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)	Using natural language processing to extract self-harm and suicidality data from a clinical sample of patients with eating disorders: a retrospective cohort study
AUTHORS	cliffe, charlotte; Seyedsalehi, Aida; Vardavoulia, Katerina; Bittar, André; Velupillai, Sumithra; Shetty, Hitesh; Schmidt, Ulrike; Dutta, Rina

VERSION 1 – REVIEW

REVIEWER	Pickrell, W. Owen Swansea University, Swansea University Medical School
REVIEW RETURNED	10-Jun-2021

GENERAL COMMENTS	Thanks for asking me to review this interesting retrospective cohort study on risk factors for suicidality and self harm in people with eating disorders. It is useful research and will be of intrerest to the community. I have some minor comments and points to improve the manuscript. 1. In the results section of the abstract, it perhaps would be better and clearer to simply state the precision and recall of positive mentions and suicidality i.e. 0.80 and 0.82 and 0.96 and 0.94. 2. The introduction is a little long. It would aid readability if it were a little shorter. 3. It would be useful to know a bit more about the clinical records (e.g. what proportion were clinical letters, clinical notes, inpatient records etc.) 4. For transparancy there needs to be more detail on the NLP algorithms, how were they written? Did they use a specific NLP system. I could not at first glance find much more detail in the supplied references 5. What is the difference between a document which contains "positive suicidality" and a document which contains "relevant suicidality" - this needs to be explained in more detail. 6. For transparency the raw numbers used to calculate the figures in table 2 should be published either in the main manuscript or as supplementary data e.g. number of true positives etc. 7. The "deprivation scores 1,2 and 3" need to be explained in more detail - how are they derived? 8. There is no need to repeat the information displayed in tables 4,5 and 6 in text preceeding the tables, please delete this to aid readability. 9. It would be useful if the authors could elaborate on why they think that precision and F1 scores lower for negative mentions?

	10. I think it would also be worth mentioning or discussing in more detail the possible heterogeniety in the recording of mentions of suicide and self harm in clinical encounters. Are some clinical encounters or some healthcare professionals more likely to discuss self harm and suicide?

REVIEWER	Dowell, Anthony Wellington School of Medicine and Health Sciences, General Practice
REVIEW RETURNED	16-Jun-2021

GENERAL COMMENTS	In all areas of health care , routinely collected clinical information can be difficult to access for research purposes. Eating disorders are an important component of psychiatric morbidity and as the authors state are at a higher risk of mortality compared to the general population, significantly due to self harm and suicide. Exploring risk factors for suicidality is hampered by the relatively rare routine recording of self harm in clinical records in previous studies. The authors have appropriately identified the use of health informatic Natural Language Processing (NLP) techniques to be able to access much larger sets of notes and datasets over time to identify self harm and suicidality in clinical Electronic Health Records (EHR) This study thus provides important information about the levels of self harm and suicidality in eating disorders, but also highlights the utility of using novel approaches to accessing clinical information. Introduction The introduction provides a clear rationale for the study by describing the relationship between suicidality and completed suicide and placing this within the context of an eating disorders population. It also appropriately describes the challenges in accessing sufficient information contained in the free text of clinical EHR about suicidality. Manual searching of sufficient EHR is not feasible, and the researchers have utilised NLP tools to overcome this. Methods. The study design and setting is appropriate to answer the research question, given population base of 2 million , and an EHR set with records dating back to 2006. The total of over 7000 patients provides a sufficiently large data set to explore the themes of interest. The inclusion criteria is determined by ICD codes. The primary outcomes of ‘patient reporting of self harm and suicidality ‘ are sufficiently discrete to be extracted by NLP, and the authors quote other work in this area using NLP. Figure 1 provides a validation diagram and shows how the NLP tools are validated. The performance of the NLP tools are appropriately tested against a ‘gold standard’ of coding on a smaller sample of randomly selected notes and inter rater agreement assessed against two

	independent coders. The number of records chose is sufficient to establish the gold standard. The gold standard coding occurs after the algorithm has been used on the chosen data set. • P 4 Line 47-52. There is no information about the initial development and testing of the NLP algorithm. Was the algorithm used the same as developed for previous research (refs 31,34,35) This should be made clear. Results The headline results are clearly presented; Table 3 gives a clear indication of the primary outcome measures in terms of algorithm derived ‘incidence’, and tables 4-6 providing information about univariate and multivariate analyses. Discussion The discussion outlines the key clinical findings and also places them within the context of the NLP methodology. The discussion of clinical findings is well referenced and usefully describes the possible relationships between ED, Comorbidity and suicidal and self harm behaviours. The strengths and limitations are appropriately described, including the importance of the NLP algorithm allowing access to information that is not available from coded and other structured data in the EHR.

