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)	Mood instability is a common feature of mental health disorders and is associated with poor clinical outcomes
AUTHORS	Patel, Rashmi; Lloyd, Theodore; Jackson, Richard; Ball, Michael; Shetty, Hitesh; Broadbent, Matthew; Geddes, John; Stewart, Robert; McGuire, Philip; Taylor, Matthew

VERSION 1 - REVIEW

REVIEWER	Elisabeth Wells University of Otago, Christchurch New Zealand
REVIEW RETURNED	26-Jan-2015

GENERAL COMMENTS	Routinely collected data from large clinical practices is useful because it is based on unselected groups of patients (compared with those in clinical trials) and it provides evidence of outcomes. However the clinical interest of these studies is often limited because the predictors of outcomes are usually only the standard demographic variables of age, sex, marital status and sometimes employment status, supplemented by diagnosis which may not necessarily be assessed in any standard way. The authors of this paper have had access to anonymized clinical notes from a large mental health provider and they have chosen to use Natural Language Processing to extract one clinical feature - mood instability - from these notes and to look at the relationship between this feature and outcomes. A quick search in Medline focused on 'Natural Language Processing' shows many articles (around a hundred articles published in the last 12 months alone) but combining 'Natural Language Processing' with 'Psychiatry' yielded only 3 articles, none related to the kind of electronic record study reported in the submitted manuscript. Therefore it seems to me that this study is quite innovative, developing a partial Natural Processing Language tool operating on the word level and conceptual level for semi-automatic encoding of free-text electronic records. Furthermore the Supplementary tables 1-3 provide quite detailed reporting of the development of the tool, including inter-annotator agreement. The sample size is extremely large with 27,704 patients included for one year outcomes. Nonetheless I have some concern about loss to follow-up. It seems likely that 27,704 is the number of patients for whom they had data for at least one year and that these patients

	are detected because they had a 'date-last-seen' more than one year after first contact. Presumably the two-year and five-year patients are defined similarly. In cohort studies the number of patients at the onset should be included and loss-to-follow-up should be included in a figure based on the Strobe guidelines: Participants 13* (a) Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed (b) Give reasons for non-participation at each stage (c) Consider use of a flow diagram I am not suggesting that the analyses reported in this paper should be redone - they are sensible, appropriate and clearly reported. Instead what I would like to know is how much drop-out there was and if this is more than perhaps 5%, an analysis of who was lost to follow-up could be added, perhaps with another Supplementary Table. There one additional predictor which was surely available, namely whether patients were outpatients or inpatients when first assessed. It may be that the service providing the data (SLam) is entirely an outpatient service. If so that should be stated. The authors have chosen to present relative rather than absolute results within the paper itself. I found that I needed to go to the relevant Supplementary Tables (5,6,7, 9 and 10) before reading Tables 2 and 3 in the paper itself. At least the text indicates the need to do this. As five tables are allowed I wondered about moving some of the Supplementary Tables into the paper itself but there are too many of them unless some were combined. For example inpatient days, compulsory admission and hospital admissions could be combined into one table and the cumulative antipsychotic and non-antipsychotic mood stabilisers into another table. This paper is well laid out, clearly written and admirably free of typographical errors. It addresses a clinical issue and the findings of are of importance, indicating the need to ascertain mood instability in patients with diagnoses other than bipolar disorder and the need to find out what can be done to ameliorate the poorer outcomes associated with mood instability.

