PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Results of a prospective, mixed methods study to assess feasibility, acceptability, and effectiveness of TRIumPH (Treatment and Recovery In PsycHosis), an Integrated Care Pathway for Psychosis, compared to usual treatment.
AUTHORS	Rathod, Shanaya; Thorne, Kerensa; Graves, Elizabeth; Phiri, Peter; Asher, Carolyn; Griffiths, Alison; Read, Tracy; Kingdon, David

VERSION 1 – REVIEW

REVIEWER	Narsimha PInninti Rowan University : SOM. USA. I personally know two of the authors and have collaborated on scientific projects with them.
REVIEW RETURNED	30-Sep-2019

GENERAL COMMENTS	Pathways for care are essential for psychiatry to standardize practice and improve uptake of evidence based practices. Demonstration of feasibility and efficacy of such pathways is helpful to improve the practice of psychiatry

REVIEWER	Margda Waern Univ of Gothenburg, Sweden
REVIEW RETURNED	16-Dec-2019

GENERAL COMMENTS	This is a pragmatic observational study that employs mixed methods to assess numerous outcomes in mental health services with and without implementation of an integrated pathway for care for persons with psychosis. Improvement of health care for persons with psychotic illness is a timely issue, and kudos to the authors for taking on this ambitious intervention project! Complex interventions are never easy to evaluate and I have some some questions and some suggestions that might improve the paper. Abstract Upon reading the abstract I felt unsure of a) how data were collected, b) what, exactly were the primary effect variables? and c) what did the authors actually find? The conclusion is a simple reiteration of the second sentence of the results section. Strengths and Limitation points The listed points echo to a large degree the points that were listed in the original study protocol. As an alternative, some considerations specific to the current study could be listed. For example, the large amount of missing patient data is a serious

	limitation. Further, it seems unclear as to whether the study was actually sufficiently powered considering the numerous outcomes that were analysed. Introduction This section reads clearly. Some of the references seem a bit old. Just as an example, regarding life expectancy, the reference is from 1996. A more up-to-date reference would be advantageous. There is an emphasis on DUP in the introduction which would make sense if DUP were one of the study outcomes, but I did not localize such data in the results section. It would be helpful if the aim could be a bit more specific in terms of variables of main interest in terms of effectiveness. Methods Would it be possible to see more clinical baseline information regarding the individuals in each study arm? For example, the diagnostic panorama? PANSS scores? Level of psychosocial function? This kind of information is necessary if the intervention is to be replicated in some other setting. It is stated in the published study protocol that differences between the implementation arm and the comparator arm will be analysed using linear regression. It seems that the authors have chosen not to do so here. If so, an explanation would be helpful. Results This section is a rather massive and a bit difficult to digest. The number of analyses is great, which is particularly problematic as no power calculation is presented. According to the original study protocol, the study question was “Is implementation of TRIumPH integrated pathway in services feasible and acceptable and does it reduce duration of untreated psychosis, hospital stay and severity of symptoms at 12 and 24 months?” Findings of qualitative analyses regarding feasibility and acceptability are very briefly described but none of the listed quantitative variables seem to be in focus in the present paper. Perhaps these findings have been presented elsewhere? If so, it would be appropriate to inform about this in the Intro section. What do the authors consider to be the primary outcome measure(s) in the current paper? This should be made clear both in the abstract, methods section, and in the discussion section as well. Discussion This section is thin and needs to be deepened. Upon reading, I was still not really sure what the most important findings were. As mentioned above there could be a power issue here. Also, the discussion section includes few references to the existing literature. Minor points Table 2: what is DNA? (abbreviations may be obvious for UK readers but not for an international readership) Is the section on safety assessments really necessary? It was made clear already in the study protocol that no safety assessments were included.

