Reviewer’s report

**Title:** Improved de-identification of physician notes through integrative modeling of both public and private medical text

**Version:** 3  **Date:** 19 August 2013

**Reviewer:** Murat Kantarcioglu

**Reviewer’s report:**

In this paper, authors discuss their machine learning based techniques for removing identifying information in medical notes. One novel aspect of this work is that they leverage the existing medical literature to model non-identifying medical terms. After this step, they use standard natural language processing and machine learning techniques to build models to classify terms in medical notes as PHI or non-PHI.

**Minor Revisions:**

Authors addressed most of comments but "the scored reported in this paper would have placed first overall in recall and last overall in precision.". It is not than in many cases you can increase recall by lowering precision and vice versa. One solution to address this issue is to use area under the curve measure (ROC AUC). It would be good to use AUC to compare against all the existing techniques.

**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’