Reviewer's report

Title: Genome Wide Association Study to Predict Severe Asthma Exacerbations in Children using Random Forests Classifiers

Version: 1 Date: 20 January 2011

Reviewer: Dana Hancock

Reviewer's report:

I commend the authors on a well-written manuscript addressing an intriguing research question. A random forests classification algorithm was used to build a predictive model of severe asthma exacerbations with genetic and clinical predictors in a population of 417 asthmatic individuals (127 with exacerbation and 290 without exacerbation). The findings indicate that incorporation of many genetic risk variants (160 SNPs) provided a prediction of asthma exacerbations that exceeded the prediction based solely on clinical variables. Other studies integrating genetic associations into prediction risk algorithms have found that the genetic predictors produce minimal improvement in risk estimation based only on phenotype/clinical variables (including Talmud et al. BMJ 2010; 340: b4348). This study is novel in that it uses random forests modeling which incorporates many more genetic predictors and considers interactions among the predictors as opposed to considering the predictors one by one.

Major Compulsory Revisions
None.

Minor Essential Revisions
1. Background, 3rd paragraph: Most of the references for the first sentence (“Asthma is a complex diseases known to be….”) do not cite the cumulative evidence supporting a genetic component to asthma, and there are no references cited to support any of the environmental risk factors.

2. Results, section on Sample characteristics: The description of Table 1 results states that “The four clinical traits (age, gender, FEV1%, and treatment group) used as covariates in the study are also similar in the two groups.” However, the patterns of the clinical traits do appear to differ between the training and testing populations. For instance, males are more likely to have an exacerbation in the training population, but males are less likely to have an exacerbation in the testing population.

3. Results, Prediction of severe asthma exacerbations: The 2nd sentence refers to Figure 1, but Figure 2 is the relevant figure. Also, please provide more explanation of the 4 comparisons in Figure 2 (permutation, training, internal cross-validation, and independent replication), given the variability in these tests especially for higher number of SNPs.
4. Discussion, 1st paragraph, 2nd sentence: It is not clear how the sensitivity, specificity, PPV and NPV values were calculated (AUC=0.66, sensitivity=0.66, specificity=0.66, PPV=0.81, and NPV=0.74). Also, the AUC estimates do not exceed 0.66, as stated in this sentence. This section seems more appropriate in Results than in Discussion.

5. Discussion, 1st paragraph, sentences pertaining to reference 17: Another interesting study for comparison is the Talmud et al. study of type 2 diabetes in >5,000 participants (BMJ, 2010; 340:b4838). They showed that addition of several replicated genetic risk factors to prediction models produced minimal improvement in risk estimation based only on phenotype/clinical variables.

6. Table 1. Please include the standard deviations for age and FEV1% in the testing population.

7. Discussion, 9th paragraph, 1st sentence: It appears that “Supplementary Figure 1” should be replaced with “Figure 4”.

Discretionary Revisions
None.

Level of interest: An article of importance in its field

Quality of written English: Acceptable

Statistical review: Yes, but I do not feel adequately qualified to assess the statistics.

Declaration of competing interests:

I declare that I have no competing interests.