REVIEWER	Prof. Dr. Martin Lambert Head of Psychoses Division, Integrated Care, Early Detection and Intervention Centre for Mental Disorders (FePS), Center of Psychosocial Medicine, Department of

	Psychiatry and Psychotherapy, University Medical Center Hamburg, Germany
REVIEW RETURNED	17-Dec-2019

GENERAL COMMENTS	BMJ Open Manuscript number: bmjopen-2019-033711 Title: Results of a multicentre study to assess feasibility, acceptability, and effectiveness of TRIumPH (Treatment and Recovery In PsycHosis): Integrated Care Pathway for Psychosis Article Type: Original article Review Abstract In the abstract, the study design, the collected parameters as well as the statistical methods are missing in the method part. In the part on the results, concrete results are missing. Due to the methods and results not mentioned in the abstract, the conclusion cannot be comprehended. Strengths and weaknesses After reading the document the mentioned strengths and weaknesses are not comprehensible. Methodology The study design is not so easy to understand. What does "pragmatic" mean? The setting should also be better described. Why did two regions become one? Have I understood correctly that in one region there are four teams, in the other first two, then one team? If TAU is the comparison, why does TAU have an EIP service? The paragraph tries to explain this in more detail, but it doesn't make it any clearer. In the chapter on the investigation methods used, parameters are mentioned but not defined. In such a chapter all parameters belong exactly defined, especially if they are scales. Results From my point of view, the first table mentioned, Table 1, contains far too many parameters. This is really hard to understand. Why are parameters divided into years? Why were the demographic data not compared statistically? What applies to Table 1 also applies to Table 2, where one should limit oneself even more and above all pay attention to multiple testing, i.e. compare only the most important parameters statistically. The same applies to Tables 3, 4 and 5 and subsequently also Tables 6. Discussion The discussion then reveals that this is not a prospective study, but a file study. This must be included in the title! The limitations of file studies with regard to reliability and validity must be explained in detail in the part on the limitations of the study, especially with regard to the comparability of the measurement by different investigators.

REVIEWER	Abigail Wright Massachusetts General Hospital & Harvard Medical School USA
REVIEW RETURNED	17-Dec-2019

GENERAL COMMENTS	Thank you for the opportunity to review this paper. This paper an interesting review of EIP pathways and was considered by the authors to be the first study to evaluate psychosis care pathways. I have major concerns about this article. Namely, the general language and writing very difficult to understand and to have faith in the results; supported by the little evidence in the actual paper (it is within tables which only provide p values). It is also very difficult to understand the paper without having read the protocol paper (this paper should be a standalone manuscript which is clear to understand itself), and much of the writing is actually directly copied from the earlier manuscript. It was also only clear in the study limitations that this was a retrospective study of medical records. This needs to be clearer. I have included some additional points below. Major comments • Abstract is very difficult to understand what was done in this study and what were the findings. • General language and writing needs to be clearer. • There are many missing references in the introduction. Please ensure every claim is supported with evidence, e.g., page 4 line 29; page 5 line 21. • The overall introduction is unstructured, without a clear flow of argument it is difficult to understand the research questions being asked in this study. • Parts of the introduction (and many parts of the manuscript) are directly copied from their earlier publication in this journal: https://bmjopen.bmj.com/content/bmjopen/6/12/e012751.full.pdf • "The Access and Waiting Time target was launched in April 2016 after the study started and will have influenced treatment as usual in each area". In what way? Please expand on this as it seems like a key issue. Results: • The paper states: "reduced significantly compared with baseline, from median 11 to 7 days, and from 20 to 11 days respectively ($p<0.0001$)". Why do you report median? This is very confusing because the p value is based on mean (not median). Is this median in the following paragraph too? • Please provide p values when you conduct any different test (regardless of outcome) • A key approach for the results is to provide as much numerical or empirical evidence as possible, without interpretation at that point. It appears there is a lot of interpretation of the results without actual numbers, e.g., see page 14, line 22 "substantial increase" without numbers and this is consistent with the overall results section. Discussion • In the discussion you state that there were pre-existing differences in the sites – is there a way to control for this? • The overall discussion could do a better job of consolidating the information and presenting this in a clearer way. I had a difficult time truly understanding the outcomes. Minor comments • The first paragraph of the introduction is not very strong. It appears there are missing links between the sentences. I would suggest starting with schizophrenia as 8th leading cause of DALYs (explain what this means in more detail). Then state the individual cost of schizophrenia (life expectancy), family burden (explain) then societal cost. • In the second paragraph you state that DUP ranges from 364-721 days. This is old research and is now incorrect and will need to be changed. DUP in the United Kingdom previously reported a mean