REVIEWER	Steven Marwaha Mental health and Wellbeing Warwick Medical School University of Warwick UK
REVIEW RETURNED	02-Feb-2015

GENERAL COMMENTS	Mood instability is a common feature of mental health disorders and is associated with poor outcomes

	Introduction: the aims at the end of this section would be aided by adding in timepoints.. eg mood instability is present at ?baseline- otherwise not very clear whether the aim was to examine the effect of MI prospectively / retrospectively / cross sectionally Methods: the search terms used by text hunter conceptualise mood instability as a very broad problem. This may be a function of the authors views or a function of the technology used for the study- either way this should be a point of discussion in the Discussion What proportion of the total dataset-sentences were the reference dataset generated from- it states 300 sentences- what proportion of the dataset is this and why was this number chosen as adequate. I'm not that clear why the authors refer to positive predictive values in the text on page 6 when they are describing validation of their methods. PPV usually relies on there being a gold standard – often a ICD/ DSM diagnosis but this isn't referred to. I think it would help clarity of this section, which is quite technical, to provide some further explanation of this Can the authors comment on whether ethical approval had been given by the participants on the use of their data. If not what is the ethical and legal framework through which the data was used. Statistical analysis The main explanatory variable at baseline is MI at admission and the main outcome is days spent in hospital. Whilst I am not a statistician it is my understanding that Cox regressions would have been the suitable methodology here. This seems to be the same for some of the secondary outcomes. If other regressions are being used then I have sometimes seen a time to event variable (outcome) been included instead. I may have got the wrong end of the stick, but if so this section needs to be better explained. Variables that I presume would be available and could reasonably be expected to influence their outcomes are drugs and alcohol use, previous deliberate self-harm, age of illness onset, extent / severity of manic / psychotic symptoms etc. As far as I can see these were not adjusted for? The authors perhaps could comment on this. Discussion It should be highlighted that the methods relies on the clinicians view of what is being said- ie the clinicians view that mood instability is being discussed and that they think it is important enough to write down. This will be governed by many factors, including whether they think they can do much about it. Whilst the reliance on clinician notes does have face validity, in my own experience, training and understanding of mood instability is quite poor. This methodology might explain the prevalence found in this study of secondary mental health services is so low- I would have expected it to be higher. Within this discussion it perhaps would be useful to consider how their results might be different or the same

	if using questionnaires or EMA methods, as their novel method adds an additional way to study MI. Page 10 anti-psychotic used for MI and then mood stabilisers increasingly used. I agree that this is one interpretation of their findings but without knowing the extent, duration of and functional consequences of the MI at the time it is also possible that these prescriptions were inappropriate or simply being used for the diagnostic groups for which they are licenced. I would recommend being more cautious with this statement (and paragraph)- Its very difficult to know from their data what the indications both formal and informal for these medications were. Summary: this study is a useful addition to the literature. Its main value is in confirming MI occurs trans-diagnostically and that it can influence a range of outcomes in a very large dataset. The main novelty is in that MI has been measured from clinical note entries using NLP. It should be published after revision

REVIEWER	Nick Craddock Cardiff University UK
REVIEW RETURNED	04-Feb-2015

GENERAL COMMENTS	This manuscript uses text analysis of computerized clinical records to seek associations between demographic and clinical variables and the occurrence of terms that suggest presence of mood instability in a defined period of time. Several associations are reported with the conclusion that outcome is generally worse for patients experiencing mood instability irrespective if diagnosis and that more attention should be paid to identifying and treating mood instability. The paper is potentially interesting but needs revisions as follows: 1) As the authors point out, the rate of mood instability identified in the patients in the study is low compared with previous self report studies. The authors need to discuss in more detail the possible biases in their data and how those might affect interpretation of their findings. For example, might clinicians who are seeing a patient with a more complex clinical picture be more likely to record mood instability? 2) Following from (1) above, it would be interesting to know the rate of mood instability identified in patients with a diagnosis of borderline personality disorder. Given that this is likely to be present in most of these patients, this would provide some "baseline" measure of how successfully the methodology identified mood instability. 3) The description of method in the abstract does not make sense and needs to be written more simply and clearly. It may require

	more than one sentence. 4) The authors need to be more careful in use of any language that suggests causality. Association is a first step that suggests the need for studies to test causality. Of course, on grounds of general clinical common sense we should encourage clinicians to perform thorough mental state assessments and record this accurately. Further, if symptoms are causing distress we should expect a clinician to provide appropriate treatment. However, the authors go farther than the data in using their findings to suggest that long-term outcome will improve with treatment of mood instability. That might very well be the case – but that is for another study to address.

