Author’s response to reviews

Title: Implementing monitoring triggers and matching of triggered and control sites in the TEMPER study: A description and evaluation of a triggered monitoring management system

Authors:
Carlos Diaz-Montana (c.diaz@ucl.ac.uk; cadiazm@yahoo.com)
William Cragg (w.cragg@ucl.ac.uk)
Rahela Choudhury (rahela.choudhury@ucl.ac.uk)
Nicola Joffe (n.joffe@ucl.ac.uk)
Matthew Sydes (m.sydes@ucl.ac.uk)
Sally Stenning (sally.stenning@ucl.ac.uk)

Version: 1 Date: 15 Feb 2019

Author’s response to reviews:

Response to reviewers: Implementing monitoring triggers and matching of triggered and control sites in the TEMPER study: A description and evaluation of a triggered monitoring management system.

We thank the editor and both reviewers for the detailed comments and suggestions on the manuscript, which have improved its content by adding more clarity, making it more accessible for a wider audience. Please see below a point-by-point response to each comment. We hope you find our responses satisfactory.

All line numbers referred in the author’s responses correspond to the clean version of the manuscript.

REVIEWERS’ COMMENTS & AUTHORS’ RESPONSES
Reviewer #1:

The manuscript supports the published TEMPER study results in Stenning et al (2018); specifically, on the logic decisions behind the matching of perceived higher-risk to lower-risk sites for matched pair monitoring visits in TEMPER. I was impressed by many features of the bespoke risk-based monitoring MS and its flexibility, e.g.: the variation in the way triggers could be set and fired; cumulative triggers that fire if a condition is met more than once; allowing thresholds to change; the addition of manual triggers; the integration of the report generation with the regular meeting to select sites for monitoring visits.

The method for selection of matched sites was very interesting. The authors chose also to match similar sites using time since opening and number of participants. Logarithmic plotting to reduce the importance of difference in time since opening when sites had been open longer was appropriate. As the matching algorithm is a critical part of the site pairing, and therefore the findings from TEMPER, I have commented on it in detail.

It should be acknowledged where data is reported here that is published in the 2018 TEMPER results (authors stated this in their covering letter):

The matching algorithm is in Stenning et al, 2018 as an Appendix

Some of Table 2.

Response: Text added at the end of Background section with a reference to Stenning et al (2018): “For ease of reference, this paper contains some details that were previously reported in [10] including the description of the matching algorithm and part of Table 2.”

Essential revisions and points for consideration:

1. Line 25: please state how many sites participated in the three trials and what proportion of the sites were visited.
Response: Text added to line 26: “with 156 UK sites in total”, and lines 225-226: “From 156 UK sites participating in at least one of the three trials, 67 different sites (43%) were visited at least once during the course of TEMPER as triggered or untriggered sites”

2. Lines 28-30: for clarity, it should be stated that the TEMPER-MS is an IT management system and not a process or paper-based system.

Response: Line 29, text changed from “an RBM system” to “an RBM tool (computer software)” to clarify this is a computer program.

3. Lines 37-40 summarise the requirements of the management system (MS). It is not stated why a bespoke system was required, or whether a search was made for an existing MS, either commercially available or within another trials unit, that would meet all or some of these requirements. Possibly, while many RBM systems exist, none would support the functionality of matching triggered to non-triggered sites in this way. Clarifying this point would support the rationale behind developing a bespoke system, which must have required considerable funding and resources.

Response: A bespoke system was regarded as the best option because it allowed flexibility to meet all study requirements, including the particular matching algorithm; also, the unit has the expertise to develop validated computer systems, and the ability to use the bespoke systems in routine practice outside of TEMPER made the investment worthwhile.

Text added, lines 50-52: “Developing a bespoke system was regarded as the best option to meet all study requirements, some of them (including the matching process) very particular to TEMPER. There was also expertise at MRC CTU to develop a validated computer system to meet these requirements”.

4. The use of the penalty p (a constant to weight poorer performing sites in order to select a non-triggered match) is unclear, since the unweighted scores "two candidate sites A (site score = 0) and B (site score = 2)" are already different. Lines 161-164 do not explain the way p was chosen sufficiently well to remove any suggestion of bias, particularly as the statistician used a subjective judgement in this process. However, it is clear that once chosen, p was applied consistently.
Response: The text the reviewer is referring to: "two candidate sites A (site score = 0) …”, was trying to explain the use of the penalty rather than the penalty factor ‘p’, and it was indeed badly positioned in the text causing confusion.

