PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (**http://bmjopen.bmj.com/site/about/resources/checklist.pdf**) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Generating high-quality data abstractions from scanned clinical records: Text mining-assisted extraction of endometrial carcinoma pathology features as proof of principle
AUTHORS	Nguyen, A; O'Dwyer, John; Vu, Thanh; Webb, Penny; Johnatty, Sharon; Spurdle, Amanda

VERSION 1 – REVIEW

REVIEWER	Ha Hoang The University of DaNang, VietNam
REVIEW RETURNED	28-Mar-2020

GENERAL COMMENTS	Authors propose a method to automatically extract clinical information from scanned clinical pathology report , that are unstructured, by using text mining. The authors have done effective analysis in the paper. Well-written paper, good piece of Work. But the paper should be improved: 1) The state of the art 2) The algorithm in page 9 (from line 3 to line 29) should be presented clearly in algorithm form (step by step like block diagram, pseudocode ...) 3) There is something wrong in line 3 (page 9): "…. using hard copies of the abstraction form." instead of "…. using hard copies of the Pathology reports" 4) This method do not mention the missing characters in the hard copies as "leiomy mas"

REVIEWER	Katrien Groenhof Julius Center for Health Sciences and Primary Care, Netherlands
REVIEW RETURNED	17-Apr-2020

GENERAL COMMENTS	I thank the editor for the opportunity to review the manuscript entitled " Generating high-quality data abstractions from scanned clinical records: Text mining-assisted extraction of endometrial carcinoma pathology features as proof of principle". This study aimed to use text mining for assisted extraction of clinical features from pathology records on endometrial features. Considering the widespread availability of scanned text and the potential goldmine of information captured herein I think research in this field is very important. Also, the detailed description of the technical process of the mining tool, including the 'training" and noise-reduction steps, are well appreciated. Lastly, adding the diagnostic accuracy to the paper shows validity of your tool is essential for uptake by clinicians and inadmissible for papers on text mining tools (though

	very frequently omitted). With minor revisions I would very much recommend to accept this paper for publication. Some small specific remarks: - Introduction: Please provide references to paragraph 1, - Methods: very clear - Results: o great visualisation of regular expression patterns o to dot the i's : please provide full text for abbreviations (such as PPV) - Discussion o The added value of OCR error correction on to op negations was insignificant. Please elaborate why you think this was the case for your proof-of-principle example. Could you think of case studies where OCR correction might have more added value? How about if you have hand-written pdf's? Could it be used for handwriting discrepancies between reporters for example? o Western-medicine has mostly transported written medical health records to EHR/EMRs. Could this technique also be applied to migration of paper records to EHR? This would significantly speed up the migration process. Relevant to for instance developing countries. Please elaborate.

VERSION 1 – AUTHOR RESPONSE

Reviewer 1

Authors propose a method to automatically extract clinical information from scanned clinical pathology report, that are unstructured, by using text mining. The authors have done effective analysis in the paper. Well-written paper, good piece of Work.

But the paper should be improved:

1) The state of the art

The proposed method integrated readily available and proven approaches from a couple of different disciplines, namely OCR error correction and text mining. The state-of-the-art in each of the different disciplines was noted in the Introduction.

The limitation of using "readily available and proven" approaches has been added to the "Strength and Limitations" section, and the "Discussion" section has been updated to place the proposed method in context with the state-of-the-art in each of the disciplines.

2) The algorithm in page 9 (from line 3 to line 29) should be presented clearly in algorithm form (step by step like block diagram, pseudocode ...)

The information on page 9 refers to the manual, paper-based data collection. It does not describe an algorithm. As a result, we left the description of the data collection unchanged.

3) There is something wrong in line 3 (page 9): "…. using hard copies of the abstraction form." instead of "…. using hard copies of the Pathology reports"

The sentence is correctly referring to "abstraction forms" but the context of the broader sentence may have caused misinterpretation. We have revised the sentence to make it clearer: "Information manually extracted from pathology reports was recorded using hard copies of the abstraction form."

4) This method do not mention the missing characters in the hard copies as "leiomy mas"

The Discussion section contains a paragraph that describes how additional search patterns could be specified in the system. It provides an example of how '?endometriosis' could be accommodated. Missing characters and other OCR error patterns could likewise be accommodated if they were affecting the system's performance. A sentence has been added to include OCR error patterns as a type of search pattern that could be specified: "Additional search patterns may include new OCR error patterns and writing variations such as medical shorthand notations."

Reviewer 2

I thank the editor for the opportunity to review the manuscript entitled " Generating high-quality data abstractions from scanned clinical records: Text mining-assisted extraction of endometrial carcinoma pathology features as proof of principle". This study aimed to use text mining for assisted extraction of clinical features from pathology records on endometrial features. Considering the widespread availability of scanned text and the potential goldmine of information captured herein I think research in this field is very important. Also, the detailed description of the technical process of the mining tool, including the 'training" and noise-reduction steps, are well appreciated. Lastly, adding the diagnostic accuracy to the paper shows validity of your tool is essential for uptake by clinicians and inadmissible for papers on text mining tools (though very frequently omitted). With minor revisions I would very much recommend to accept this paper for publication.

Some small specific remarks:

- *Introduction: Please provide references to paragraph 1,*

An additional 3 references have been added to paragraph 1.

- *Methods: very clear*
- *Results:*

- *great visualisation of regular expression patterns*

- *to dot the i's : please provide full text for abbreviations (such as PPV)*

PPV was defined on first use in the Evaluation section. Tables 5 and 6, however, were not self-sufficient in explaining what PPV was, so the full term has been added as a table footnote to each of the tables.

- *Discussion*

- *The added value of OCR error correction on top negations was insignificant. Please elaborate why you think this was the case for your proof-of-principle example. Could you think of case studies where OCR correction might have more added value? How about if you have hand-written pdf's? Could it be used for handwriting discrepancies between reporters for example?*

The value of OCR error correction is dependent on the quality of the OCR software employed and the type of artifacts present in the scanned versions of the pathology reports. This statement has now been added to the respective paragraph. However, we believe that regardless of how good the OCR software or scanned reports are, improvements in extraction performances due to OCR error correction would still be of value.

The documents of concern in this study were typewritten reports as opposed to handwritten reports. The use of typewritten documents has been made explicit throughout the manuscript (see "typewritten" in Abstract, Strengths and Limitations, Introduction and Discussion). As such we have not explored handwritten documents, but further study would be necessary to evaluate a role of our system on handwritten documents. We have added this in the Discussion section (second last paragraph).

- *Western-medicine has mostly transported written medical health records to EHR/EMRs. Could this technique also be applied to migration of paper records to EHR? This would significantly speed up the migration process. Relevant to for instance developing countries. Please elaborate.*

EMRs, as mentioned in the second last paragraph in the Discussion section, can contain a substantial amount of scanned medical records. This paragraph also elaborates on the role of our system on both handwritten and typewritten documents.

Its use in "clinical settings" to provide searchable databases of medical records is also important, and this has now been added in the last paragraph of the Discussion.