VERSION 1 – AUTHOR RESPONSE

Reviewer Name J Elisabeth Wells
Institution and Country University of Otago, Christchurch
New Zealand
Please state any competing interests or state 'None declared': None declared
Please leave your comments for the authors below

Routinely collected data from large clinical practices is useful because it is based on unselected groups of patients (compared with those in clinical trials) and it provides evidence of outcomes. However the clinical interest of these studies is often limited because the predictors of outcomes are usually only the standard demographic variables of age, sex, marital status and sometimes employment status, supplemented by diagnosis which may not necessarily be assessed in any standard way. The authors of this paper have had access to anonymized clinical notes from a large mental health provider and they have chosen to use Natural Language Processing to extract one clinical feature - mood instability - from these notes and to look at the relationship between this feature and outcomes.

A quick search in Medline focused on 'Natural Language Processing' shows many articles (around a hundred articles published in the last 12 months alone) but combining 'Natural Language Processing' with 'Psychiatry' yielded only 3 articles, none related to the kind of electronic record study reported in the submitted manuscript. Therefore it seems to me that this study is quite innovative, developing a partial Natural Processing Language tool operating on the word level and conceptual level for semi-automatic encoding of free-text electronic records. Furthermore the Supplementary tables 1-3 provide quite detailed reporting of the development of the tool, including inter-annotator agreement.

/*Thank you for your supportive comments on the methods employed in our study.*/

The sample size is extremely large with 27,704 patients included for one year outcomes. Nonetheless I have some concern about loss to follow-up. It seems likely that 27,704 is the number of patients for whom they had data for at least one year and that these patients are detected because they had a 'date-last-seen' more than one year after first contact. Presumably the two-year

and five-year patients are defined similarly. In cohort studies the number of patients at the onset should be included and loss-to-follow-up should be included in a figure based on the Strobe guidelines:
Participants 13* (a) Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed
(b) Give reasons for non-participation at each stage
(c) Consider use of a flow diagram

I am not suggesting that the analyses reported in this paper should be redone - they are sensible, appropriate and clearly reported. Instead what I would like to know is how much drop-out there was and if this is more than perhaps 5%, an analysis of who was lost to follow-up could be added, perhaps with another Supplementary Table.

/*We obtained mood instability data on a total of 27,704 patients in our study. The study included all adult patients (between the ages of 16 and 65 years) who presented to SLaM between 1st April 2006 and 31st March 2013 with a psychotic, affective or personality disorder. Outcome data were obtained up to 31st March 2014. This period was chosen in order to allow for at least one year of follow-up data to be analysed for all participants from the point of first presenting to SLaM. We also analysed data up to five years following presentation to SLaM. However, as you have pointed out, only a proportion of the 27,704 participants had five years of follow-up available, depending on when they first presented to a SLaM clinical service.

Presenting between 1st April 2006 and 31st March 2012: 2 years follow-up (n=24848)

Presenting between 1st April 2006 and 31st March 2011: 3 years follow-up (n=21188)

Presenting between 1st April 2006 and 31st March 2010: 4 years follow-up (n=17130)

Presenting between 1st April 2006 and 31st March 2009: 5 years follow-up (n=13032)

We have amended the "Participants" section of the Methods (Page 6, paragraph 1) to clarify this.