	DUP of 74 days (Reichert & Jacobs, 2018) but this is even lower now as the UK has a two-week rule from referral to treatment (as you mention below) and, equally, Australia have very low DUP from their EIP system. • Be consistent with UK or USA spelling. • I cannot see diagram one and the intervention needs more detail – not many readers will refer to the other paper. • You do not need the information about sample size. • "The development of the pathway was tailored to the needs of people with psychosis". How? Please provide more information and it is very difficult to understand. • Descriptive statistics are reported in a strange fashion. Please provide numerical information, e.g., % of males, or just refer them to the table. • Associated with this, please try to provide as much as evidence as possible with the data throughout all the results. E.g., instead of stating "most commonly", provide the actual %, and show evidence for the rise in patients • Please provide age standard deviation. Finally, with revisions, this paper could be a great addition to the field and demonstrating the key importance of EIP/FEP programs.

VERSION 1 – AUTHOR RESPONSE

3. Reviewer: 2
3.1. Abstract

3.1.1. Upon reading the abstract I felt unsure of
a) How data were collected
This has been clarified for both the quantitative and qualitative data collected
Quantitative data consisted of routinely collected clinical data retrieved from patient records to asses whether the implementation of TRIumPH achieved better concordance to NICE standards. These included: time to access services, physical health assessments, clinical outcomes based timeliness of delivery, acute data. The controlled trial has evaluated the effect of TRIumPH (Intervention) with Care As Usual (Comparator). Qualitative measures consisted of questionnaires, interviews and focus groups to assess acceptability and satisfaction. Outcome measures were compared within the baseline, year 1 and year 2 cohorts and between the two sites (implementation and comparator).

b) What, exactly were the primary effect variables ?
This has been clarified in the objectives and outcome measures section of the paper.

c) what did the authors actually find? The conclusion is a simple reiteration of the second sentence of the results section.
The results sectioned has been substantially amended to clarify this query.

3.1.2. Strengths and Limitation points: The listed points echo to a large degree the points that were listed in the original study protocol. As an alternative, some considerations specific to the current study could be listed. For example missing data is a limitation.
This section has been re-written completely and there is no duplication with the protocol paper. Missing data has been added as a limitation.

3.1.3. Further, it seems unclear as to whether the study was actually sufficiently powered considering

the numerous outcomes that were analysed .
This has been clarified in the paper with a section on sample size that reads:
As this was a prospective and pragmatic study, no a priori power and sample size calculations were performed or required as routinely collected and available data for all patients and staff during the study period was used.

3.2. Introduction
This section reads clearly. Some of the references seem a bit old. Just as an example, regarding life expectancy, the reference is from 1996. A more up-to-date reference would be advantageous.
This has been updated:
Chang CK, Hayes RD, Perera G, Broadbent MT, Fernandes AC, Lee WE, et al. Life expectancy at birth for people with serious mental illness and other major disorders from a secondary mental health care case register in London. PLoS One. 2011; 6(5): e19590.

3.2.1. There is an emphasis on DUP in the introduction which would make sense if DUP were one of the study outcomes, but I did not localize such data in the results section.
The relevance of this has been further clarified, linking DUP to the need for speedier access to services and treatment which is a core part of the pathway and national standards implemented in 2016. This timeliness of access as a key measure has been made more clear throughout the methods and results

3.3. It would be helpful if the aim could be a bit more specific in terms of variables of main interest in terms of effectiveness.
This has been clarified by breaking down the outcome measures into timeliness of access, NICE recommended treatments offered and clinical outcomes.

3.4. Methods
3.4.1. Would it be possible to see more clinical baseline information regarding the individuals in each study arm? For example, the diagnostic panorama? PANSS scores? Level of psychosocial function? This kind of information is necessary if the intervention is to be replicated in some other setting.
Due to the pragmatic nature of the study, only routinely collected clinical data recorded in the medical records were available. This meant that HoNoS scores were used to assess symptomatology but due to the large amounts of missing data, full exploration of these was not possible. Patients in EIP services often do not receive a diagnosis early on as they may have a prolonged assessment and on occasion diagnosis can remain unclear. PANSS was not conducted as we only collected routinely collected data as explained and as would be the norm in clinical settings.