The text has changed (lines 163-167) to make a bit clearer the use of the penalty and the penalty factor ‘p’:

“matching score =distance+penalty

penalty= ([site score]*p)

By increasing the matching score value of the candidate site, the penalty decreased its eligibility as an untriggered site, proportionally to the site’s score. A penalty factor ‘p’, a proportionality constant, was introduced to determine the weight of the site score in the final matching score calculation.”

We have also tried to clarify the process of selecting the values of ‘p’, and why this could have varied by trial. Ultimately, it was a subjective assessment, but only during testing; as the reviewer notes, in the live study, the values were not changed and thus manipulation of matched sites by this means was avoided.

Changed text in lines 167-175: “The optimal value of p would not necessarily be the same for each trial as the number of triggers assessed, and frequency with which each trigger fired, varied across trials. The value of p for each participating trial was determined by the TEMPER Statistician (SS), based on testing a range of values of p for each trial and making a subjective assessment of the adequacy of the matches selected in terms of the matching factors, the matched site score and the difference in site scores within the p pairs. While p was chosen in a subjective manner, it was then fixed at the end of testing, and applied consistently to all selections in the live study; it could not therefore be used to manipulate matched site selection”

5. Lines 145-149: please state whether this calculation used the actual time since first opening or the natural logarithm.

Response: Line 156. Text added to clarify it is the natural logarithm of months since first randomisation

6. As each untriggered site was visited as one of a matched pair, and could (if each site was visited no more than once) be removed from the list of sites still available to visit, did the
reducing number of available sites for matching cause any problem with subsequent matching? How closely matched were the matched pairs, by the end of the study?

Response: Once a site was selected as Untriggered site, it was removed from the ‘pool’ of eligible sites for that trial, however this did not appear to lead to problems finding adequate matches, perhaps because, at any given time, the number of potential untriggered sites was generally much larger than the number of potential triggered sites.

7. Line 183 - 189: It is not stated how many sites were included in the three trials, therefore these appear to the reader to be large numbers (Figure 4 also). It is not stated whether triggers were only evaluated when a trigger meeting was due (every 3-6 months). Without this information it is hard to see the significance of comparing the number of evaluations to the number of firings, and the reason for including Figure 4.

Response: Number of sites and trigger meetings per trial, added to legend of Figure 4 (end of the document). We feel this figure is of interest because, for example, triggers that never (or always) fire are not discriminatory.

Triggers were evaluated in full before each planned trigger meeting. In addition, ad hoc trigger reports could be run if required. Text added to lines 200-202: “triggers were evaluated before each planned trigger meeting and additionally as required to find matches for a site chosen for a triggered visit between planned meetings (as might occur if, for example, a serious protocol or GCP breach was identified).”

Discretionary revisions and questions:

1. Line 166-169: It is not clear why the algorithm was not sufficient for the allocation of sites to matched pairs; i.e. why there had to be a process of choosing (although explained in Stenning et al, 2018 and examples are given); the implication from the wording is that bias in site selection may have been introduced here.

Response: Text added (lines 178-180) to make clear the highest ranked candidate site for visiting was normally selected, unless there was a documented reason not to, with a reference to Stenning
et al (2018): “The highest ranked candidate (with lowest matching score) was selected by default as the untriggered match; exceptions are described in Stenning et al [10]”

2. Line 187 may be misleadingly worded, as not every trigger fired, according to Figure 4.

Response: Text has changed (lines 199-200) from “All triggers fired between 19 Apr 2013 and 13-Nov 2015…” to “During the course of TEMPER (19 Apr 2013 to 13-Nov 2015) triggers were evaluated…”

3. In the Discussion, ”Evaluation of triggers” is fair. The MS was able to isolate the contribution of each trigger for prognostic analysis. Traffic light assessments are potentially a useful refinement and are used in the presentation of data from commercial RBM IT systems.

No response required.

4. Line 304-312: Some of this functionality may be available already in a commercial system.

Response: True. We intend to research on existing systems that have this functionality to use as a benchmark. However, we have the resources and the expertise to develop our own tailored computer systems, which will give us the flexibility to adapt it and evolve to our own needs, and the compatibility to link with our current systems, which are mostly also built in-house. We have therefore not amended the text at present.

5. Please check the pages cited in Reference 1.

Response: Reference type changed to “Web page”. This now includes the URL and removes the pages shown before.
Reviewer #2

This report serves a dual purpose, it details a triggered site monitoring approach to trial monitoring - relevant to trialists potentially interested in adopting such an approach - and it describes how this approach was used to pair-match sites within the TEMPER trial, which is of independent methodological importance. The report is well written and timely, especially so, given the persistent uncertainties around risk based monitoring despite it mandated use in ICH GCP E6[R2].

