

## 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

<b>TITLE (PROVISIONAL)</b>	Can process mining automatically describe care pathways of patients with long-term conditions in UK primary care? A study protocol
<b>AUTHORS</b>	Litchfield, Ian; Hoye, Ciaron; Shukla, David; Backman, Ruth; Turner, Alice; Lee, Mark; Weber, Phil

### VERSION 1 – REVIEW

<b>REVIEWER</b>	Carlos Fernandez Llatas Universitat Politecnica de Valencia
<b>REVIEW RETURNED</b>	10-Nov-2017

<b>GENERAL COMMENTS</b>	<p>The authors presents a protocol analyze the events in primary care for the discovery of clinical pathways using process Mining techniques over T2D patients.</p> <p>The application of Process Mining techniques over real health environments is a so demanding problem and I consider the study presented very interesting and exciting. However, I see some points that, in my humble opinion, can be improved.</p> <p>There are few studies about Process Mining in Healthcare, but I miss some works, also in diabetes, that you can find available in literature. In [1] I suggest a very recent literature review that the authors can use to improve the related work.</p> <p>There are not too much information about the methodology used, but It seems that you are applying a Interactive model methodology. If this is the case, I think the paper will improve in scientiphic sound if the authors try to formalice the metodology followed</p> <p>However, in my opinion, the main lack of the paper is that the authors does not explain the limitatons of the study due to the Spaghetti Effect and how this problem will be addressed. The Spaghetti Effect is the most common problem in the application of Process Mining in high variability environments like Healthcare. In [2] I suggest an article that is collected in the review [1] that cover some points in interactive process mining methodology for reducing Spaghetti effect in diabetes problems.</p> <p>Also other limitations should be treated in discussion part i.e: How to deal with paralelism? How about Data quality?...</p>
-------------------------	---

	<p>[1] E Rojas, J Munoz-Gama, M Sepúlveda, D Capurro Process mining in healthcare: A literature review Journal of biomedical informatics 61, 224-236 (<a href="https://www.ncbi.nlm.nih.gov/pubmed/27109932">https://www.ncbi.nlm.nih.gov/pubmed/27109932</a>)</p> <p>[2] C. Fernandez-Llatas, A. Martinez-Millana, A. Martinez-Romero, J. M. Benedi, and V. Traver, Diabetes care related process modelling using Process Mining techniques. Lessons learned in the application of Interactive Pattern Recognition: coping with the Spaghetti Effect, in 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2015, pp. 2127–2130. (<a href="https://www.ncbi.nlm.nih.gov/pubmed/26736709">https://www.ncbi.nlm.nih.gov/pubmed/26736709</a>)</p>
--	--

<b>REVIEWER</b>	Jan Vanthienen KU Leuven, Faculty of Economics and Business, Leuven Institute for Research on Information Systems
<b>REVIEW RETURNED</b>	20-Dec-2017

<b>GENERAL COMMENTS</b>	<p>Comments to authors: The protocol describes a methodology for applying process analytics in the healthcare domain and specifically in the TIIDM domain. I believe process mining and analysis can indeed contribute to healthcare. As such, this study could be an interesting contribution to the domain. Moreover, the motivation is well discussed. However, the authors claim to be the first, while I believe process mining has already been applied multiple times for healthcare purposes (in primary care) before. Furthermore, no clear methodology is presented, the concepts are not always clearly described and it is unclear how the authors will exactly contribute to the domain besides providing an additional case study. In order for the paper to be acceptable, I recommend to address the following issues:</p> <p>*With regards to the methodology, I recommend to base yourself on an existing framework. You could focus on design science research [4], process mining frameworks [3] or BPM frameworks [5] for example.</p> <p>*Abstract, Introduction: we already performed process mining for Type II Diabetes Mellitus patients using government sickness fund data (also including primary care), see [1]. Perhaps you can also take a look at [2]. Moreover, you already mention a couple of process mining papers in healthcare yourself.</p> <p>*Abstract, methods: which maps? Please be more specific and provide an explanation. The difference between traditional process maps based on qualitative data and process maps as defined in process mining literature is unclear.</p> <p>*What are possible limitations or challenges of applying process mining for this purpose? These are not adequately addressed.</p> <p>*p. 6: You mention a couple of examples of how process mining can contribute to organizational aspects. I like this, but the examples are not clear enough in their current version.</p> <p>*How would you position your techniques in the general process mining framework, see also [3]?</p> <p>*You mention that you want to develop your own methods and algorithms. How do they compare to existing methods and frameworks in process mining, and specifically to those in process mining for healthcare? E.g., in what sense does your data deviate from the data used in previous work? See also our references in [1].</p> <p>*p. 7: which two techniques are you planning to apply?</p>
-------------------------	--

