Development of a predictive model for Ross River virus disease in Brisbane, Australia

Author
Hu, Wenbiao, Nicholls, Neville, Lindsay, Mike, Dale, Patricia, J. McMichael, Anthony, S. Mackenzie, John, Tong, Shilu

Published
2004

Journal Title
American Journal of Tropical Medicine and Hygiene

Copyright Statement
Copyright 2004 American Society of Tropical Medicine and Hygiene. The attached file is reproduced here in accordance with the copyright policy of the publisher. Please refer to the journal's website for access to the definitive, published version.

Downloaded from
http://hdl.handle.net/10072/5671

Link to published version
http://www.ajtmh.org/content/71/2/129.abstract
DEVELOPMENT OF A PREDICTIVE MODEL FOR ROSS RIVER VIRUS DISEASE IN BRISBANE, AUSTRALIA

WENBIAO HU, NEVILLE NICHOLLS, MIKE LINDSAY, PAT DALE, ANTHONY J. McMICHAE L, JOHN S. MACKENZIE, and SHI LU TONG

School of Public Health, Queensland University of Technology, Brisbane, Queensland, Australia; Bureau of Meteorology Research Centre, Brisbane, Queensland, Australia; Department of Health, Perth, Western Australia, Australia; National Centre for Epidemiology and Population Health, Australian National University, Canberra, Australia; Department of Microbiology and Parasitology, University of Queensland, Brisbane, Queensland, Australia

Abstract. This paper describes the development of an empirical model to forecast epidemics of Ross River virus (RRV) disease using the multivariate seasonal auto-regressive integrated moving average (SARIMA) technique in Brisbane, Australia. We obtained computerized data on notified RRV disease cases, climate, high tide, and population sizes in Brisbane for the period 1985–2001 from the Queensland Department of Health, the Australian Bureau of Meteorology, the Queensland Department of Transport, and Australian Bureau of Statistics, respectively. The SARIMA model was developed and validated by dividing the data file into two data sets: the data between January 1985 and December 2000 were used to construct a model, and those between January and December 2001 to validate it. The SARIMA models show that monthly precipitation ($\beta = 0.004, P = 0.031$) was significantly associated with RRV transmission. However, there was no significant association between other climate variables (e.g., temperature, relative humidity, and high tides) and RRV transmission. The predictive values in the model were generally consistent with actual values (root mean square percentage error $= 0.94\%$). Therefore, this model may have applications as a decision supportive tool in disease control and risk-management planning programs.

INTRODUCTION

There has been global resurgence of arboviral diseases.$^1$ Ecologic changes and human behavior are important in the spread of these diseases.$^2$ Infection with Ross River virus (RRV) is the most prevalent vector-borne disease in Australia. This virus was isolated from mosquitoes in Australia in 1963,$^3$ from people in 1972,$^4$ and from mosquitoes and people in the Pacific islands in 1979–1980 during epidemics of epidemic polyarthritis.$^5–8$ The virus causes epidemic polyarthritis, with rash and fever in some patients, and joint symptoms in 95% of the cases.$^9–11$

The arthritic symptoms may persist for months and can be severe and debilitating. There is no effective treatment of the disease and, in the absence of a vaccine, prevention remains the sole public health strategy. Over the last 11 years (1992–2002), a total of 48,242 laboratory confirmed cases of RRV disease have been identified in Australia.$^{12}$ In general terms, RRV activity appears to have increased in Australia in the past decade.$^{13,14}$ but causes for this increase remain largely unknown.$^{13–15}$

The ecology of RRV disease is complex, and the mechanism of RRV transmission remains unclear. The virus is maintained in a primary mosquito–mammal cycle involving macropods (kangaroos and wallabies), possibly other marsupials (e.g., possums), flying fox, and native rodents.$^{16}$ A human–mosquito cycle may occur in explosive outbreaks, but in Australia does not appear to extend over more than three or four cycles (Lindsay M, Mackenzie JS, unpublished data). The incubation period in humans ranges from 5 to 21 days,$^{16}$ which may explain why the disease is transmitted rapidly once an outbreak occurs. Extreme weather events (e.g., heavy rainfall) often trigger outbreaks of RRV disease at a lag of one to two months.$^{17–19}$ Such lags may assist disease control managers in effectively planning public health interventions in advance. Nevertheless, a quantitative relationship between climate variability and the RRV transmission remains to be determined.