I have several minor comments, none of which I consider critical but which may enhance what the reader can take from the paper.

1. Across the 3 studies 23 Trigger Meetings were held at between "3-6 month intervals". Was the frequency of meetings pre-determined or what factors influenced its frequency (baseline risk assessment/results of prior site visits/available resources etc)?

Response: The trigger meetings were pre-determined to be held approximately on a 3-monthly basis as this reflected typical practice by the trial teams. This in turn reflected past experience and resources. Some aspects of site conduct are effectively monitored continuously – for example those aspects reflected in the manual triggers – and an event such as a serious protocol breach could lead to an urgent triggered visit. On such occasions an ad-hoc trigger report could be run to identify a control (untriggered) site.

Additional wording added lines 59-60: “This frequency reflected typical practice by the trial teams according to the stage of the trial (e.g. in recruitment or follow-up).”

2. What was the extent of the monitoring that took place at the selected sites (100% source data verification or otherwise)

Response: We have added reference to Stenning et al (2018), in which we state: “To maximise similarity between triggered and untriggered monitoring visits, they were all conducted according to the trial’s monitoring plan with the same planned checks at all visits ... These were broadly similar across the trials in the study: monitoring usually included SDV on a sample of
patients and review of consent forms, pharmacy documents and facilities, and Investigator Site Files.”

New text added in lines 32-33 to point the reader to more detailed information on the study: “The results of the TEMPER study have been reported by Stenning et al [10] and further details of the study conduct and included trials are explained there.”

3. The selection of appropriate Triggers play a critical role in the described approach. As different triggers fired between 0 and 79% of occasions, it would be interesting to add the frequency of firing to Table 1 or to categorize this frequency e.g. common/uncommon/rare. Providing the additional trigger parameters, of Trigger Weight and Frequency might also be of interest to the reader.

Response: Table 1 has now a column showing the firing rate of the trigger.

Trigger weight functionality, as mentioned in the Discussion, was “available… (but) not used by any of the trial teams during the TEMPER study”, apart from High Recruitment triggers which have a weight =0, and were used as for-information-only triggers.

Frequency was also underused, with a couple of triggers using a value = 0.5 (and only one trigger at the end of the study), while the rest using the default (=1).

4. The authors indicate that the trial specific triggers were based on existing trial practises and in their discussion describe the potential future development and evaluation processes for triggers. However any additional insight, based on their experience within TEMPER, that the authors can provide to the reader as to factors that may or should determine the selection of triggers would be of use to trialists adopting their approach.

Response: We are conscious of not wanting to repeat elements of the discussion that were included in the main results paper, and feel this is such an area. However we have now explicitly mentioned that this topic is addressed in line 240: “see [10] for further discussion of this point”

5. I found the description of the inequality rule L72-L79 confusing, in part because I wasn't initially sure of what was meant by the sample and population and found L73 -74 particularly opaque, also I am not sure that "<>" has an standard definition as compared for example to "vs.".
This may be easier to follow when in the published version, Figure 2 will be in proximity to this description. However, this might be clearer if L72 was clarified, for example "…relationship between a given trigger threshold and the quotient of the sample meeting this threshold over the relevant total number of assessments of that trigger."

Response: Text has been modified and reordered (lines 78-83) to give a bit more clarity about the sample, population and the inequality symbol ‘<>’:

“(inequality rule) … is the relationship between a given trigger threshold and the quotient of a metric Sample over a Population,

Sample/Population<>Threshold

where the Population is the relevant total number of assessments of the observed metric, the Sample, generally a subset of the Population, is a sample of the metric, and the inequality symbol ‘<>’ denotes either ‘<’, ‘≤’, ‘>’, or ‘≥’.”

As the reviewer suggested, the proximity of Figure 2 may also help in making the concepts less obscure.

6. Which triggers if any were assigned a weight of 0 for the study, and did their performance justify the future use of these triggers with a more influential weight.

Response: Exploratory high recruitment triggers were used in two trials, with trigger weight of 0, to allow further assessment of “high recruitment” as a predictor of findings in secondary analyses.

“We note that high recruitment … although commonly used as triggers, were not of clear prognostic value.” (Stenning et al 2018)

7. In calculating the score how was repeated occurrences of the same trigger during the same assessment captured e.g. if 2 medications errors had occurred since last assessment in the same trial.

Response: It all depends on how the narrative and the threshold of the trigger are defined. Having a “medication error” metric, we could specify the trigger narrative “If N or more
medication errors occurred in the site”, where N is the trigger threshold. If N=1, the trigger would fire for either 1 or 2 occurrences during the same assessment.