	<p>*Research design: you mention that you will work with data from CMS. How does it relate to EHR data? And to the event logs, e.g. how granular are the timestamps? Are you able to follow all activities of a specific patient?</p> <p>*Which technique did you use for Figure 1? This process is very structured already, which is unusual for process mining. Is this a realistic example? If so, did you preprocess the data?</p> <p>*Phase 3: which techniques will you use to compare the traditional method with yours? You mention focus groups later on, but still the methods are not discussed extensively.</p> <p>*p. 10, line 8: this phrase is unclear. Moreover, what do you mean with “with whom to build a process model”? I also find the explanation of event logs unclear, perhaps an example can help. E.g. you don’t mention the concept of a trace and resource, what is an activity exactly, etc.</p> <p>*The overall structure of the document could be improved. E.g. several subheadings are repeated, and information related to one topic is scattered across the paper.</p> <p>*Spelling, lay-out and grammar: punctuation and formatting is not always correct which affects comprehensibility. Furthermore, the phrases themselves are also not always correct and some words contain typos.</p> <p>*Note that the dates of the study are not included, which is necessary for a protocol.</p> <p>[1] Jasmien Lismont, Anne-Sophie Janssens, Irina Odnoletkova, Seppe vanden Broucke, Filip Caron, Jan Vanthienen, A guide for the application of analytics on healthcare processes: A dynamic view on patient pathways, In Computers in Biology and Medicine, Volume 77, 2016, Pages 125-134, ISSN 0010-4825, <a href="https://doi.org/10.1016/j.combiomed.2016.08.007">https://doi.org/10.1016/j.combiomed.2016.08.007</a>.</p> <p>[2] Denis Klimov, Alexander Shknevsky, Yuval Shahar, Exploration of patterns predicting renal damage in patients with diabetes type II using a visual temporal analysis laboratory, In Journal of the American Medical Informatics Association, Volume 22, Issue 2, 2015, Pages 275-289, <a href="https://doi.org/10.1136/amiajnl-2014-002927">https://doi.org/10.1136/amiajnl-2014-002927</a>.</p> <p>[3] W. van der Aalst, Process Mining: Discovery, Conformance and Enhancement of Business Processes, Springer-Verlag Berlin Heidelberg, Schleiden, Germany (2011), <a href="https://doi.org/10.1007/978-3-642-19345-3">https://doi.org/10.1007/978-3-642-19345-3</a>.</p> <p>[4] Alan R. Hevner, Salvatore T. March, Jinsoo Park, Sudha Ram, Design science in information systems research, In MISQ Quarterly, Volume 28, Issue 1, Pages 75-105, 2004.</p> <p>[5] Marlon Dumas, Marcello La Rosa, Jan Mendling, Hajo A. Reijers, Fundamentals of Business Process Management, Springer, 2013.</p> <p>[6] Filip Caron, Jan Vanthienen, Kris Vanhaecht, Erik van Limbergen, Jochen De Weerd, Bart Baesens, Monitoring care processes in the gynecologic oncology department. Comp. in Bio. and Med. 44: 88-96, 2014</p> <p>[7] Filip Caron, Jan Vanthienen, Bart Baesens, Clinical Pathway Analytics. JITR 7(1): 12-26, 2014.</p>
--	--

<b>REVIEWER</b>	Arianna Dagliati University of Manchester, UK
<b>REVIEW RETURNED</b>	21-Dec-2017

<b>GENERAL COMMENTS</b>	Overall, the presented study protocol is clear and the methodology appropriate. However, there are several points that should be
-------------------------	--

	<p>better explained. A more careful revision of the literature and insightful considerations of possible drawbacks in the application of process mining methods to T2D care flows are also needed. The date of the study is missing.</p> <p>Regarding the literature review, authors should apply a broader approach, which encompasses other fields very much correlated (and sometimes used as synonyms), to process mining: workflow and careflow mining. In this regard, authors might consider the Rojas et al. review about process mining in health care (<a href="https://www.ncbi.nlm.nih.gov/pubmed/27109932">https://www.ncbi.nlm.nih.gov/pubmed/27109932</a>), in order to better assess the applied mining methods. Upon this revision, authors should also carefully reconsider their statement about this being the first time process mining is applied to primary care. The employed terminology has to be consistent across process maps and process mining, the results of process mining discovery algorithms are usually referred to as “process models”.</p> <p>The research question is well posed, but it misses two important steps: the first one related to a precise data model definition, and the last one related to the effects of different processes on patients’ outcome, which should entail the ultimate validation of the protocol. Data modelling and pre-processing in the healthcare context poses several challenges, which authors should consider. For example, depending on the level of chosen description, results might consistently vary. This might be particularly crucial when information about medications or Read coded diagnosis are included in the process. If data are represented through detailed concepts (e.g. specific diagnosis or drug names) its intrinsic high variability can produce so-called “spaghetti-like” models; otherwise, if data representation is too general (e.g. classes of disease or active principles) important information might be lost.</p> <p>Another issue regarding the representation, and handling within the discovery phase, of event type and measure type. Authors should better explain how they are planning to manage heterogeneous data and might check this two-step approach proposed by Dagliati et al. (<a href="https://www.ncbi.nlm.nih.gov/pubmed/28057564">https://www.ncbi.nlm.nih.gov/pubmed/28057564</a>) based on process and clinical data.</p> <p>The same considerations apply to associated data, in particular locations and clinicians, which illustrate resource consumption. Their integration within the framework might be crucial, and authors should clearly explain how they are planning to integrate this information into the framework.</p> <p>Considering the frequency of variant iterations as a part of the process is a very good point, especially for T2D populations, where long chronic disease histories entail recursive paths. However, these long and complex processes often include acute events and complications that produce process deviations. These deviations have to be included in the model, as they reflect the ultimate outcomes of the process and explain different variants (e.g. changes in treatments, test prescriptions). Authors should explain if and how the proposed approach will be able to take into account these events.</p>
--	--