A number of studies have examined the relationship between climate variation and RRV disease.$^{17–22}$ Several models have been developed to predict the likelihood of RRV epidemics using weather and environmental data.$^{23,24}$ However, some important methodologic issues such as stationarity and auto-correlation of time series data have not been formally addressed in previous research.

Time series analyses have been increasingly used in epidemiologic research.$^{25–38}$ Since the early 1970s, time series methods, in particular seasonal autoregressive integrated moving average (SARIMA) models, which have the ability to cope with stochastic dependence of consecutive data, have become well established in the commercial and industrial fields.$^{28,29}$

This study examines the potential impact of climate variability on the transmission of RRV disease and explores the possibility of developing an epidemic forecasting system for RRV disease using the multivariate SARIMA technique in Brisbane, Australia.

MATERIALS AND METHODS

Study area. Brisbane, the capital of Queensland State, is a subtropical city situated on the east coast of Australia and covers approximately 1,142 km$^2$ (Figure 1). The coastal areas are flat, with extensive mangrove forests, salt marshes, and mudflats. Within the administrative boundaries of Brisbane City Council, which also determines the study area of this investigation, the population was 888,449 on June 30, 2001.$^{39}$ Brisbane had the highest number of RRV cases notified (4376 cases) in Queensland between 1985 and 2001. The number of cases was highest in 1996 (849 cases), and lowest in 1988 (41 cases). Figure 2 shows that most of the cases occurred from February to May. The average annual incidence was 29.14/100,000.

Data collection. Since infection with RRV is a notifiable disease, positive test results have, by law, to be reported by laboratories to the Queensland Department of Health, where
they are archived by the Communicable Diseases Unit. The requirement for notification of RRV disease is based on a demonstration of IgM antibodies in blood, a four-fold or greater change in serum antibody titers between acute and convalescent phase sera, isolation of RRV, or demonstration of arboviral antigen or genome in blood.40 The reported place of onset for each case was generally regarded to reflect the geographic distribution of RRV infection.41

We obtained the computerized data set on the notified RRV disease cases in Brisbane for the period of 1985 to 2001 from the Queensland Department of Health. Data provided for each notification case included a unique record reference number, disease code, date and place of onset, sex, age, and the confirmation status of the report. Climate and population data were obtained for the period 1985–2001 from the Australian Bureau of Meteorology and the Australian Bureau of Statistics, respectively. There are three main meteorologic stations in Brisbane. The one with the longest history of meteorologic recordings was chosen. The information is reported to the Australian Bureau of Meteorology regularly. Climate data comprised monthly mean maximum and minimum temperature (°C), total precipitation (mm), and mean relative humidity (%). Data on monthly mean high tidal levels (cm) along the coastal regions were supplied by the Queensland Department of Transport. The reason for using monthly parameters was that the incidences of RRV disease were too low if weekly or daily indices were used.

**Data analysis** Univariate/bivariate analyses were conducted for each independent variable. Cross-correlations were used to compute a series of correlations between climate variables and the incidence of RRV disease over a range of time lags (a time lag was defined as the time span between climatic observation and the incidence of RRV disease).42 For adequate modeling, a time series should be stationary with respect to mean and variance. If the mean increases or decreases over time, the series may need to be transformed (e.g., differenced) to make it stationary, before being modeled.37 A simple inspection of the graph of the untransformed series is the most useful approach. Similarly, if the variance (as indicated by the excursions around the mean becoming smaller or larger over time) increases or decreases, some transformation (logarithm or square root, etc.) should also be applied. A time series with seasonal non-stationary may be transformed to stationary data by taking seasonal differences into account.25,24

Since both RRV incidence and climate variables exhibited strong seasonal variation and fluctuations in their yearly means, we adjusted for seasonality by first seasonally differencing the series (i.e., each observation is replaced by the difference between it and the observation a year before) in the analysis. However, for variables without a clear secular trend, regular differencing (each observation is replaced by the difference between it and the previous observation) was not used.