With respect to loss to follow-up, it was not possible to accurately determine the degree to which patients were lost from routine clinical follow-up during the period of the study. This is because patients may have been discharged from clinical services during the period of the study for a number of reasons including improvement in symptoms (i.e. planned discharge to primary care), disengagement from services and moving outside the catchment area of SLaM. It was not possible to extract reason for discharge as this information was not routinely documented. This is a limitation of analysing routine healthcare data in comparison to data from an observational or interventional study involving prospective patient recruitment where there is a standardised plan for follow-up, allowing for attrition to be measured. We have highlighted this limitation in the discussion section (page 11, paragraph 5).*/

There one additional predictor which was surely available, namely whether patients were outpatients or inpatients when first assessed. It may be that the service providing the data (SLam) is entirely an outpatient service. If so that should be stated.

/*SLaM includes both inpatient and outpatient services. 3221 patients (11.6%) presented to inpatient services when first assessed. We have updated the "Participants" section of the Methods with this information (page 6, paragraph 1).*/

The authors have chosen to present relative rather than absolute results within the paper itself. I found that I needed to go to the relevant Supplementary Tables (5,6,7, 9 and 10) before reading Tables 2 and 3 in the paper itself. At least the text indicates the need to do this. As five tables are allowed I wondered about moving some of the Supplementary Tables into the paper itself but there are too many of them unless some were combined. For example inpatient days, compulsory admission and hospital admissions could be combined into one table and the cumulative antipsychotic and non-antipsychotic mood stabilisers into another table.

/*Thank you for your helpful suggestion. We have restructured the tables in the main manuscript and supplementary material in order to allow absolute results to also be presented in the main manuscript.*/

This paper is well laid out, clearly written and admirably free of typographical errors. It addresses a clinical issue and the findings of are of importance, indicating the need to ascertain mood instability in patients with diagnoses other than bipolar disorder and the need to find out what can be done to ameliorate the poorer outcomes associated with mood instability.

/*Thank you.*/

Reviewer Name Dr Steven Marwaha
Institution and Country Mental health and Wellbeing
Warwick Medical School
University of Warwick
UK
Please state any competing interests or state 'None declared': non declared
Please leave your comments for the authors below

Mood instability is a common feature of mental health disorders and is associated with poor outcomes

Introduction: the aims at the end of this section would be aided by adding in timepoints.. eg mood instability is present at ?baseline- otherwise not very clear whether the aim was to examine the effect of MI prospectively / retrospectively / cross sectionally

/*In order to clarify our aims to investigate the association of mood instability at presentation with subsequent clinical outcomes, we have amended the introduction to state, "We tested the hypothesis that mood instability is present across a wide range of mental disorders at presentation to mental health services…" (page 5, paragraph 3).*/

Methods: the search terms used by text hunter conceptualise mood instability as a very broad problem. This may be a function of the authors views or a function of the technology used for the study- either way this should be a point of discussion in the Discussion

/*We performed a search in free text clinical records for the two words before or after any instance of "mood", "affect" or "emotion". We then chose modifier words to be included in each NLP application based on those which implied "instability" of each construct. The modifier words were reviewed by RP, TL and MT to agree on which to include. This method was chosen in order to develop NLP applications for our study which were relevant to the data upon which they were applied. We have updated the Methods and Discussion section to explain the rationale for using this data-driven NLP development method for our study (page 6, paragraph 3; page 10, paragraph 1).*/

What proportion of the total dataset-sentences were the reference dataset generated from- it states 300 sentences- what proportion of the dataset is this and why was this number chosen as adequate. I'm not that clear why the authors refer to positive predictive values in the text on page 6 when they are describing validation of their methods. PPV usually relies on there being a gold standard – often a ICD/ DSM diagnosis but this isn't referred to. I think it would help clarity of this section, which is quite technical, to provide some further explanation of this

/*Supplementary table 2 provides further details of the number of references sentences for each application and the total number of sentences which were analysed by the NLP applications. Around 300 reference sentences were manually annotated by two human annotators (RP and TL). This is based on methods used in previous studies for information extraction from mental health records (which we have referenced in our updated manuscript):