3.4.2. It is stated in the published study protocol that differences between the implementation arm and the comparator arm will be analysed using linear regression. It seems that the authors have chosen not to do so here . If so, an explanation would be helpful.
In the protocol we mentioned that in addition to analysing data by comparing means (or ranks) or proportions (depending on the data), we shall additionally use regression to compare groups (for effect sizes and predictive models). However the extent of the missing data for many outcome variables meant that we did not have enough numbers to test such models in valid and reliable way. Thus, we decided to restrict ourselves to exploratory analysis rather than measuring effects and developing models using regression approach.

3.5. Results
3.5.1. This section is a rather massive and a bit difficult to digest.
We have amended the results section to make it easier to read. Quantitative outcome measures have been broken down as noted above to aid clarity throughout the methods and results section.

3.5.2. The number of analyses is great, which is particularly problematic as no power calculation is presented .
No power calculations were performed due to the exploratory nature of the study. This is noted as a key limitation. Further replication and perhaps hypothesis testing in future research may be needed to ensure the validity of results. However this pragmatic trial gives a good indication as to the areas of improvement such a pathway may influence, the feasibility and acceptability of such an implementation.

3.5.3. According to the original study protocol, the study question was “Is implementation of TRIumPH integrated pathway in services feasible and acceptable and does it reduce duration of untreated psychosis, hospital stay and severity of symptoms at 12 and 24 months?” Findings of qualitative analyses regarding feasibility and acceptability are very briefly described but none of the listed quantitative variables seem to be in focus in the present paper. Perhaps these findings have been presented elsewhere? If so, it would be appropriate to inform about this in the Intro section.
We have rewritten the results section and it clarifies the outcomes measures and what they indicate. The timeliness of access impacts on DUP, outcomes and number and length of hospital stays. These are now reported. Severity of symptoms is addressed using the HoNoS scores.

3.5.4. What do the authors consider to be the primary outcome measure(s) in the current paper? This should be made clear both in the abstract, methods section, and in the discussion section as well.
We have rewritten these sections as suggested.

3.6. Discussion

3.6.1. This section is thin and needs to be deepened. Upon reading, I was still not really sure what the most important findings were. As mentioned above there could be a power issue here. Also, the discussion section includes few references to the existing literature.
We have incorporated the changes and added some references as suggested.

3.7. Minor points
3.7.1. Table 2: what is DNA ? (abbreviations may be obvious for UK readers but not for an international readership)
DNA term was only used in table 2 therefore abbreviation removed and full explanation added as a replacement

3.7.2. Is the section on safety assessments really necessary? It was made clear already in the study protocol that no safety assessments were included.
Reporting on the lack of AEs/SAEs due to the nature of the study is considered an important point. So this has been shorted and clarified within the Research ethics approval and safety section

4. Reviewer: 3
4.1. Abstract
4.1.1. In the abstract, the study design, the collected parameters as well as the statistical methods are missing in the method part. In the part on the results, concrete results are missing. Due to the methods and results not mentioned in the abstract, the conclusion cannot be comprehended.
We have added all the above suggested parameters but due to constraints of 300 words it has had to be brief.
The results sections has been expanded.

4.1.2. Strengths and weaknesses
After reading the document the mentioned strengths and weaknesses are not comprehensible .
As mentioned earlier, we have re written this section

4.2. Methodology
4.2.1. The study design is not so easy to understand. What does "pragmatic" mean?
Pragmatic research methods align with the pragmatic epistemological view that to best understand the world around us we need to explore the usefulness of research outcomes, rather than how a phenomenon is researched. In real world research this means taking the best and most feasible approach to a answer question (for example using non-randomised, medical note data) rather than designing a complex study that may not be suitable in initial or exploratory research and/or give very similar answers to a pragmatic study. A reference has been added to the paper for readers who may wish to have further clarity on this point.

4.2.2. The setting should also be better described. Why did two regions become one? Have I understood correctly that in one region there are four teams, in the other first two, then one team? If TAU is the comparison, why does TAU have an EIP service? The paragraph tries to explain this in more detail, but it doesn't make it any clearer.
This has been further clarified. In the NHS there is a constant reorganisation of services. As this is a real world pragmatic trial, the normal changes in systems have not been interfered with.