The multivariate SARIMA model with environmental variables was used to estimate the independent contribution of each climate variable and of high tide in this study. Four steps were undertaken in the modeling of the relationship between climate variation and the RRV transmission. First, the monthly incidences of RRV infection were calculated using monthly counts of RRV disease cases as a numerator and population size in the middle of each year as a denominator and, then, both RRV disease incidences and environmental variables were transformed to become stationary input series with respect to yearly periodicity by seasonally differencing, before being modeled. Second, SARIMA models were developed using the monthly incidence of RRV as a response variable and monthly climate variables and high tides as explanatory variables. Six main parameters were selected when fitting the SARIMA model: the order of autoregressive (p) and seasonal autoregressive (P), the order of integration (d) and seasonal integration (D), and the order of moving average (q) and seasonal moving average (Q). The process is called SARIMA (p,d,q) (P,D,Q). (s is the length of seasonal period). The equation of this model is $y_t = \Theta_p(B)(1 - B)^d(1 - B)^D + \phi_p(B)^s(1 - B)^d(1 - B)^D + \phi_P(B)^s(1 - B)^D + P\text{ precipitation at lag 2 months},$ where $\phi_p(B)$ is seasonal autoregressive operator, $\phi_p(B)$ is autoregressive operator, $\Theta_p(B)$ is the operator of moving averages, $\Theta_P(B)^s$ is seasonal operator of moving averages, $a_t$ is white noise, $y_t$ is the dependent variable, and $P$ is explanatory variables’ regressive coefficients). The selection of SARIMA processes was conducted using Akaike’s information criterion (AIC) which measures how well the model fits the series.43 Of all the models tested, an SARIMA (1,0,1) (1,1,1)12 model was found to best fit the data. Meanwhile, we used a stepwise regression method to select the environment variables. Additionally, a comparison of the SARIMA with and without environmental variables was also conducted. Third, the goodness of fit of the models was checked for adequacy, using both time series (auto-correlation functions of residuals) and classic tools (to check the normality of residuals). Finally, the model developed was verified by dividing the data file into two data sets: the data between January 1985 and December 2000 were used to construct a SARIMA model and those between January and December 2001 were used to validate the model. Such a validation method is widely used in time series analyses.44,45

In addition, the predictive validity of the models was evaluated by using the root mean square (RMS) error and RMS percentage error criterion (RMS error $= \sqrt{\frac{\sum(y_t - \hat{y}_t)^2}{N}}$; RMS percentage error $= \sqrt{\frac{\sum(y_t - \hat{y}_t)/|y_t|}{N}}$), where $\hat{y}_t$ is the predicted values and $y_t$ is the observed values for
month \( t \), \( N \) is the number of observations). The smaller the RMS error, the better the model in terms of the ability of forecast. We used the 12 monthly forecasts and calculated the RMS error of these 12 forecasts, and then compared this with the RMS error if the historical average number of RRV cases each month was used as a forecast. All analyses were performed with SPSS software using the Trends procedure.

RESULTS

Summary statistics for each independent variable are shown in Table 1. The monthly mean maximum and minimum temperatures, precipitation, relative humidity, and high tide were 25.30°C, 15.60°C, 90.76 mm, 65.78%, and 201.01 cm, respectively, between 1985 and 2001 in Brisbane. Bivariate
analyses show that there appeared to be positive relations between climate variability and the transmission of RRV disease (Figure 2). The results of the cross-correlations adjusted for season show that the incidence of RRV disease was significantly associated with maximum temperature at the current month, precipitation at lags of 0–3 months, relative humidity at lags of 0–3 months, and high tide at a lag of two months (Figure 3). However, there was no significant association between minimum temperature and RRV transmission in any lag.

