Reviewer's report

Title: Predicting Disease Risks from Highly Unbalanced Data using Random Forest

Version: 3 Date: 21 April 2011

Reviewer: W. Edward Hammond

Reviewer's report:

General Comments. (Discretionary Revisions)

The authors have done an excellent job of responding to the critiques. The value of the paper is the discussion of using Random Forest. I also like the use of a publically available database. However, I think the limitations of the HCUP database severely limit the value of the research. The fact that future diagnoses may be included in the analyses has a significant impact on the conclusions. The fact that significant data is missing that would logically be used in predicting the occurrence of a disease seems important. For example, obesity is a major causative factor in many diseases. It also is not clear to me the degree of bias in your study. For example, it seems that unbalance was determined by whether a disease existed or not. If the analysis was focused on whether a disease was present or not, the data used for analysis would have to be restricted to those encounters in which the disease was not present. I would think the analysis would be based on pre-documentation of a disease, a risk factor calculated, and then determine if that disease was ever documented for that patient. It appears that the HCUP database would not permit such an analysis.

You also make conclusions but do not support those conclusions with evidence. An example is how you deal with demographic parameters. Logically age would be a factor because of the evidence that more disease occurs in older people. You state you do not understand the influence of age, and, in the conclusions, you state that age was less a factor than expected. Was gender a factor in diseases that were more gender based – for example, breast cancer has a much higher prevalence in females?

I would be interested in seeing your research repeated on a population of patients with complete data and observe these patients in a longitudinal study. Clearly, age would be a factor in how long one would wait for the disease to show up. In fact, an interesting component of the study would be if you could predict a time-line, that is, a risk factor that had a time component. The risk is 15% for the next 10 years, 30% for the next 10 years; and 50% for the next 10 years. Has such a study been done?

Since you did not use the diagnoses codes but the disease categories, how do you know, as stated in the conclusions, that disease categories work better than disease codes? If you believe that, I also would like a comment about why you think that might be. Does it, perhaps, have anything to do with the precision of the diagnosis?
You are still careless in the use of disease versus disease categories. I really don’t know to which you are talking about when you do that. I think you must be consistent. You also use unbalance and imbalance – are they the same? Why not be consistent?

When you describe the data elements you are using, procedures, and diagnoses, I also would like to know what additional clinical data elements are used. If I am to predict the risk for a disease, it must be based on data elements such as physical findings, laboratory results and such prior to the onset of a disease. Else it is not a prediction. I fail to see this point made anywhere in your paper. My earlier comments apply.

It is interesting to note that in Figure 1, the diseases you use for breast cancer include gender in the diagnoses with exception of DX code 233.0.

Tables 1 and 2 deal with specific diseases (?), yet you state that you deal only with disease categories. Should not these tables reflect disease categories rather than diseases? You are missing x-axis labels for race.

You misspell eldercare on page 6.

With prevalence being such a strong factor, I believe that other approaches such as a Bayesian approach is more appropriate.

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

**Quality of written English:** Acceptable

**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.