Reviewer’s report

Title: Supervised Learning Methods in Modeling of CD4+ T Cell Heterogeneity

Version: 1

Date: 7 February 2015

Reviewer: Brett McKinney

Reviewer’s report:

Some details about the data and analysis were unclear to me from the description:

1. Typically, Random Forest has one output variable. How were the multiple output cytokines handled in RF?

2. When output cytokines are predicted, does it not matter what the baseline levels are for prediction? Is it presumed they are near zero because of cell type?

3. What time point is chosen for determining steady-state levels?

4. In Table 1, it appears that four simulated instances were used to train and test the learners with different combinations of initial cytokine levels. This seems like a small number. Is this a problem when doing the bootstrap bagging in Random Forests?

5. In real data, what is a realistic number of training instances? How independent are they?

6. The noise added to the simulated data seems low. Are these noise levels realistic?

7. Some of the paper suggests that the supervised method can predict cell differentiation. Perhaps it can indirectly? I think a more accurate characterization might be that the method predicts the levels of cytokines that help characterize some cell types. I think some clarification of this distinction would be helpful.

Minor:

Table 1 column headings say Training and Testing, but do you mean input and output cytokines?

Spelling: “as oppose to” should be “as opposed to.”

Word choice: “which approved that our models are capable.” Instead of “approved,” a more appropriate word might be “demonstrated.”

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 have no competing interests.