**VERSION 1 – AUTHOR RESPONSE**

Reviewer(s)' Comments to Author:

Reviewer: 1

Reviewer Name: Carlos Fernandez Llatas  
Institution and Country: Universitat Politecnica de Valencia  
Please state any competing interests or state 'None declared': None Declared.

Please leave your comments for the authors below

The authors presents a protocol analyze the events in primary care for the discovery of clinical pathways using process Mining techniques over T2D patients.

The application of Process Mining techniques over real health environments is a so demanding problem and I consider the study presented very interesting and exciting. However, I see some points that, in my humble opinion, can be improved.

Thank you.

There are few studies about Process Mining in Healthcare, but I miss some works, also in diabetes, that you can find available in literature. In [1] I suggest a very recent literature review that the authors can use to improve the related work.

Many thanks for this we have now produced a section entitled "Knowledge Review" which follows the Introduction. It consists of two parts. In the first we discuss the existing literature describing the use of process mining in healthcare, the second describes the issues around mining complex and entangled processes. This begins on page 6.

There are not too much information about the methodology used, but It seems that you are applying a Interactive model methodology. If this is the case, I think the paper will improve in scientific sound if the authors try to formalize the methodology followed

We agree and thank you for this useful advice. We will follow what seems to be the emerging "standard" process mining project methodology in healthcare as proposed by Bozkaya [2009] while particularly emphasising iterative, interactive exploration of the data and the modelling results with clinical experts. We summarise our approach in the "Research Design" section and provide more theoretical detail in the "Research Methodologies" section.

REF:

Process diagnostics: a method based on process mining. M. Bozkaya, J. Gabriels, J. Werf, , in: International Conference on Information, Process, and Knowledge Management, 2009. eKNOW'09, IEEE, 2009, pp. 22–27

However, in my opinion, the main lack of the paper is that the authors does not explain the limitations of the study due to the Spaghetti Effect and how this problem will be addressed. The Spaghetti Effect is the most common problem in the application of Process Mining in high variability environments like Healthcare. In [2] I suggest an article that is collected in the review [1] that cover some points in interactive process mining methodology for reducing Spaghetti effect in diabetes problems.

It is indeed true that the so-called "spaghetti" processes arising from complex and heterogeneous data are a particular problem when using process mining in healthcare settings. Thank you for the suggested article. We have now included an extensive discussion of the literature in the "Knowledge Review" section which includes a discussion on the options available to mitigate the problem.

Also other limitations should be treated in discussion part i.e: How to deal with parallelism?

We have added to the “Strengths and Limitations” a summary point on limitations of process mining in healthcare. The “Knowledge Review” section includes a more extensive discussion of the potential complications associated with process mining including the complexity of the data and underlying processes. Potential limitations with respect to primary care (additional variety in data, events, and disease types) and long-running chronic diseases are also referred to.

Dealing with parallelism is standard in process mining, particularly in unstructured and flexible environments like healthcare and we do not feel that a detailed discussion fits well in a protocol paper. We would however expect such a discussion to be an important part of describing any novel algorithms which we produce in the future.

How about Data quality?...

We will be using the data held on the clinical management system of each of the study practices. Our data will be drawn from either EMIS Web, SystemOne or InVision I management systems. Incidentally all provide consensual access to data held on a sample of practices using SystemOne e.g THIN or Research One.

We have now included a line in the first paragraph page 16 which reads.

“We have some experience of using these datasets and the data they contain forms the basis of pseudonymised datasets used in other examples of primary care research [THIN 2017]. We have also successfully used the data held on the CMS in exploring prescribing behaviours in multimorbid patients in primary care [Backman et al 2017]”

REFS:

<https://www.birmingham.ac.uk/research/activity/mds/projects/HaPS/PCCS/THIN/index.aspx>.

Backman R, Weber P, Turner AM, et al Assessing the extent of drug interactions among patients with multimorbidity in primary and secondary care in the West Midlands (UK): a study protocol for the Mixed Methods Multimorbidity Study (MiMMS) BMJ Open 2017;7:e016713. doi: 10.1136/bmjopen-2017-016713

Reviewer: 2

Reviewer Name: Jan Vanthienen

Institution and Country: KU Leuven, Faculty of Economics and Business, Leuven Institute for Research on Information Systems

Please state any competing interests or state ‘None declared’: None declared

Please leave your comments for the authors below Comments to authors:

The protocol describes a methodology for applying process analytics in the healthcare domain and specifically in the TIIDM domain. I believe process mining and analysis can indeed contribute to healthcare. As such, this study could be an interesting contribution to the domain. Moreover, the motivation is well discussed.