4.3. In the chapter on the investigation methods used, parameters are mentioned but not defined. In such a chapter all parameters belong exactly defined, especially if they are scales.
Most quantitative outcome measures where based on medical record data. The parameters used to define these outcomes, such as days/weeks, interventions offered or not etc, have been added to the section on outcome measures. HoNoS as a measure of symptomatology has had a reference added but not further described as it is a nationally used measure in the UK and there are multiple measures being described.

4.4. Results
4.4.1. From my point of view, the first table mentioned, Table 1, contains far too many parameters. This is really hard to understand. Why are parameters divided into years?
This section has been amended and has been further clarified in the paper by explaining that the participants are split into and compared between cohorts, baseline, year 1 and year 2 to explore the impact of the implementation of the pathway at the beginning of year 1.

4.4.2. Why were the demographic data not compared statistically?
What applies to Table 1 also applies to Table 2, where one should limit oneself even more and above all pay attention to multiple testing, i.e. compare only the most important parameters statistically. The same applies to Tables 3, 4 and 5 and subsequently also Tables 6.
The authors acknowledge that there is a lot of data and as such have tried to present it clearly. However, all data analysed is reported so that reviewers and readers can be aware of the exploratory nature of the study and we acknowledge that the conclusions should be read in light of this.

4.5. Discussion
The discussion then reveals that this is not a prospective study, but a file study . This must be included in the title! The limitations of file studies with regard to reliability and validity must be explained in detail in the part on the limitations of the study, especially with regard to the comparability of the measurement by different investigators.
This is a prospective study, data collection started in 2015 at the beginning of the implementation of the pathway and qualitative data was collected at three points, baseline, 12 and 24 months. This has been further clarified throughout the paper. The use of mixed methods, qualitative and quantitative (from medical records) has been added to the title.
Use of data collected from the medical records and use of one dedicated staff member to complete collection is detailed in the study limitations section.

5. Reviewer: 4
5.1. Namely, the general language and writing very difficult to understand and to have faith in the results; supported by the little evidence in the actual paper (it is within tables which only provide p values).
The language and writing style has been reviewed. With regards to the use of p values only, the authors feel that the addition of confidence intervals or effect sizes, to an already large amount of data presented would not add much, but may rather create an even more complex picture. Rather, it has been emphasised throughout that this is an exploratory study and therefore any results should be interpreted with caution and replicated.

5.2. It is also very difficult to understand the paper without having read the protocol paper (this paper should be a standalone manuscript which is clear to understand itself), and much of the writing is actually directly copied from the earlier manuscript . It was also only clear in the study limitations that this was a retrospective study of medical records. This needs to be clearer.
This is a prospective study as data collection started in 2015 at the beginning of the implementation of the pathway and qualitative data was collected at three points, baseline, 12 and 24 months. This has been further clarified throughout the paper and title.
We have also included some elements from the protocol paper but are limited by word count.

5.3. I have included some additional points below.
Major comments
•Abstract is very difficult to understand what was done in this study and what were the findings.
The abstract has been revised to aid clarity and understanding

•General language and writing needs to be clearer.
The whole paper has been re-read and reviewed in light of these comments

•There are many missing references in the introduction . Please ensure every claim is supported with evidence, e.g., page 4 line 29; page 5 line 21.
These have been added
•The overall introduction is unstructured, without a clear flow of argument it is difficult to understand the research questions being asked in this study.
We have revised this to ensure there is flow.

•Parts of the introduction (and many parts of the manuscript) are directly copied from their earlier publication in this journal:
https://eur03.safelinks.protection.outlook.com/?url=https%3A%2F%2Fbmjopen.bmj.com%2Fcontent%2Fbmjopen%2F6%2F12%2Fe012751.full.pdf&data=01%7C01%7CD.Kingdon%40soton.ac.uk%7C23137058d63c41a26fb808d78f5f9033%7C4a5378f929f44d3ebe89669d03ada9d8%7C0&sdata=8UpG5cb5aElEBppm9GfToYQR6yudaoXbiAXdE9zLLM8%3D&reserved=0
We have now revised the paper to include relevant parts of the protocol paper as suggested by the previous reviewer. We are aware of the conflicting review on this point and have tried to get a balance.

•“The Access and Waiting Time target was launched in April 2016 after the study started and will have influenced treatment as usual in each area”. In what way? Please expand on this as it seems like a key issue.
This has been further clarified in the paper. In summary, as all NHS organisations would have to meet this standard, access times would change as a result in the TAU group as well.