REVIEWER	Leung, Yvonne University Health Network
REVIEW RETURNED	19-Jun-2021

GENERAL COMMENTS	The aim of this study was to evaluate two NLP approaches that identify both self-harm and suicidality in the EHR of patients who suffered from eating disorders. Authors compared the performance of the NLP tools against a gold standard set of manually annotated documents, using previously defined coding rules. They then used the tools to identify positive mentions of either self-harm or suicidality on a patient level, to evaluate the incidence of self-harm and suicidality. This is a very innovative study and has important implications for mental health search if we can identify a reliable way to identify medical conditions from medical notes using NLP approaches. However there are some places needing clarifications before it can be moved forward for publication. First, it was unclear what are the two NLP approaches the authors were comparing against the gold standard. Authors should list them out in the objectives. As the authors mentioned GATE and NLP applications, I am guessing GATE vs NLP algorithms? What are the mechanisms behind each approach? I am guessing GATE is a keyword matching with Part of Speech Tagging? But I am not sure about the other approach. Authors should detail the differences.

	Second, authors could include a diagram about the workflow or pipeline of the NLP approaches to illustrate how the texts and documents are processed. For example, what techniques did the algorithms use to pick up self-harm or suicidal behaviours from the open text fields? One approach contains a Bag of Words model, and the other contains keyword matching using medical dictionaries? How did it deal with negations if the doctor said "no suicidal behaviours were shown”? In this case, will the system be affected by typos? Authors should provide a description summary of the systems. Readers will like to learn about the coding rules of self-harm or suicide behaviours in the gold standard. As there many ways to describe self-harm, can authors provide examples of what expressions or phases were used by the doctors in the open text fields of the EHR? What kind of mistakes that the NLP applications made? Without this information it is hard to judge how good the NLP applications were. Lastly, can authors provide the coding guide including an example for each category in the confusion matrix? Before we are clear about these information it is hard to judge how reliable the incident rates are.

VERSION 1 – AUTHOR RESPONSE

#	**Comments**	**Author responses**
	Reviewer: 1	
	Thanks for asking me to review this interesting retrospective cohort study on risk factors for suicidality and self harm in people with eating disorders. It is useful research and will be of interest to the community. I have some minor comments and points to improve the manuscript. 1. In the results section of the abstract, it perhaps would be better and clearer to simply state the precision and recall of positive mentions and suicidality i.e. 0.80 and 0.82 and 0.96 and 0.94. .	We thank the reviewer for their comments about the topic area and their recommendation for publication. I have updated the abstract as follows: *The attribute agreements for precision of positive mentions of self-harm were 0.96 and for suicidality were 0.80;*

	The introduction is a little long. It would aid readability if it were a little shorter.	We thank the reviewer for this valuable comment and have edited to be more concise and readable.
2	It would be useful to know a bit more about the clinical records (e.g. what proportion were clinical letters, clinical notes, inpatient records etc.)	We analysed the data as 'event notes' within the clinical data, agnostic to whether they were created during an inpatient admission, during clinical follow up or a telephone call. I have updated the manuscript with the above information.
3	For transparency there needs to be more detail on the NLP algorithms, how were they written? Did they use a specific NLP system. I could not at first glance find much more detail in the supplied references	I have updated the reference (38) explaining the NLP tool and it's previous validation and use. The detail of how the NLP algorithm was written is beyond the scope of the paper and previously derived which I have now referenced. I have updated the text as follows: *The first application used rule-based linguistic processing to identify positive mentions of self-harm (SH) in clinical texts, this included historic and current episodes, but did not include self-harm ideation. The second application, also rule-based and using lexical resource, included suicidal ideation (SUI) of both a passive and active nature; both of these were recorded as a binary outcome. A detailed description of the development of both NLP tools used to identify mentions of self-harm and suicidality are described in previous studies (36)(35).*
4	What is the difference between a document which contains "positive suicidality" and a document which contains "relevant suicidality" - this needs to be explained in more detail.	I have included the below paragraph to explain further: *Any mention of self-harm within the document was coded as positive, negative and whether relevant or non-relevant, for example a positive code refers to the note referring to an act of self-harm by the individual, negative refers to a denial or negated act of self-harm. If the mention was about a friend or family member that was not relevant to the patient non-relevant was coded. (see Figure 1).*