Thank you

However, the authors claim to be the first, while I believe process mining has already been applied multiple times for healthcare purposes (in primary care) before.

We have now produced a Knowledge review that more thoroughly describes previous research into process mining in healthcare environments. We do acknowledge that process mining has already been applied many times to healthcare, yet in primary care, and especially in the UK, its application is much more limited.

We have also clarified that this is the first time that to the best of our knowledge process mining has been used in primary care in the UK in the second paragraph of page 7.

“Our study is the first time that process mining has been used in primary care settings in the UK to describe the care processes used by individual practices.”

Furthermore, no clear methodology is presented,

Thank you for this advice. We have clarified our methodology in the “Research Methodology” section and summarised it where appropriate (Introduction and “Research Design”). In summary we base our approach on the emerging “standard” methodology, while further emphasising iterative and interactive development with healthcare experts during this project. We aim to produce a process mining approach which is as automated as possible, while retaining flexibility for users and general applicability across heterogeneous sites.

the concepts are not always clearly described

We now begin the “Research Methodology” section by explaining the core concepts of process mining and how they apply to our study, summarised in Table 1.

and it is unclear how the authors will exactly contribute to the domain besides providing an additional case study.

This proof of concept study will demonstrate that automated discovery of care processes in UK primary care is viable using the data held on the clinical management systems of individual practices. In particular the study will build on existing developments which deal with heterogeneous data, complex processes, etc. to apply process mining in the complex and flexible primary care environment in the UK.

The algorithms we will develop will enable the comparison of processes across practices and relative to existing clinical guidance. Our partners BXCCG are committed to systematically understanding more about the care processes employed across their Group in particular for Hypertension and TIIDM and how they might be optimised to improve the quality and efficiency of care. In doing so we will apply and develop principles proposed in our previous work [Weber et al 2013] to bring rigour and justifiability to our methods for mining, modelling, simplifying, generalising and visualising primary care processes. We summarise this in a paragraph in the Introduction.

Weber P, Bordbar B, Tiño P. A Framework for the Analysis of Process Mining Algorithms. IEEE T. Systems, Man, and Cybernetics: Systems 2013;43:303-17.

In order for the paper to be acceptable, I recommend to address the following issues:

\*With regards to the methodology, I recommend to base yourself on an existing framework. You could focus on design science research [4], process mining frameworks [3] or BPM frameworks [5] for example.

Thank you for these recommendations. We take a process mining methodology approach, based on the emerging “standard” approach (see also answers above) but focussing on interaction and iterative development. We explain this under “Research Methodologies” Section beginning on Page 12.

\*Abstract, Introduction: we already performed process mining for Type II Diabetes Mellitus patients using government sickness fund data (also including primary care), see [1]. Perhaps you can also take a look at [2]. Moreover, you already mention a couple of process mining papers in healthcare yourself.

Thank you this is now clarified and the amended text reads

“Here and for the first time in the UK we will apply them to primary care to gain a greater understanding of how patients with Type II Diabetes Mellitus (TIIDM) and Hypertension are managed.”

As mentioned above we have also extended the literature review to incorporate more detail on the previous use of process mining in healthcare.

\*Abstract, methods: which maps? Please be more specific and provide an explanation. The difference between traditional process maps based on qualitative data and process maps as defined in process mining literature is unclear.

We have now clarified in the text which reads

“Third, with the aid of staff and patients at each practice we will compare and contrast the process models produced by process mining with the process maps produced via manual techniques, review differences and similarities between them and the relative importance of each.”

We have also clarified the terminology, using “process models” for the results of process mining, and “process maps” for the outputs of manual process mapping exercises. See also response to Reviewer 3.

\*What are possible limitations or challenges of applying process mining for this purpose? These are not adequately addressed.

We have added a bullet point in the strengths and limitations section and under “Knowledge Review” added a detailed review of issues, relating to the challenges of dealing with complexity of healthcare data and processes. Specific challenges relating to primary care in the UK, motivating the need for this study, include a) heterogeneous data sources, b) flexibility in applying clinical guidelines to the local context, and c) diversity in site-specific size and demographics, and also d) long-running processes for patients with chronic diseases.

\*p. 6: You mention a couple of examples of how process mining can contribute to organizational aspects. I like this, but the examples are not clear enough in their current version.

I'm sorry I'm not entirely clear which sentences you are referring to. We will happily clarify following your further advice. Meanwhile I have added text with regards the challenges of applying process mining in the primary care environment (as opposed to secondary care). The text now reads in the section headed "Research Question"

"...the overarching aim of our study is to determine whether process mining techniques can be applied equally effectively to primary care with its challenges of scale and diversity"

\*How would you position your techniques in the general process mining framework, see also [3]?