5.4. Results:

5.4.1. The paper states: “reduced significantly compared with baseline, from median 11 to 7 days, and from 20 to 11 days respectively (p<0.0001)”. Why do you report median? This is very confusing because the p value is based on mean (not median). Is this median in the following paragraph too?
This is due to the data being non-parametric. We report median because the data is positively skewed (most waiting times are small). The p value is from Mann-Whitney U test as stated in the statistical analysis section, which is a non-parametric test.

5.4.2. Please provide p values when you conduct any different test (regardless of outcome)
All p values are reported. As noted in the table heading some p values relate to chi square tests exploring large tables for example 4x5 tables of data. P values were not calculated for some tests because the number of participants in some categories was low (<5) as noted in the statistical analysis section.

5.4.3. A key approach for the results is to provide as much numerical or empirical evidence as possible, without interpretation at that point. It appears there is a lot of interpretation of the results without actual numbers, e.g., see page 14, line 22 “substantial increase” without numbers and this is consistent with the overall results section.
We have now included all the results but also acknowledge as pointed out by the previous reviewer that this is a lot of data.

5.5. Discussion
5.5.1. In the discussion you state that there were pre-existing differences in the sites – is there a way to control for this?
This was a pragmatic, exploratory trial and therefore no controls or randomisation were in place. Results should be and are interpreted based on this. Reporting of demographics showed little difference between participants but this was not overly emphasised as other differences could be the staff at sites, data recording practises etc. This is noted in the study limitations. The original aim was to have four sites to allow for more robust findings however as noted two sites withdrew from the study due to inability to provide data. Again this is not uncommon in NHS organisations and a real world trial like this highlights the real problems.

5.5.2. The overall discussion could do a better job of consolidating the information and presenting this in a clearer way. I had a difficult time truly understanding the outcomes.
We have now revised this.

5.6. Minor comments
5.6.1. The first paragraph of the introduction is not very strong. It appears there are missing links between the sentences. I would suggest starting with schizophrenia as 8th leading cause of DALYs (explain what this means in more detail). Then state the individual cost of schizophrenia (life expectancy), family burden (explain) then societal cost.
This has been revised as suggested.

5.6.2. In the second paragraph you state that DUP ranges from 364-721 days. This is old research and is now incorrect and will need to be changed. DUP in the United Kingdom previously reported a mean DUP of 74 days (Reichert & Jacobs, 2018) but this is even lower now as the UK has a twoweek rule from referral to treatment (as you mention below) and, equally, Australia have very low DUP from their EIP system.
A newer reference from 2014 has been added, however as this study started in 2015 the references are used to explain the need and background of the current study. DUP is very likely to have been reduced based on the government initiatives mentioned in the paper.

5.6.3. Be consistent with UK or USA spelling .

This has been amended

5.6.4. I cannot see diagram one and the intervention needs more detail – not many readers will refer to the other paper.
We had uploaded the diagram. We have revised the paper to include information from the protocol paper although we are aware of the conflicting reviews from different reviewers.

5.6.5. You do not need the information about sample size.
This relates to the concern raised by other reviewers, and are aware of the conflicting reviews

5.6.6. “The development of the pathway was tailored to the needs of people with psychosis”. How? Please provide more information and it is very difficult to understand.
We have included some information, referenced and reported in detail in the protocol paper. It is not fully explained here due to lack of space and would be a replication. We are also aware of the conflicting reviews from different reviewers.

5.6.7. Descriptive statistics are reported in a strange fashion. Please provide numerical information, e.g., % of males, or just refer them to the table.
Numerical information has been added

5.6.8. Associated with this, please try to provide as much as evidence as possible with the data throughout all the results. E.g., instead of stating “most commonly”, provide the actual %, and show evidence for the rise in patients
Numerical information has been added

•Please provide age standard deviation.
Interquartile ranges have been reported to aid understand of the spread of data, however data was non-normal so SDs would not provide further clarity.