Jackson, R., Ball, M., Patel, R., Hayes, R., Dobson, R., & Stewart, R. (2014). TextHunter - A User Friendly Tool for Extracting Generic Concepts from Free Text in Clinical Research. Proceedings of the American Medical Informatics Association, 729–738. doi:10.13140/2.1.3722.9121

The term "annotation" refers to classifying whether a sentence represents the presence or absence of construct of interest (in this study, instability of mood/affect/emotion). The human-annotated reference sentences formed the "gold" standard upon which the NLP applications were tested to generate precision (positive predictive value) and recall (sensitivity) statistics. We have added a supplementary figure to better explain the definition of the precision and recall statistics in the context of NLP application development.*/

Can the authors comment on whether ethical approval had been given by the participants on the use of their data. If not what is the ethical and legal framework through which the data was used.

/*We have added further details on the ethical framework for our study in the methods section (page 6, paragraph 2).*/

Statistical analysis
The main explanatory variable at baseline is MI at admission and the main outcome is days spent in

hospital. Whilst I am not a statistician it is my understanding that Cox regressions would have been the suitable methodology here. This seems to be the same for some of the secondary outcomes. If other regressions are being used then I have sometimes seen a time to event variable (outcome) been included instead. I may have got the wrong end of the stick, but if so this section needs to be better explained.

/*We chose to analyse number of days spent in hospital as the primary outcome as this represents an important measure of illness severity as well as cost to healthcare services whereas time to hospital admission (in a Cox regression analysis) does not distinguish between patients who have short vs long periods of stay in hospital. We have updated the methods section to better explain the rationale for our choice of analysis methods (page 7, paragraph 4).*/

Variables that I presume would be available and could reasonably be expected to influence their outcomes are drugs and alcohol use, previous deliberate self-harm, age of illness onset, extent / severity of manic / psychotic symptoms etc. As far as I can see these were not adjusted for? The authors perhaps could comment on this.

/*We agree that these are important variables which could have influenced outcomes. Unfortunately, it was not possible to extract data on these variables from routine electronic health records. We have highlighted this limitation (and the need for further study into these factors in relation to mood instability) in the discussion section (page 11, paragraph 4).*/

Discussion
It should be highlighted that the methods relies on the clinicians view of what is being said- ie the clinicians view that mood instability is being discussed and that they think it is important enough to write down. This will be governed by many factors, including whether they think they can do much about it. Whilst the reliance on clinician notes does have face validity, in my own experience, training and understanding of mood instability is quite poor. This methodology might explain the prevalence found in this study of secondary mental health services is so low- I would have expected it to be higher. Within this discussion it perhaps would be useful to consider how their results might be different or the same if using questionnaires or EMA methods, as their novel method adds an additional way to study MI.

/*We agree that the use of routinely recorded clinical data may have resulted in a lower prevalence of mood instability in our study compared to data obtained from direct patient interview using a structured questionnaire. However, the benefit of analysing routinely recorded clinical data is the representativeness to everyday clinical practice. The clinicians recording information were agnostic to the possibility that their data would be later analysed using the NLP techniques employed in our study, and were not specifically seeking to elicit symptoms of mood instability. It is therefore likely that while clinicians may not have comprehensively documented symptoms in clinical records, where they were documented, it was because they were perceived to be clinically relevant. While it is not possible to establish whether this is the case, we have updated the discussion section to highlight the importance of future studies to investigate the degree to which mood instability is routinely documented in comparison to data obtained from structured questionnaires (page 10, paragraph 2).*/

Page 10 anti-psychotic used for MI and then mood stabilisers increasingly used. I agree that this is one interpretation of their findings but without knowing the extent, duration of and functional consequences of the MI at the time it is also possible that these prescriptions were inappropriate or simply being used for the diagnostic groups for which they are licenced. I would recommend being more cautious with this statement (and paragraph)- Its very difficult to know from their data what the indications both formal and informal for these medications were.