Considering the three broad areas of process mining (process discovery, conformance, extension) we are working within process discovery. However, we anticipate identifying needs in the conformance area (e.g. to formally compare processes from different general practices and with national guidance), and extension (connecting processes implemented with patient outcomes, making predictions, etc.) which would be the subject of future research. Methodologically, we expect to work across the full scope of obtaining, cleaning, aggregating, clustering data, mining, validating – iteratively – as described in responses to previous comments.

\*You mention that you want to develop your own methods and algorithms. How do they compare to existing methods and frameworks in process mining,

This is a proof of concept study exploring the effectiveness of applying process mining in primary care in the UK. We state that in Phase One we will begin by applying existing algorithms and only extending them or developing our own algorithms as necessary. We believe this is likely, in which case our algorithms will be underpinned by a machine learning theoretic mindset, i.e. closely related to machine learning concepts of learning distributions over traces of activities, and metrics to measure / control the learning and generalisation. The existing literature (see Knowledge Review section) shows that most case studies either apply existing algorithms in a limited context, reporting some initial successes and future challenges, or develop new algorithms for a specific context, which are yet to be widely accepted.

and specifically to those in process mining for healthcare? E.g., in what sense does your data deviate from the data used in previous work? See also our references in [1].

We do not assume that our data deviates enormously from other routinely collected healthcare data held in other countries. However it is reasonable to assume that differences in the organisational cultures of different healthcare systems might mean that the equivalent data is not recorded in an identical fashion. The data held on the clinical management systems in the UK might present challenges and opportunities that similar datasets collated elsewhere do not – for example in the future we may be able to use natural language processing to interpret the data entered freehand by GPs to add to the richness of the dataset.

\*p. 7: which two techniques are you planning to apply?

We have clarified the text to better convey our intention to compare 1) process models resulting from automatic process mining, with 2) process maps arising from manual process mapping exercises.

\*Research design: you mention that you will work with data from CMS. How does it relate to EHR data? And to the event logs, e.g. how granular are the timestamps? Are you able to follow all activities of a specific patient?

The key data being collected is described in Table 2. Specifically with regards to timestamps yes we should be able to follow all of the activities of a specific patient. We cannot however make broad assumptions as yet on how complete the data is and there may be differences between practices or the way in which between physicians at the same practice records patient- related information.

Prior work in the process mining literature would lead us to expect to encounter data issues in an exploratory study of this nature, particularly involving data from diverse sources. These cannot be precisely identified in advance but we are aware of (and now review in the “Knowledge Review” section) previous approaches to mitigate the effects of data missing or logged out of sequence, inconsistent granularity, etc

\*Which technique did you use for Figure 1? This process is very structured already, which is unusual for process mining. Is this a realistic example? If so, did you pre-process the data?

This was a consciously stylised version of a process; aware of the readership of the BMJ Open we wanted to provide a straightforward example of what the resulting process model might look like while also highlighting some of the problems with regard to noise, and complex or entangled processes.

\*Phase 3: which techniques will you use to compare the traditional method with yours? You mention focus groups later on, but still the methods are not discussed extensively.

As stated in the article we will explore the differences between the maps qualitatively. We are interested in how closely they resemble the experiences of staff and patients and also the relative importance of provider and patient perspectives on the differences and similarities in the resultant maps.

We have now included further text

“Focus groups were chosen as the primary method of data collection as the interaction between participants can serve to challenge any over idealised statements and produce realistic accounts of what people actually do.[ Morgan DL. 1993]; They also offer an opportunity for participants to reflect and test ideas rather than formulate ideas on the spot and the uninhibited discussion can remind participants and generate new thoughts [Gill P, Stewart K, et al 2008].”

REFS:

Morgan DL. Future directions in focus group research. Successful Focus Groups. London: Sage; 1993

Gill P, Stewart K, Treasure E, Chadwick B. Methods of data collection in qualitative research. British Dental Journal 2008; 204:291-295.

\*p. 10, line 8: this phrase is unclear. Moreover, what do you mean with “with whom to build a process model”?

We apologise that this section was grammatically unclear; however we have now expanded and clarified this passage in the text on page 10.

I also find the explanation of event logs unclear, perhaps an example can help. E.g. you don't mention the concept of a trace and resource, what is an activity exactly, etc.

We now begin our “Research Methodologies – Process Mining” section with a brief description of these concepts, summarised in Table 1.

\*The overall structure of the document could be improved. E.g. several subheadings are repeated, and information related to one topic is scattered across the paper.

Thank you, we have now re-structured the document. Adding a Knowledge Review and reducing repetition by collating the technical detail in the Research Methodologies section.

\*Spelling, lay-out and grammar: punctuation and formatting is not always correct which affects comprehensibility. Furthermore, the phrases themselves are also not always correct and some words contain typos.

We have again proof read the manuscript and believe we have eliminated these errors.

\*Note that the dates of the study are not included, which is necessary for a protocol.

These have now been added to the text.

“The study will begin in June 2018 and last for 12 months.”

Note to the editor: Note that I don't have experience with reviewing study protocols. However, I have focused on the methodology as suggested in the note from the editors.