VERSION 2 – REVIEW

REVIEWER	Margda Waern Univ of Gothenburg, Sweden
REVIEW RETURNED	27-Jan-2020

GENERAL COMMENTS	Thank you to the authors; the paper is much improved. Still, there is further room for improvement. The bullet points are numerous and rather wordy; please consider refining to better highlight the main issues. Thank you for clarifying the study outcomes. Time between EIP referral and assessment is stressed as an important one. Looking at it from a non-UK perspective, I was positively surprised at the very short duration between EIP referral to EIP assessment, already at baseline! Would it be expected that a change from a median of 11 to 7 days actually have any clinically meaningful impact on DUP? Another outcome stressed in the study objectives is “clinical outcomes based on HoNOS scores”. If I understand correctly, there were 168 patients in the year 1 cohort at the comparator site, and HoNoS data were available for only 16 of these.

	Corresponding figures for the year 2 cohort data were 11 out of 102. Rather than retaining results of statistical analyses based on these extremely small subgroups, it would be more up front to simply state that it was not meaningful to test for changes among cohorts at the comparator sites due to the fact that 90% had missing data. Specifically relating to my reviewer point 3.6.1: The discussion section is improved but would benefit from further deepening. A few thoughts on this: Little attention is placed on the qualitative component of the study. The lack of qualitative data for the comparator site is of course problematic. The authors might consider strengthening the scientific contribution of this study component or, alternatively, consider saving the qualitative component for a separate study where results could be presented and discussed in more detail. The issue of site differences at baseline is mentioned in the discussion (and, importantly, as a bullet point), but there is no discussion on how this might have impacted differentially on outcomes at the two sites. Just as an example: Considering the severe nature of the illness, I found it odd that only 56% of the baseline cohort at the comparator site received any of the listed interventions. The proportion at the implementation site was more in line with what could be expected. Does this tell us something about baseline differences in services at the two sites? And how might this effect outcomes? Not all outcomes were positive for the intervention site. The significant decrease in the proportion with a crisis plan (paralleled by a significant improvement in the comparator site) came as a surprise, and could be discussed. The issue of multiple testing remains and could be specifically pointed out as a limitation. Minor points Abstract, first sentence in Results: This is not a result, but rather a description of a statistical method, which would be better placed in the methods section. Page 8, Setting, line 5: "who were working to principles..." I think perhaps they mean "who were working according to principles..." line 1 page 10 "limited to that routine collected" should probably read "routinely collected". page 11, line 3. Please show full name of scale the first time HoNoS is used in the manuscript text. Please check p value on page 20, third line: p=037 may be a typo? Page 24, last sentence: Please check wording "..reduction in patients admitted to inpatients and the subsequent reduction of relapse". Might they mean "admitted to inpatient wards"? Finally, (sorry, I missed this last time) it is stated in the text that there was no significant change in staff experiences across time

	points or between the sites. I could not find results of significance testing for these analyses.

REVIEWER	Abigail Wright Massachusetts General Hospital and Harvard Medical School. USA.
REVIEW RETURNED	28-Jan-2020

GENERAL COMMENTS	Thank you for the opportunity to re-review this manuscript. The manuscript is clearer and the abstract now helpfully explains the study. The introduction is clearer. The results have been revised as suggested. However, I do still suggest that the authors control for differences at the site . This is an exploratory study and can be included as a sub-analysis. If not, the authors need a section in the limitations to explain these site differences, e.g. as the authors noted in the response "other differences could be the staff at sites, data recording practices. The original aim was to have four sites to allow for more robust findings however as noted two sites withdrew from the study due to inability to provide data. Again this is not uncommon in NHS organisations and a real world trial like this highlights the real problems." This is a valuable response and explains the real-world differences compared to controlled research studies.

VERSION 2 – AUTHOR RESPONSE

The bullet points are numerous and rather wordy; please consider refining to better highlight the main issues.
Strengths and Limitations bullet points have been amended

1. Time between EIP referral and assessment is stressed as an important one. Looking at it from a non-UK perspective, I was positively surprised at the very short duration between EIP referral to EIP assessment, already at baseline! Would it be expected that a change from a median of 11 to 7 days actually have any clinically meaningful impact on DUP?
This has been further explained in the discussion section: From a patient and carer perspective a reduction in waiting times and DUP even of a few days, especially when acutely unwell, could be meaningful for example the potential impact being unwell could cause on relationships and employment.

