwwdnrstatemnus/climate/indexhtml). Accessed 9 September 2009.
38. **DeLurgio SA.** *Forecasting Principles and Applications*, 1st edn. St Louis, MO: Irwin McGraw-Hill, 1997, pp. 802.
39. **SAS Institute.** Statistical analysis systems (SAS), version 9.2. Cary, North Carolina, USA: SAS Institute Inc., 2008.
40. **United States Census Bureau.** (http://factfindercensusgov/servlet/DatasetMainPageServlet?_program = PEP&_submenuId = &_lang = en&_ts = :). Accessed 5 June 2009.
41. **Hintze J.** NCSS, PASS and GESS. In: *NCSS*. Kaysville, Utah, USA, 2006.
42. **Alwan LC.** Time series modelling for statistical process control. *Journal of Business and Economic Statistics* 1988; **6**: 87–95.
43. **Box EP, Jenkins GM.** *Time Series Analysis: Forecasting and Control*, 3rd edn. Upper Saddle River, New Jersey: Prentice Hall, 1994, pp. 592.
44. **Reichert TA, *et al.*** Influenza and the winter increase in mortality in the United States, 1959–1999. *American Journal of Epidemiology* 2004; **160**: 492–502.
45. **Patrick M, *et al.*** Effects of climate on incidence of Campylobacter spp. in humans and prevalence in broiler flocks in Denmark. *Applied and Environmental Microbiology* 2004; **70**: 7474–7480.
46. **Bi P, *et al.*** Climate variability and Ross River virus infections in Riverland, South Australia, 1992-2004. *Epidemiology and Infection*. Published online: 20 March 2009. doi: 10.1017/S0950268809002441.
47. **Singer JD, Willett JB.** *Applied Longitudinal Data Analysis*. New York: Oxford University Press Inc., 2003, pp. 644.