The SARIMA models show that auto-regression ($\beta = 0.364, P < 0.001$), moving average ($\beta = -0.544, P < 0.001$), seasonal auto-regression ($\beta = -0.354, P < 0.001$), seasonal moving average ($\beta = 0.641, P < 0.001$), and monthly total...
precipitation ($\beta = 0.004$, $P = 0.031$) were significantly associated with RRV transmission. It suggests that there may be 50 more cases a year for an increase of 100 mm precipitation on average in Brisbane. However, the maximum and minimum temperatures, relative humidities, and the high tides were not significantly associated with the monthly incidence of RRV disease after adjustment for auto-correlation, seasonality, and other covariates. The model estimated with the climate variables was a better fit than the model without these variables (i.e., the log-likelihood increased, while the values of AIC decreased) (Table 2).

Figure 4 shows that there was no significant auto-correlation between residuals at different lags in the SARIMA model. The graphic analysis shows that the residuals in the model appeared to fluctuate randomly around zero with no obvious trend in variation as the predicted incidence values increased. Thus the goodness-of-fit analyses showed that the model fits the data reasonably well.

The model constructed with the data between January 1985 and December 2000 was used to predict the transmission of RRV disease in Brisbane between January and December 2001, and was then validated by the actual observations (Figure 5). The validation analyses indicate that the model had reasonable accuracy over the predictive period, even though the predicted spike was slightly ahead of the actual peak. The validity of the model increased with the inclusion of precipitation (RMS error = 1.96 and RMS error percentage = 0.94%) compared with the model without this variable (RMS error = 2.01 and RMS error percentage = 1.73%).

**DISCUSSION**

This is the first attempt to develop an epidemic forecasting model for predicting RRV transmission in a metropolitan area according to our knowledge. The results of this study suggest that climatic variability, particularly precipitation, may have played a significant role in the transmission of RRV disease in Brisbane, given that other socioecologic conditions have been unlikely to change dramatically on a monthly time scale in this city over the past two decades. The key determinants of the RRV disease transmission included auto-regression, moving average, seasonal auto-regression, seasonal moving average, and precipitation at lag two months. These variables may be used to assist in forecasting outbreaks of RRV disease in Brisbane.

Changes in climate and the environment may influence the abundance and distribution of vectors and intermediate hosts of RRV disease.\(^{48,49}\) Precipitation is important in the transmission of mosquito-borne diseases including RRV infection. All mosquitoes have aquatic larval and pupal stages and therefore require water for breeding. Quantity, timing, and pattern of precipitation would first affect the breeding of mosquitoes, especially fresh water ones, because sufficient amounts of precipitation will assist in maintaining the breeding habitats further into the summer months. In addition, precipitation can also markedly affect the breeding of salt

**TABLE 1**

Characteristics of explanatory variables

<table>
<thead>
<tr>
<th>Variables</th>
<th>N</th>
<th>Mean</th>
<th>SD</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximum temperature</td>
<td>204</td>
<td>25.30</td>
<td>3.04</td>
<td>18.97</td>
<td>30.73</td>
</tr>
<tr>
<td>Minimum temperature</td>
<td>204</td>
<td>15.60</td>
<td>4.24</td>
<td>7.37</td>
<td>22.72</td>
</tr>
<tr>
<td>Precipitation (mm)</td>
<td>204</td>
<td>90.76</td>
<td>89.27</td>
<td>0.00</td>
<td>616.80</td>
</tr>
<tr>
<td>Relative humidity (%)</td>
<td>204</td>
<td>65.78</td>
<td>6.69</td>
<td>46.30</td>
<td>81.37</td>
</tr>
<tr>
<td>High tide (cm)</td>
<td>204</td>
<td>201.01</td>
<td>7.09</td>
<td>177.43</td>
<td>228.82</td>
</tr>
</tbody>
</table>
marsh mosquitoes, allowing breeding to be prolonged after periods of high tides. In this study, monthly total amount of precipitation was positively correlated with the monthly incidence of RRV infection in Brisbane over the study period, with two month lagged effect. Some outbreaks of RRV infection were predominantly precipitation associated. Similar findings have been reported for other vector-borne diseases. For example, precipitation along with humidity and temperature were found to be related to epidemics of malaria in Pakistan.