8. Can the authors comment on the choice of variables used for the matching process in pairing triggered with non-triggered sites. Is there any validation of the predictive power of the used variables with regard to trial performance or protocol compliance over time.

Response: The choice of variables was based on discussion within the study development team. This has been clarified in the text, lines 141-142: “These ‘matching’ factors were chosen through discussion by the study development team.”

We were not aware of any reliable pre-existing validation of their predictive power. As some checks carried out on monitoring visits are done on all patients (e.g. checks on informed consent), matching on this factor avoided bias in outcome assessments within pairs due to number of patients at site. Matching on time since opening meant sites had experienced the same protocol amendments, which may in turn have affected compliance.

9. The number of sites chosen for monitoring per trigger meeting was 1.83, was the number of sites chosen during any one Trigger Meeting dependant on the absolute site score or what determined this.

Response: There was no explicit limit on the number of sites that could be chosen at any given meeting; the number of sites to be visited as triggered sites was guided by the site score but also influenced by trial team resources for planning, conducting and following up site visits. Clarified in lines 208-210: “the number of triggered sites chosen at a given meeting was predominantly guided by the absolute site scores, but also took account of the trial team resources”.

10. To what extent did sites within a trial vary in their scores over time. Did such temporal trends influence triggering.

Response: This analysis of temporal trends in trigger scoring was not done in the study so we cannot report at this time. We agree that it might be interesting and are considering a project looking into this area, possibly with one of our MSc students this summer.
11. As the 'penalty' was fixed per trial (L164], it does influence the ordering of site scores at a given Trigger Meeting or the selection of the matched site - as the ordering of potential matching sites by their scores would be the same for [distance +site score] as for [distance +(site score * 'penalty' factor)], given that the penalty factor is a constant. The purpose of penalty factor would therefore appear to be exclusively to compensate for across-trial heterogeneity and to facilitate the combining of data from the 3 different trials used in the study. However the description of how the penalty factor was derived in L161-164 is quite vague and appears subjective. How effectively it standardises the results from across the 3 trials and allows for a fair aggregation of the trial specific data is unclear. Given the difficulty in validating the effectiveness of the 'penalty' it would seem preferable to present trial specific data for each of the 3 trials in Table 2 as well as the aggregate data derived from the use of the penalty as is currently shown.

Response: The purpose of the penalty factor ‘p’ was to reduce the chance of a candidate site being proposed as a match, which had too high a trigger score, to be genuinely considered a site not of immediate concern, i.e. a potential untriggered site.

This proportionality constant influences how much the candidate site’s score weighted in the final matching score and therefore it did influence the ranking of the candidate sites to be chosen as untriggered sites.

The ordering of candidates by [distance + site score] is the same as for [distance + (site score * 'penalty' factor)], only if the penalty factor equals to 1, and it is increasingly different as its value decreases.

Lines 163-167 have been changed to make it a bit clearer the use of the penalty and the penalty factor ‘p’. Please also see the response to Reviewer #1, comment #4 (of essential revisions), which hopefully gives some further clarification.

12. L233 "Trigger scores informed rather than mandated the selection of triggered sites" What additional factors were taken into account in deciding this and what percentage of cases was the triggered site determined by team decision rather than by the trigger-score approach. Was there a difference in the matched monitoring findings between triggered sites and those in which the selection was manually informed?

Response: Some additional notes on the factors that affect which and how many sites were selected for visits has been added (lines 254-255): “(team … decide how many sites to visit at that time, depending) on the trigger scores, any additional external information on sites (such as staff turnover or concerns raised in other trials) and …”
Trial teams would use trial-specific context to help pick sites to visit, both from within the trigger data (e.g. amongst sites with the same trigger score, those with a particular trigger such as general concern may be prioritised)) and outside (e.g. known recent protocol violations, or high turnover of staff). Resources for conducting visits at a given time also played a part. TEMPER was designed to test a pragmatically used triggering system, of which this “human” element in the process was a part. It is not possible for us to retrospectively assign a site as having been chosen purely on score or with the aid of other information.

13. L335 Conclusion, the statement (L335) ....."Temper -MS implemented an approach of standardizing current practise triggers...", is misleading. Trigger formats and triggering processes were standardized but the triggers themselves varied considerably between studies which, as mentioned in the future direction, remains an important challenge to the proposed approach.

Response: Text added to line 354: “The TEMPER-MS implemented an approach of standardising the automation of current-practice triggers...”