Reviewer: 3

Reviewer Name: Arianna Dagliati

Institution and Country: University of Manchester, UK

Please state any competing interests or state 'None declared': None

Please leave your comments for the authors below

Overall, the presented study protocol is clear and the methodology appropriate.

Thank you.

However, there are several points that should be better explained. A more careful revision of the literature and insightful considerations of possible drawbacks in the application of process mining methods to T2D care flows are also needed.

We have added a bullet point to the Strength and Limitations section and present an extended literature review in the "Knowledge Review" section that focuses on process mining projects in healthcare, and the strategies available for mitigating the so-called "spaghetti problem" arising from complex data and processes. We have also added a paragraph focussing on potential limitations and difficulties with respect to long-running processes for chronic diseases, and the UK primary care landscape.

The date of the study is missing.

Now included in the text.

Regarding the literature review, authors should apply a broader approach, which encompasses other fields very much correlated (and sometimes used as synonyms), to process mining: workflow and careflow mining. In this regard, authors might consider the Rojas et al. review about process mining in health care (<https://www.ncbi.nlm.nih.gov/pubmed/27109932>), in order to better assess the applied mining methods.

We have extended the literature review in particular addressing the questions of methodology and dealing with complex and heterogeneous data and underlying processes (leading to the problem of so-called "spaghetti" models). The Rojas review was very useful and our search uncovered relevant articles using the terms "workflow mining" and "careflow mining".

Upon this revision, authors should also carefully reconsider their statement about this being the first time process mining is applied to primary care.

Now clarified in the text that this is the first time this has been used in primary care in the UK. Please also see the additional material included in the review.

The employed terminology has to be consistent across process maps and process mining, the results of process mining discovery algorithms are usually referred to as "process models".

Our intention was to standardise the terminology to refer to the outputs of both process mining and (manual) process mapping, as "maps", but on reflection this is confusing. As suggested, we now refer

to the results of process mining, as process models, and manual mapping, as process maps, to be consistent with the relevant literature.

The research question is well posed, but it misses two important steps: the first one related to a precise data model definition, and the last one related to the effects of different processes on patients' outcome, which should entail the ultimate validation of the protocol.

We have extended the data overview in Table 2 to highlight the events recorded that will be the basis of our process mining activities. At this stage we are not in a position to report a precise data model definition and how it relates to the algorithms developed. The study is predicated on an iterative exploration of the data with experts to produce the optimal combination of data, algorithms and visualisation for our primary care landscape.

At this point we are not looking at patient outcome. Our work is a proof of concept that process mining can be a valid tool in understanding (and modelling) the treatment processes for patients with complex chronic conditions. We can confirm however that during the next phase we intend to explore the outputs of these various processes and how they might impact on service utilisation and patient outcome and how they might be amended to improve both.

Data modelling and pre-processing in the healthcare context poses several challenges, which authors should consider. For example, depending on the level of chosen description, results might consistently vary. This might be particularly crucial when information about medications or Read coded diagnosis are included in the process. If data are represented through detailed concepts (e.g. specific diagnosis or drug names) its intrinsic high variability can produce so-called "spaghetti-like" models; otherwise, if data representation is too general (e.g. classes of disease or active principles) important information might be lost.

Thank you for this advice. We have we have reviewed the approaches for mitigating the spaghetti effect in the "Knowledge Review" section and have summarised in the "Research Methodologies" section our initial ideas for dealing with them, particularly with regard to UK primary care data. We also note that since this is a protocol paper, we expect that the final solutions for mitigating these challenges, and even the challenges themselves, may differ significantly from those anticipated at this stage.

Another issue regarding the representation, and handling within the discovery phase, of event type and measure type. Authors should better explain how they are planning to manage heterogeneous data and might check this two-step approach proposed by Dagliati et al. (<https://www.ncbi.nlm.nih.gov/pubmed/28057564>) based on process and clinical data.

Thank you for this advice and for bringing this paper to our attention. We review some of the existing approaches to try to address this problem, including aggregating events into "event types", trace clustering, semantic notions, etc.

The same considerations apply to associated data, in particular locations and clinicians, which illustrate resource consumption. Their integration within the framework might be crucial, and authors should clearly explain how they are planning to integrate this information into the framework.

Again thank you for these comments, and we agree that it may be important to incorporate such data into our framework. However this being a protocol it is not possible at this stage to describe in detail all the data which we may need to work with and exactly how it will be integrated. We refer to our previous work with data-enriched process models [Weber et al 2013] and related literature covering different approaches taken in the past.

REF:

Weber P, Bordbar B, Tiño P. A Principled Approach to Mining From Noisy Logs Using Heuristics Miner. In Proc. IEEE Symposium on Computational Intelligence and Data Mining (CIDM). 2013. p. 119-26.

Considering the frequency of variant iterations as a part of the process is a very good point, especially for T2D populations, where long chronic disease histories entail recursive paths. However, these long and complex processes often include acute events and complications that produce process deviations. These deviations have to be included in the model, as they reflect the ultimate outcomes of the process and explain different variants (e.g. changes in treatments, test prescriptions). Authors should explain if and how the proposed approach will be able to take into account these events.

