Reviewer's report

Title: Combined Mathematical Model in predicting incidence of hepatitis E using 2000-2012 morbidity data from Shanghai, China

Version: 1 Date: 21 April 2013

Reviewer: Tassanee Silawan

Reviewer's report:

Minor Essential Revisions

Overall
The paper should be properly checked for English corrections before being published

Abstract

Background:
“Therefore, some improved tools based on mathematical mode were developed to forecast hepatitis E epidemics.”

- model?
- Do the authors want to forecast “epidemics” or “incidence”? Methods for forecast epidemics are different from the methods for forecast incidence.

“Due to both linear and nonlinear patterns of time series morbidity data of infectious disease such as hepatitis E, applicable statistical model cannot be easily determined.”

- This should directly mention only for hepatitis E, focus on the reasons for using combined mathematical model.

Methods:
The following details should be added:
- Source (surveillance system?) and characteristics (laboratory confirmed?) of data
- Data points (units) for analysis (Annual/ Monthly/ Cases/ Rates?)
- Explain more about (and how to develop) the combined model
- Time horizon for training (Fitted) set and testing (predicted) set for model diagnostic check
- Time horizon for fitted data and forecasted data

Results:
- The equation of ARIMA would be skipped, few people can understand and we rather want to know the equation of the combined model

…. of single ARIMA model and combined model….
-Specify the combined model
The predicted incidence of hepatitis E in 2013 would keep fluctuating within a narrow range from 0.095 to 0.372 per 100,000 persons with a seasonal variation.
-The word “predicted” ……revise to “forecasted”
-Clearly explain about the seasonal variation
Conclusion:
-Conclusions should not beyond the findings.

Background
The last paragraph:
“Although ARIMA models are quite flexible in that they can represent several different types of time series, their major limitation is the pre-assumed linear form of the model. On the contrary, artificial neural networks (ANNs) have a good ability to learn and to describe the highly-nonlinear and strongly-coupled relationships between multi-input and multi-output variables. With ANNs, there is no need to specify a particular model form. But ANNs alone is also not able to handle both linear and nonlinear patterns equally well.
-Add evidences or findings from previous researches which indicated an inaccuracy (not so fit) of those models

Materials and Methods
1st paragraph:
-Explain how to calculate the monthly rates. Was the same population a denominator for all 12 months in the same year?

Results
Should explain more details of the results. The focus should be on comparison between ARIMA and ARIMA-BPNN combined model.

Data description
-It should be “General patterns of hepatitis E”
-Graphic presentation (line graph of monthly rates) should be presented instead of table. The characteristics of trend and seasonal variations should be clearly explained.
“‘The male morbidity was significantly higher than that of female (t = 8.951, P<0.001)”
-What were the reasons for comparison of morbidity between male and female? Were these differences taken into account in forecast?
“Meanwhile, X-12-ARIMA seasonal adjustment program showed that the monthly morbidity of hepatitis E had seasonal and trend variations (F = 40.02, P <0.001)”
- No evidence support.
The best-fitting ARIMA model
-Is logarithmic transformation needed?
- The following results should be added and clearly explained;
  • error of the fit model
  • predicted values (from model identifying), and how the predicted values close to the actual values.
  • forecasted values in 2013 (line graph by month should be presented).
ARIMA-BPNN combined model
  - should explain on predicted values (from model identifying), and how the predicted values close to the actual values.
  - Line graph by month should be presented (can be compared with ARIMA).
Note: If possible, some of the actual data in year 2013 should be compared with the forecasted values, both ARIMA and ARIMA-BPNN combined model.
Discussion
  “Compared with other infectious diseases, the epidemic of hepatitis E is influenced by individual, environmental and social factors”
  - Other infectious diseases are also influenced by those factors.
  “These multiple factors will lead us to the difficulties attempting to employ variables regression models to forecast the epidemics. Time series analysis method takes advantage of associations in the seasonally and sequentially lagged relationships. Without independent variables, it can also automatically determine the best-fitting model.”
  - These are not reasonable.
  The discussion should cover the following key findings of the current research by comparing with the previous research and giving the reasons for those phenomenon or findings.
  - Pattern; trend and seasonal
  - Model, and
  - Forecast.
  - The effectiveness of the model; strength, weakness, points need to be concern
  - Implications
Conclusions
  - Conclusions should not beyond the findings.
Table 1
  - Should present monthly cases or rates (per 100,000 pop.) of...from...to... using line graph (to show the general patterns, trends and seasonality)
Table 2
  - Can be omitted. The significant estimated parameters should be > 2.0.
Table 3
- Forecasted rates from Jan to Dec 2013 should be separated from table 3 and better presented using line graph to show the magnitude and seasonality.
- Table 3 Predicted and error rates of ..... from ..... to ..... 
- Date ..... revise to Month
- Morbidity (per 100,000) ..... revise to Morbidity (per 100,000 pop.)
- Validation ..... revise to Predicted rate
- Des 2012 ..... revise to Dec 2012
- The last line should present the mean error rate (MER).

The following line graphs should be presented.
- Line graph of monthly cases or rates from ..... to ..... (to show general patterns, trend, and seasonal of the disease).
- Line graph of monthly cases or rates from ..... to ..... (to show forecasted values).

Discretionary Revisions

To make the article more attractive, the words “using 2000-2012 morbidity data from” should be removed from the title. The title may be
- Combined Mathematical Model in predicting incidence of hepatitis E in Shanghai, China. Or
- Combined Mathematical Model to forecast hepatitis E incidence in Shanghai, China. Or
- Forecast for hepatitis E incidence in Shanghai of China using Combined Mathematical Model.

Level of interest: An article whose findings are important to those with closely related research interests

Quality of written English: Needs some language corrections before being published

Statistical review: No, the manuscript does not need to be seen by a statistician.

Declaration of competing interests:

I declare that I have no competing interests.