5	For transparency the raw numbers used to calculate the figures in table 2 should be published either in the main manuscript or as supplementary data e.g. number of true positives etc.	I have included this under Table 2 within the results section.
6	. The "deprivation scores 1,2 and 3" need to be explained in more detail - how are they derived?	I have now explained this in the methods section. Please see the text below: ’ *The deprivation score was grouped into tertiles (33rd percentiles) and converted into a categorical variable. Previous studies have used this method of categorical definition using the same data source (2).*
7	There is no need to repeat the information displayed in tables 4,5 and 6 in text preceding the tables, please delete this to aid readability.	I have now re-written the start of the Results section to this: *Table 1 summarises the different types of ED diagnosis by age. The mean age was 26.0 (SD 11; range 10-90)*
8	It would be useful if the authors could elaborate on why they think that precision and F1 scores lower for negative mentions?	I have now explained this in the discussion. Please see the text below: *This is likely due to errors in the linguistic pre-processing needed to identify negation*.

9	I think it would also be worth mentioning or discussing in more detail the possible heterogeneity in the recording of mentions of suicide and self harm in clinical encounters. Are some clinical encounters or some healthcare professionals more likely to discuss self harm and suicide?	I have added this to the discussion to read as below: *Finally, the data relies on recording of suicidality and self-harm following a clinical encounter. This is likely to result in some heterogeneity at a document level, as some healthcare professionals may be more likely to discuss or record self-harm or suicidal thoughts depending on their level of experience, clinical background or their prior knowledge of the patient. However as there only needed to be one positive mention of self-harm or one positive mention of suicidality, at a patient level, the threshold was low for detection of either outcome.*
	Reviewer: 2	
1	Introduction The introduction provides a clear rationale for the study by describing the relationship between suicidality and completed suicide and placing this within the context of an eating disorders population. It also appropriately describes the challenges in accessing sufficient information contained in the free text of clinical EHR about suicidality. Manual searching of sufficient EHR is not feasible, and the researchers have utilised NLP tools to overcome this.	We thank the reviewer for their comments regarding the introduction.

12	Methods The study design and setting is appropriate to answer the research question, given population base of 2 million , and an EHR set with records dating back to 2006. The total of over 7000 patients provides a sufficiently large data set to explore the themes of interest. The inclusion criteria is determined by ICD codes. The primary outcomes of ‘patient reporting of self harm and suicidality ‘ are sufficiently discrete to be extracted by NLP, and the authors quote other work in this area using NLP. Figure 1 provides a validation diagram and shows how the NLP tools are validated. The performance of the NLP tools are appropriately tested against a ‘gold standard’ of coding on a smaller sample of randomly selected notes and inter rater agreement assessed against two independent coders. The number of records chose is sufficient to establish the gold standard. The gold standard coding occurs after the algorithm has been used on the chosen data set. ◦ • P 4 Line 47-52. There is no information about the initial development and testing of the NLP algorithm. Was the algorithm used the same as developed for previous research (refs 31,34,35) This should be made clear.	I have now explained in more detail about how the NLP tools were previously derived. *The first application used rule-based linguistic processing to identify positive mentions of self-harm (SH) in clinical texts, this included historic and current episodes, but did not include self-harm ideation. The second application, also rule-based and using lexical resource, included suicidal ideation (SUI) of both a passive and active nature; both of these were recorded as a binary outcome. A detailed description of the development of both NLP tools used to identify mentions of self-harm and suicidality are described in previous studies (36)(35).*

	Results The headline results are clearly presented; Table 3 gives a clear indication of the primary outcome measures in terms of algorithm derived ‘incidence’, and tables 4-6 providing information about univariate and multivariate analyses.	There are no suggested changes for the results.
	Discussion The discussion outlines the key clinical findings and also places them within the context of the NLP methodology. The discussion of clinical findings is well referenced and usefully describes the possible relationships between ED, Comorbidity and suicidal and self harm behaviours. The strengths and limitations are appropriately described, including the importance of the NLP algorithm allowing access to information that is not available from coded and other structured data in the EHR.	We thank the reviewer for their comments on the discussion section.
	Reviewer: 3	