Thank you for highlighting these issues. Our approach is for an iterative development cycle with process/data/clinical experts. In this way the capabilities of the process mining algorithms, the support from the data, and the nuances of the clinical processes that emerge, will become apparent. Healthcare experts will drive the requirements for what should and should not be visualised, and how this should be controlled, while the computational aspects will ensure validity of the results from a machine learning point of view.

As part of the final phase we will be seeking the feedback of providers and patients and this will further inform which aspects of the process are important or less important or that have otherwise been missed. We are hopeful that the events you describe will be accounted for by the process mining and some measure of their frequency obtained.

### VERSION 2 – REVIEW

<b>REVIEWER</b>	Carlos Fernandez Llatas Universitat Politecnica de Valencia, Spain
<b>REVIEW RETURNED</b>	17-Feb-2018
<b>GENERAL COMMENTS</b>	<p>The authors present a study about the application of Process Mining techniques over Diabetes primary care patients to know how these technologies can deal with discovery of processes</p> <p>In my opinion, in general the paper is well written and easy to read, the awareness of the literature is adequate, The Study process is correct</p> <p>As a minor issue, in my opinion the word "determine" in the title is too hard and the study probably is not enough to demonstrate that. In interactive models is the human expert who has the real potentiality to determine. I suggest changing the word "determine" for "support the determination" or something like that.</p>

<b>REVIEWER</b>	Jan Vanthienen KU Leuven, Belgium
<b>REVIEW RETURNED</b>	17-Feb-2018

<b>GENERAL COMMENTS</b>	<p>Comments:</p> <p>The proposal has improved significantly. The authors have added an extensive literature section, clarified their methods and structured their paper better. Nevertheless, I have some small final remarks.</p> <p>*The spelling and grammar have improved a lot, but still some grammatical errors occur, e.g. "data held on CMS" (abstract), "flow of material and resource" (p.5, line 24). There are also inconsistencies in the applied acronyms.</p> <p>*methods and analysis, abstract: do you apply process mining in phase 1? It's not entirely clear from this section in which phase you apply automated process mining.</p> <p>*Please correct "not hypertension, although such data" (p.8, line 5)</p> <p>*Research questions: can you, in the third objective, clarify the desired final outcome? Do you, as an end result, wish to present a tool or rather a framework?</p> <p>*Section methods and analysis: "We will then use the algorithms we discover as a basis for mining the processes of TIIDM." Please explain better what you mean here. What do you mean with the "algorithms we discover"?</p> <p>*Which technology/ies will you use for the process mining techniques?</p> <p>*Table 1: wouldn't a case be defined as a specific patient? Or a specific patient within a specific practice? As far as I know, a trace will also follow a specific case. Maybe also add 'resource'? Plus, this table definitions are not really consistent with what is said on p.18.</p> <p>*Table 2: how granular is the time data? Because this can cause problems with identification of sequence.</p> <p>*Table 2, Practice: what does 'wte' mean? And CCG?</p> <p>*How many patients do you aim to recruit for each case?</p>
-------------------------	---

<b>REVIEWER</b>	Arianna Dagliati University of Manchester, UK
<b>REVIEW RETURNED</b>	23-Feb-2018

<b>GENERAL COMMENTS</b>	<p>Authors addressed my previous comments and did a satisfactory job improving the manuscript, especially in the knowledge review and methods sections.</p> <p>Minor comment and typos to be corrected.</p> <p>Page 2, line 34: I would cite them here. Eventually specifying that the first "pilot" will regard hypertension and the secondo diabetes.</p> <p>Page8, line 6: check the sentence "not hypertension, although such data" (?)</p> <p>Page 9, line 22: remove the bracket</p> <p>Page 9, lines 32/33: at this point this has been claimed various time. Although this is a key message, I would suggest limiting repetition of the same concept too many times.</p> <p>Page 11, line 27: be more precise in defining "mixed" or use the term "focus group"</p> <p>PAGE 18, LINE 40: "patients with TYIIDM" check and be consistent with terminology through the <sup>11</sup><sub>SEP</sub> whole manuscript.</p>
-------------------------	--

## VERSION 2 – AUTHOR RESPONSE

Reviewer(s)' Comments to Author:

Reviewer: 1

Reviewer Name: Carlos Fernandez Llatas

Institution and Country: Universitat Politecnica de Valencia, Spain

Please state any competing interests or state 'None declared': None Declared

Please leave your comments for the authors below The authors present a study about the application of Process Mining techniques over Diabetes primary care patients to know how these technologies can deal with discovery of processes

In my opinion, in general the paper is well written and easy to read, the awareness of the literature is adequate, The Study process is correct

As a minor issue, in my opinion the word "determine" in the title is too hard and the study probably is not enough to demonstrate that. In interactive models is the human expert who has the real potentiality to determine. I suggest changing the word "determine" for "support the determination" or something like that.

