Reviewer's report

Title: Mining FDA Drug Labels for Medical Conditions

Version: 2  Date: 8 January 2013

Reviewer: Erik van Mulligen

Reviewer's report:

The authors describe a method based on CRFs to extract known side effects from FDA drug labels. It is not completely clear why this method is innovative and different from what others have done.

Minor Essential Revisions

With respect to the setup of the experiment that is some essential information lacking. First of all the authors extracted 7024 tokens and 74 annotated entities for manual error analysis. How was this selection made? Randomly? Stratified for the different label sections? The error analysis is not provided in this manuscript, where it would be very helpful.

Why did you choose for a 4-fold cross validation and not for a 10-fold or so?

With respect to the preprocessing it is unclear (and important) whether a stopword list has been used and stemming. It is known that drugs can be written in various different writing styles. From the Overdosage and Usage sections was also the dosage information extracted?

From table 5 I don't understand that an exact match gives a lower precision score than a partial match. I understand the increasing recall. This needs more clarification, for instance an error analysis.

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