Nearly 40 species of mosquitoes belonging to six genera have been confirmed as vectors of RRV in Australia. Ochlerotatus vigilax and Ochlerotatus camptorhynchus are probably the major vectors in Brisbane region. Some freshwater breeding species, such as Culex annulirostris, Coquillettidia linealis, and Ochlerotatus normanensis, are also the vectors of RRV. These freshwater mosquitoes are closely associated with human habitation. How hosts, vectors, and humans interact with socioenvironmental changes remains to be determined.

The lagged effect of precipitation, at a lag of two months, on the incidence of RRV infection is very important. Such delays are consistent with the development of mosquitoes, the external period of incubation of RRV within mosquitoes, and the incubation period of the virus in the host. Lags of 1−3 months were considered because of biologic plausibility (i.e., the time period for the development of mosquitoes, the extrinsic incubation period, and the incubation period of the virus within host). Longer lags (e.g., ≥ 4 months) were found less important in this study.

It is crucial to use adequate research methodology in the assessment of possible impacts of environmental variability on disease transmission. Recently, increasing attention has

Figure 3. Cross-correlation function between Ross River virus and climate variables after seasonal differencing. CI = confidence interval.
focused on the use of the Box-Jenkins modeling strategy to construct SARIMA models for vector-borne disease.\textsuperscript{27,45,53} The modeling strategy analyses a long series of values in a stationary mode. However vector-borne disease and climatic variables of interest are not stationary, and analysts have to resort to preliminary transformations, such as time series differencing or variance stabilizing to achieve stationarity. In this study, a multivariable SARIMA model was used to determine the independent effects of climate variables on the transmission of RRV. The steps of model identification, parameter estimation, and diagnostic checking were performed as recommended.\textsuperscript{27,44}

The SARIMA modeling is a useful tool for interpreting and applying surveillance data in disease control and prevention. Once a satisfactory model has been obtained, it can be used to forecast expected numbers of cases for a given number of future time intervals.\textsuperscript{37} Since the SARIMA model has the capacity to forecast when and where an outbreak is likely to occur, it therefore has great potential to be used as a decision-supportive tool for planning public health interventions.

This study has three strengths. First, a sophisticated time-series model was used in this attempt to develop an epidemic forecasting system for the control and prevention of RRV disease in metropolitan areas. Second, detailed infor-

<table>
<thead>
<tr>
<th>Variables</th>
<th>Model without climate variables\textsuperscript{*}</th>
<th>Model with climate variables\textsuperscript{†}</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\beta$</td>
<td>SE</td>
</tr>
<tr>
<td>Autoregression</td>
<td>0.383</td>
<td>0.090</td>
</tr>
<tr>
<td>Moving average</td>
<td>–0.510</td>
<td>0.084</td>
</tr>
<tr>
<td>Seasonal autoregression</td>
<td>–0.395</td>
<td>0.083</td>
</tr>
<tr>
<td>Seasonal moving average</td>
<td>0.607</td>
<td>0.081</td>
</tr>
<tr>
<td>Precipitation at lag 2 month</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

\textsuperscript{*} Log likelihood = –488, Akaike’s information criterion = 984.

