Reviewer’s report

Title: A decision aid to rule-out pneumonia and reduce unnecessary prescriptions of antibiotics in primary care patients with cough and fever

Version: 5 Date: 4 February 2011

Reviewer: Feng Gao

Reviewer’s report:

In general, this is a well written manuscript. It raises a clinical question of broad interest. The data are collected prospectively and sufficiently representative the targeting population. The methods are properly used and the results are well interpreted. The authors are well aware of the strength and limitation for each method.

- Major Compulsory Revisions

1. Because there are only n=127 cases with pneumonia, the logistic model is likely over-fit with all 25 variables included. So, a variable selection (say, backward or forward selection procedure) is needed to identify a more parsimonious model and to assess its performance (validation, AUC, etc.)

2. The big algebra formula provided in the Box (4th paragraph in the Results) is not helpful. Either a nomogram (if the purpose of the BOX is to develop a prediction rule) or a calibration plot (if the purpose of the BOX is to provide information of an internal validation) will be more informative. These methods can be found in the book “Regression Modeling Strategies” by Frank Harrell, and can be easily implemented using the free-download Design library in the statistical package R.

- Minor Essential Revisions

A few typos are noted:

1. Last paragraph in Introduction: “… pneumonia t in patients with …”;

2. Last paragraph in Statistical Analysis: “…… could have been avoided in this group. I.e. The diagnostics tool was ……”

- Discretionary Revisions

1. Two statistical methods (logistic regression and CART) are used in this paper, but the purpose of logistic regression is not clear. Is it serving to pre-select the factors used in CART, or to develop a separate prediction rule. Either way is OK, but need to be clearly stated.

2. It is not a good idea to test the sensitivity only using the 4 imputed data. A better approach is to use the 10-fold cross-validation. This is similar to what the authors did for the validation of logistic regression, except that leave 10th
individuals out. This remains an internal validation, but it provides better information on model stability than using different imputed data.

3. The criteria of variable-selection (third paragraph of Statistical Analysis) for using in CART are not clear. Are they selected after the logistic regression analysis, or simply pre-selected based on the easy availability and reliability as indicated in the paper.

4. Actually pre-selection may not be necessary for CART at all. As an exploratory tool, one of the important features of CRAT is its capability to identify unexpected yet potentially meaningful interactions among predictors. With the selected subset, one will lose this advantage. Since the data in this study is collected prospectively which is not frequently seen in a diagnostics study like this, a CART with full data may provide useful information for the relationship among predictors.

5. An alternative choice for variable pre-selection is to use so called “random forest”. The underlying idea of “forest” is very similar to CART, except that it runs hundreds or thousands trees in the same time, and thus allows one to rank the relative importance of predictors. Personally, this reviewer uses forest to select factors and then use CART to develop decision rules. Again, this method can be easily implemented using the corresponding library in the R package.

**Quality of written English:** Acceptable

**Statistical review:** Yes, and I have assessed the statistics in my report.

**Declaration of competing interests:**

I declare that I have no competing interests