2. Another outcome stressed in the study objectives is “clinical outcomes based on HoNOS scores”. If I understand correctly, there were 168 patients in the year 1 cohort at the comparator site, and HoNoS data were available for only 16 of these. Corresponding figures for the year 2 cohort data were 11 out of 102. Rather than retaining results of statistical analyses based on these extremely small subgroups, it would be more up front to simply state that it was not meaningful to test for changes among cohorts at the comparator sites due to the fact that 90% had missing data.
Analysis of the comparator site has been removed from table 5 and the results section and we have stated this in analysis section.

3. Little attention is placed on the qualitative component of the study. The lack of qualitative data for the comparator site is of course problematic. The authors might consider strengthening the scientific contribution of this study component or, alternatively, consider saving the qualitative

component for a separate study where results could be presented and discussed in more detail.
There is an additional full report of the qualitative analysis planned as there is limited space within this article. Note regarding the lack of qualitative data at the comparator site has been added to the methods section: Satisfaction and acceptability were assessed using questionnaires, interviews and focus groups. The later two were only conducted at the intervention site to enable a process evaluation of the implementation of the pathway at this site.

4. The issue of site differences at baseline is mentioned in the discussion (and, importantly, as a bullet point), but there is no discussion on how this might have impacted differentially on outcomes at the two sites. Just as an example: Considering the severe nature of the illness, I found it odd that only 56% of the baseline cohort at the comparator site received any of the listed interventions. The proportion at the implementation site was more in line with what could be expected. Does this tell us something about baseline differences in services at the two sites? And how might this effect outcomes?
It is noted that this is an issue that has come across from reviewers. However this was a factor of the pragmatic nature of this study. Further explanation of this has been added to the discussion: This reflects the variations in service provision landscape in the UK which can be geographically determined and can potentially impact on outcomes. There are additional factors like data recording at
sites, staffing etc. This study was not designed to explore these differences and their potential impact due to the pragmatic nature of the study meaning that some of the differences seen may have been due to other factors.

5. Not all outcomes were positive for the intervention site. The significant decrease in the proportion with a crisis plan (paralleled by a significant improvement in the comparator site) came as a surprise, and could be discussed.
This has been further highlighted in the discussion: Not all outcomes for the intervention site were positive for example the decrease in the recording of crisis plans, paralleled by the significant increase
in the comparator site.

6. The issue of multiple testing remains and could be specifically pointed out as a limitation.
This has been further highlighted in the study limitations section.

7. Abstract, first sentence in Results: This is not a result, but rather a description of a statistical method, which would be better placed in the methods section.
This has been moved

8. Page 8, Setting, line 5: "who were working to principles..." I think perhaps they mean "who were working according to principles..."
Changed accordingly

9. line 1 page 10 "limited to that routine collected" should probably read "routinely collected".
Changed accordingly

10. page 11, line 3. Please show full name of scale the first time HoNoS is used in the manuscript text.
Changed accordingly

11. Please check p value on page 20, third line: p=037 may be a typo?
Changed accordingly

12. Page 24, last sentence: Please check wording "..reduction in patients admitted to inpatients and the subsequent reduction of relapse". Might they mean "admitted to inpatient wards"?
Changed accordingly

13. Finally, (sorry, I missed this last time) it is stated in the text that there was no significant change in staff experiences across time points or between the sites. I could not find results of significance testing for these analyses.
Wording changed to no notable differences

14. I do still suggest that the authors control for differences at the site . This is an exploratory study and can be included as a sub-analysis. If not, the authors need a section in the limitations to explain these site differences, e.g. as the authors noted in the response "other differences could be the staff at sites, data recording practices. The original aim was to have four sites to allow for more robust findings however as noted two sites withdrew from the study due to inability to provide data. Again this is not uncommon in NHS organisations and a real world trial like this highlights the real problems." This is a valuable response and explains the real-world differences compared to controlled research studies.
As above, it is noted that this is an issue that has come across from reviewers. However this was a factor of the pragmatic nature of this study. Further explanation of this has been added to the discussion: This reflects the variations in service provision landscape in the UK which can be geographically determined and can potentially impact on outcomes. This study was not designed to explore these differences and their potential impact due to the pragmatic nature of the study meaning that some of the differences seen may have been due to other factors such as data recording at sites.

Please let us know if any further clarity is needed. We look forward to hearing from you.