	First, it was unclear what are the two NLP approaches the authors were comparing against the gold standard. Authors should list them out in the objectives. As the authors mentioned GATE and NLP applications, I am guessing GATE vs NLP algorithms? What are the mechanisms behind each approach? I am guessing GATE is a keyword matching with Part of Speech Tagging? But I am not sure about the other approach. Authors should detail the differences.	The NLP tools have been developed in previous studies and previously tested; which I have now updated the text to ensure this is more clear. This study then tested their performance, using manual annotations for comparison.
	◦ Second, authors could include a diagram about the workflow or pipeline of the NLP approaches to illustrate how the texts and documents are processed. For example, what techniques did the algorithms use to pick up self-harm or suicidal behaviours from the open text fields? One approach contains a Bag of Words model, and the other contains keyword matching using medical dictionaries? How did it deal with negations if the doctor said "no suicidal behaviours were shown"? In this case, will the system be affected by typos? Authors should provide a description summary of the systems. ◦	As the tools have been previously developed, this explanation is beyond the scope and beyond the objectives of the paper. We have now referenced the papers that outline the NLP tool development so that it would be possible for readers to access this information.

	◦ dictionaries? How did it deal with negations if the doctor said "no suicidal behaviours were shown”? In this case, will the system be affected by typos? Authors should provide a description summary of the systems. ◦ Readers will like to learn about the coding rules of self-harm or suicide behaviours in the gold standard. As there many ways to describe self-harm, can authors provide examples of what expressions or phases were used by the doctors in the open text fields of the EHR? ◦	This is beyond the scope of this paper; as the tools were previously developed.
	◦ What kind of mistakes that the NLP applications made? Without this information it is hard to judge how good the NLP applications were. ◦	We have outlined this in the strength and limitation section. As we took one mention of either self-harm or suicidality for each patient, there were multiple opportunities to capture the information required. Please see below: *Furthermore, given that EHRs include routine clinical data not primarily collected for research purposes, the study relies on clinician documentation which could include non-grammatical errors, jargon and idiosyncratic abbreviations; all of these could increase the chance of NLP misclassification (35). However, this was mitigated by using all documents available for each patient. Therefore, there were multiple opportunities to capture suicidality information to compensate for lack of sensitivity of the tool.*
	◦ Lastly, can authors provide the coding guide including an example for each category in the confusion matrix? Before we are clear about these information it is hard to judge how reliable the incident rates are. ◦	This is beyond the scope of this paper; as the tools were previously developed.

VERSION 2 – REVIEW

REVIEWER	Pickrell, W. Owen Swansea University, Swansea University Medical School
REVIEW RETURNED	13-Sep-2021

GENERAL COMMENTS	Thanks for addressing my comments. I only have one further comment (point 6 in your table). You explain that deprivation scores were grouped into tertiles which is fine but what deprivation score did you use please?

REVIEWER	Leung, Yvonne University Health Network
REVIEW RETURNED	14-Sep-2021

GENERAL COMMENTS	The only comment I have is that authors should include sufficient information on how to interpret the precision findings of the NLP tool. E.g. precision equals to 0.96 and 0.80. I understand that this new tool was developed previously by some other groups but the current study must include sufficient information for the readers to interpret the findings and ensure reproducibility.

VERSION 2 – AUTHOR RESPONSE

Reviewer: 1
Dr. W. Owen Pickrell, Swansea University

Thank you for your comment, I have now updated the text with a more detailed explanation. Please see below:

We used the 'multiple deprivation score' which is a small-area-level measure of socioeconomic status, based on the individual's address closest to the diagnosis of the eating disorder in the observation window, covering seven components: employment, income, education, health, barriers to housing and services, crime and the living environment with specific weightings. The index of multiple deprivation is a well-established measure that has been widely used as a regional indicator for socioeconomic status in previous studies; the scores are transformed into percentiles (1-100) with higher scores indicating greater deprivation. In this analysis the deprivation socre was grouped into tertiles.

Reviewer: 3
Dr. Yvonne Leung, University Health Network

Thank you for your comment, I have now updated the table 2 with the two studies that evaluated the tools and their attribute agreement scores. Please see below:

Table 2 ; Attribute agreements: attribute agreements reflect the comparison of the NLP tool output to the gold standard set of manually annotated documents. Annotations are document-level for suicidality and mention-level for self-harm. The results from the study that developed and evaluated the suicidality tool reported 0.58-0.72 precision, 0.70-0.87 for recall and 0.69-0.75 F1-score (35) and the results from the study that developed and evaluated the self-harm tool reported 0.88-0.96 precision, 0.88-0.96 recall and 0.88-0.96 F1 score (40).