Thank you for your suggestion. We feel "support the determination of " may be too long a phrase to include in the title and so have substituted determine with describe.

Reviewer: 2

Reviewer Name: Jan Vanthienen

Institution and Country: KU Leuven, Belgium

Please state any competing interests or state 'None declared': None declared

Please leave your comments for the authors below

Comments:

The proposal has improved significantly. The authors have added an extensive literature section, clarified their methods and structured their paper better.

Thank you.

Nevertheless, I have some small final remarks.

\*The spelling and grammar have improved a lot, but still some grammatical errors occur, e.g. "data held on CMS" (abstract), "flow of material and resource" (p.5, line 24). There are also inconsistencies in the applied acronyms.

Apologies for any confusion but the version submitted does not contain these phrases. Instead the abstract reads "data held on the clinical management system" and on page 5 it reads "flow of materials and resource".

\*methods and analysis, abstract: do you apply process mining in phase 1? It's not entirely clear from this section in which phase you apply automated process mining.

Thank you – now clarified in the text

\*Please correct “not hypertension, although such data” (p.8, line 5)

Thank you the phrase has been removed

\*Research questions: can you, in the third objective, clarify the desired final outcome? Do you, as an end result, wish to present a tool or rather a framework?

We have now included the text

“This will allow us to develop a framework to optimise the use of process mining to automatically describe complex care pathways in primary care.”

\*Section methods and analysis: “We will then use the algorithms we discover as a basis for mining the processes of TIIDM.” Please explain better what you mean here. What do you mean with the “algorithms we discover”?

Thank you. This phrase was included in error and has been removed.

\*Which technology/ies will you use for the process mining techniques?

Apologies but we are unsure as to what the reviewer is referring to. Technology only appears once in the text and this in the literature review section in reference to Associated Rule Mining.

To confirm we will develop our algorithms as plug-ins for the ProM framework for process mining in line with the current approach.

\*Table 1: wouldn't a case be defined as a specific patient? Or a specific patient within a specific practice? As far as I know, a trace will also follow a specific case. Maybe also add 'resource'?

Thank you. For this exploratory study we are considering a case as a specific patient's interaction with a practice. This could be all interactions with the practice but more realistically will focus on activities related to a particular process (sub-process), e.g. yearly medication review. However, we note that process mining is a broad field and allows for differing viewpoints to be taken on the data, e.g. to take the perspective of staff in response to specific questions.

We have amended and extended Table 1 to clarify as suggested and understand 'trace' to refer to the recorded events relating to a specific case, rather than the events themselves.

Plus, this table definitions are not really consistent with what is said on p.18.

Apologies for our confusion but we are not sure that we have understood this comment correctly. On p18 we are discussing the data contained in Table 2. The text relating to Table 1 can be found on pages 12-13 where it describes process mining more from the business perspective, summarising the relationship with healthcare, whereas the table gives the specifics. We have modified the text slightly to clarify the focus of a case.

\*Table 2: how granular is the time data? Because this can cause problems with identification of sequence.

We have a minimum of date though there may be further detail in terms of which clinic (such as morning or evening for example) contained in the data set. We have added some text to clarify this under "Data Management and Analysis" and also suggesting ways in which we may mitigate problems arising if on occasion the granularity is too low.

\*Table 2, Practice: what does 'wte' mean? And CCG?

Thank you. Wte means "whole time equivalent". This has now been included in the table

CCG means Clinical Commissioning Group (the clinically-led statutory NHS bodies responsible for the planning and commissioning of health care services for their local area.) This is included in the table in full and now in the body of the article.

\*How many patients do you aim to recruit for each case?

We have added a sentence "This will include data for an estimated 4000 patients."

Reviewer: 3

Reviewer Name: Arianna Dagliati

Institution and Country: University of Manchester, UK

Please state any competing interests or state 'None declared': None

Please leave your comments for the authors below Authors addressed my previous comments and did a satisfactory job improving the manuscript, especially in the knowledge review and methods sections.

Minor comment and typos to be corrected.

Page 2, line 34: I would cite them here. Eventually specifying that the first "pilot" will regard hypertension and the secondo diabetes.

Thank you, we have now included the text

"The first pilot study will be on Hypertension and the second for patients diagnosed with Type II Diabetes."

Page8, line 6: check the sentence "not hypertension, although such data" (?)

Now removed thank you.

Page 9, line 22: remove the bracket Page 9, lines 32/33: at this point this has been claimed various time. Although this is a key message, I would suggest limiting repetition of the same concept too many times. Page 11, line 27: be more precise in defining "mixed" or use the term "focus group"

Now amended thank you.

PAGE 18, LINE 40: "patients with TYIIDM" check and be consistent with terminology through the whole manuscript.

Thank you now corrected.

### VERSION 3 – REVIEW

<b>REVIEWER</b>	Jan Vanthienen KU Leuven, Belgium
<b>REVIEW RETURNED</b>	15-Apr-2018
<b>GENERAL COMMENTS</b>	All remarks from the previous review have been answered.