/*This is a fair point – while we showed an association of mood instability with antipsychotic and non-antipsychotic mood stabiliser treatment, this does not necessarily demonstrate a causal association and there may be other factors which resulted in this association (such as medications being indicated to treat underlying disorders which are associated with mood instability). We have updated the discussion section accordingly (page 11, paragraph 2).*/

Summary: this study is a useful addition to the literature. Its main value is in confirming MI occurs trans-diagnostically and that it can influence a range of outcomes in a very large dataset. The main novelty is in that MI has been measured from clinical note entries using NLP. It should be published after revision

/*Thank you for your supportive comments.*/

Reviewer Name Nick Craddock
Institution and Country Cardiff University
UK
Please state any competing interests or state 'None declared': None declared
Please leave your comments for the authors below

This manuscript uses text analysis of computerized clinical records to seek associations between demographic and clinical variables and the occurrence of terms that suggest presence of mood instability in a defined period of time. Several associations are reported with the conclusion that outcome is generally worse for patients experiencing mood instability irrespective if diagnosis and that more attention should be paid to identifying and treating mood instability. The paper is potentially interesting but needs revisions as follows:
1) As the authors point out, the rate of mood instability identified in the patients in the study is low compared with previous self report studies. The authors need to discuss in more detail the possible biases in their data and how those might affect interpretation of their findings. For example, might clinicians who are seeing a patient with a more complex clinical picture be more likely to record mood instability?

/*We agree – as stated in the response to Dr Marwaha, the use of routinely recorded data biases documentation towards symptoms which are perceived to be clinically relevant. We have highlighted this point in our revised discussion (page 10, paragraph 2).*/

2) Following from (1) above, it would be interesting to know the rate of mood instability identified in patients with a diagnosis of borderline personality disorder. Given that this is likely to be present in

most of these patients, this would provide some “baseline” measure of how successfully the methodology identified mood instability.

/*17.8% of patients with personality disorder had documented mood instability. This compared with 22.6% of patients with bipolar disorder and 15.5% of patients with schizophrenia (Table 1). The association of documented symptoms and diagnosis is complex and multifactorial. It is possible that consideration of a particular diagnosis may have influenced the degree to which symptoms of mood instability were documented. However, it is noteworthy that in our study, mood instability was represented across a range of diagnoses. One way to address this in future studies would be to perform standardised diagnostic and mood instability interviews and compare these with the data recorded in electronic health records. This would help to establish the degree to which routinely documented symptoms are influenced by the clinician’s working diagnosis (and vice versa). We have updated the Discussion section to explore this area further (page 10, paragraph 2).*/

3) The description of method in the abstract does not make sense and needs to be written more simply and clearly. It may require more than one sentence.

/*We have reworded the abstract (within the 300 word limit) to clarify the exposure and outcome measures investigated in our study.*/

4) The authors need to be more careful in use of any language that suggests causality. Association is a first step that suggests the need for studies to test causality. Of course, on grounds of general clinical common sense we should encourage clinicians to perform thorough mental state assessments and record this accurately. Further, if symptoms are causing distress we should expect a clinician to provide appropriate treatment. However, the authors go farther than the data in using their findings to suggest that long-term outcome will improve with treatment of mood instability. That might very well be the case – but that is for another study to address.

/*We accept that there are several limitations of observational studies, particularly those which analyse routinely recorded clinical data, and that it is not possible to infer causation based solely on the association of a predictor with outcome. However, we have not made any statement in our manuscript to suggest that “long-term outcome will improve with treatment of mood instability”. Instead, we state that “direct treatment of this symptom [mood instability], irrespective of a patient’s working diagnosis, could have considerable health economic benefits”. We also suggest that our findings “highlight the need for interventional studies across a range of mental disorders to better understand which pharmacological and psychosocial interventions are most successful in reducing the impact of mood instability.” In so doing, we have not made any statement which implies that treatment of mood instability would necessarily result in better clinical outcomes as we agree that this cannot be inferred from our findings and requires further investigation with interventional studies.*/