\textsuperscript{†} Log likelihood = –455, Akaike’s information criterion = 920. Other climate variables (i.e., minimum temperature, maximum temperature, relative humidity, and high tides) were adjusted.

\begin{figure*}[h]
\centering
\includegraphics[width=\textwidth]{figure4.png}
\caption{Auto-correlation, partial auto-correlation of residuals, and scatterplot of residuals at different lags in the seasonal auto-regressive integrated moving average model. CI = confidence interval.}
\end{figure*}
ation on a range of climate and tidal variables was incor-
porated in the model. Third, the model developed in this
study appeared to have a high degree of accuracy and may
have implications in the disease control and risk-management
planning.

The weaknesses of this study must be acknowledged. First,
this is a broad, ecologic assessment of the relationship be-
tween climate and tide variability and the transmission of
RRV at the city level. More detailed risk assessment at com-
munity and individual levels may also be required if a com-
prehensive and systematic risk assessment is to be made. In-
clusion of other information (e.g., virus strain, mosquito
population densities and survival, human behaviors, popula-
tion immunity, housing characteristics, and other mosquito-
relevant environmental information) may further improve the
model. Second, the model may only be applicable to Brisbane
and areas with a similar socioecologic background, since only
local data were used in the construction of the model. Third,
under the Notifiable Diseases Surveillance System, the place
of onset is assumed as the place of acquisition of infection.
However, two places may differ for some cases, particularly
during holiday periods. If these risk factors could be included
in this model, a more effective and accurate model would be
expected (e.g., predicted errors would be further improved).

The development of epidemic forecasting systems is impor-
tant in the control and prevention of infectious disease out-
breaks in the future. Should an outbreak of RRV disease
occur, a large-scale public health intervention is usually re-
quired. Early warning based on forecasts from the model can
assist in improving vector control and personal protection.
Increasing insecticide spraying during high-risk periods and
decreasing it during low-risk periods will improve cost-
effectiveness of operations. Disease control programs, if ant-
icipating an increase in RRV disease, can increase vigilance,
e.g., by alerting district health offices, filling vacant positions
of health staff, and requesting more frequent reporting to
facilitate early identification of problem areas. Additionally,
the disease surveillance data can be integrated with social,
biologic, and environmental databases. These data may pro-
vide additional input into the development of epidemic fore-
casting models. These attempts, if successful, may have sig-
nificant implications in environmental health decision-making
and practices, and may help health authorities determine pub-
lic health priorities more wisely and use resources more effec-
tively and efficiently.

Received January 7, 2004. Accepted for publication March 10, 2004.

Acknowledgments: We thank the Queensland Department of
Health, the Australian Bureau of Meteorology, Queensland trans-
port, and the Australian Bureau of Statistics for providing the data on
notified RRV disease cases, climate, sea level, and population growth
between 1985 and 2001, respectively.

Financial support: This study was partly funded by the National
Health and Medical Research Council, the Queensland Department
of Health, and the Queensland University of Technology

Authors’ addresses: Wenbiao Hu and Shilu Tong, School of Public
Health, Queensland University of Technology, Brisbane, Queens-
land 4059, Australia, Telephone: 61-7-3864-9745, Fax: 61-7-3864-
3369, E-mail: s.tong@qut.edu.au. Neville Nicholls, Bureau of Meteor-
ology Research Centre, Melbourne, Victoria 3001, Australia. Mike
Lindsay, Department of Health, Perth, Western Australia 6001, Aus-
tralia. Pat Dale, Australian School of Environmental Studies, Griffith
University, Brisbane, Queensland 4111, Australia. Anthony J. Mc-
Michael, National Centre for Epidemiology and Population Health,
Australian National University, Canberra 0200, Australia. John S.
Mackenzie, Department of Microbiology and Parasitology, Univer-
sity of Queensland, Brisbane, Queensland 4072, Australia.

REFERENCES

1. Reiter P, 2001. Climate change and mosquito-borne disease. En-
third group A arbovirus in Australia, with preliminary obser-
vations on its relationship to epidemic polyarthritis. Aust J Sci
26: 183–184.


