PEER REVIEW HISTORY

BMJ Paediatrics Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Validation of a classification system for treatment-related mortality in children with cancer
AUTHORS	Hassan, Hadeel; Rompola, Melpomeni; Kinsey, Sally; Glaser, Adam; Phillips, Bob

VERSION 1 - REVIEW

REVIEWER	Sung, Lillian The Hospital for Sick Children, Toronto, Canada Competing interests: Developed TRM system and have frequently collaborated with Dr. Phillips
REVIEW RETURNED	27-May-2017

GENERAL COMMENTS	This is a nice study by Phillips et al. that evaluates a newly developed TRM classification and cause-of death attribution system. Please see the following comments. 1. The major issue is that the TRM system does not apply a primary cause of death because of the difficulties in making such a designation. Rather, the system assigns whether a factor is a probable or a possible cause of death. In the system, it is typical for a patient to have multiple probable causes of death. Thus, Table 2 should be redone to show all the probable causes of death. If in this study respondents were asked to identify one cause of death, the manuscript needs to be transparent that this is not how the TRM classification was meant to be applied. I suggest the manuscript finding and discussion be re-framed given this information. For example, the suggestion to have “respiratory infection” as a cause of death does not make sense since in the TRM attribution system, both respiratory and infection can (and usually are) concurrently listed as probable causes of death. 2. Many of the challenges identified can be rectified with the use of SOPs. It may be useful to highlight that such SOPs have now available on-line: https://www.sungresearch.com/trm-training-manual/ 3. Table 1 is not faithful to the original system. Please remove this Table and rather, refer to the original publication or replicate it as this summarization will likely cause confusion for users in the future. I have the same concern regarding Figure 1. 4. Discussion is a fair reflection of the challenges faced in developing a system which is reliable and valid but not perfect. I also agree that modification for use in palliative care may be useful but such a process should have input from palliative care physicians and researchers.

REVIEWER	Harron, Katie LSHTM, UK Competing interests: No competing interests
REVIEW RETURNED	11-Jul-2017

GENERAL COMMENTS	This manuscript aims to evaluate a classification system for treatment-related mortality, in a new population. The aim is clearly stated and justified. Overall the paper is well written, but the way the statistics are presented could be clearer. I have a few specific points. 1. "validation" in the title and "evaluation" throughout the text are used interchangeably – It would be helpful to be specific about what the objective is – i.e. to evaluate criterion validity (i.e. how well the system works at predicting the correct outcome) or reliability (between raters / time scales). If the former, it must be clear what the gold-standard is. The text states that the consultant results were considered gold-standard, but what about the two cases they disagreed on? 2. In relation to the point about, it is unclear how the 10 TRM deaths were classified as such, given the disagreement between consultants. (e.g. second paragraph of results). 3. It is not appropriate to present an overall kappa statistic for all 60 reviews, since each review is counted twice for each reviewer (using 2 or 4 weeks). 4. The order in which the records were given to the reviewers should be stated – i.e. was only the 2 weeks of data presented first, followed by additional data from the previous 4 weeks? This is important as it could influence the intra-rater reliability, especially since all cases were reviewed on the same day. 5. It is difficult to tell how the results presented in Table 3 are derived – this could be improved by a clearer description of the statistical methods for each comparison, and by presenting the number classified as TRM deaths by each reviewer. It is not clear what is meant by "independent reviewers" – this could be made more clear by providing a footnote to indicate which kappa statistic is being used for which comparison. 6. It is also unclear how the comparison between CRAs and consultants was performed, given there were disagreements within each group. 7. The sample size calculation needs to be re-written, as it is unclear what is meant by the "null hypothesis". 8. Where percentages are presented, numerators should able be provided. 9. Table 3 and Table 2 are in the wrong order. 10. The discussion should include a strengths/limitations section.

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1

1. The major issue is that the TRM system does not apply a primary cause of death because of the difficulties in making such a designation. Rather, the system assigns whether a factor is a probable or a possible cause of death. In the system, it is typical for a patient to have multiple probable causes of death. Thus, Table 2 should be redone to show all the probable causes of

death. If in this study respondents were asked to identify one cause of death, the manuscript needs to be transparent that this is not how the TRM classification was meant to be applied. I suggest the manuscript finding and discussion be re-framed given this information. For example, the suggestion to have "respiratory infection" as a cause of death does not make sense since in the TRM attribution system, both respiratory and infection can (and usually are) concurrently listed as probable causes of death.

Thank you for highlighting the differences between the systems used in our study to attribute a cause of death in comparison to the reality of clinical work, where multiple causes of death can be identified. We have amended the manuscript to highlight that reviewers were asked to identify the primary cause of death, rather than multiple causes, in the following:

Lines 31-32 (methods in abstract): "When TRM occurred, reviewers applied the cause-of-death attribution system to identify the primary cause of death".

Lines 37-39: "Reviewers disagreed on the primary cause of death (e.g. respiratory versus infection) when applying the cause-of-death attribution system in 6 cases and probable and possible causes in 4 cases."

Lines 104-105 (methods in main manuscript): "For cases assessed as TRM, the reviewers were asked to apply the cause-of-death attribution system (supplemental file 1) to identify a primary cause of death."

We have also amended the discussion section to highlight the differences of the studies stating the following:

Lines 267-268 amended to state: "Reviewers failed to agree on a primary cause of death in 6 episodes and probable and possible causes in 4 cases".

Lines 214-223 amended to state: "In this study reviewers attributed death to one probable or possible primary cause. Initially, the cause-of-death attribution system was developed to list concurrent causes of death. During the development of this study, we decided to limit the number of causes identified for simplicity. However, reviewers found it difficult to identify a primary cause of death and to distinguish between probable and possible causes.

Since the development of this study, a standard operating procedure TRM web-based tool has been published (https://www.sungresearch.com/trm-training-manual/) and includes working examples. Use of this tool when delivering the training package should help clarify how to use the cause-of-death attribution system and minimise misunderstanding. Currently, the web-based tool is available in English, having the tool available in other languages could potentially reduce confusion and improve harmonisation across clinical trials."

2. Many of the challenges identified can be rectified with the use of SOPs. It may be useful to highlight that such SOPs have now available on-line:

https://www.sungresearch.com/trm-training-manual/

Thank you for identifying the SOP TRM training manual. We have highlighted this in the discussion section in lines 220-223: "Since the development of this study a standard operating procedure TRM web-based tool has been developed (https://www.sungresearch.com/trm-training-manual/) and includes working examples. Use of this tool when delivering the training package should help clarify how to use the cause-of-death attribution system and minimise misunderstanding."

3. Table 1 is not faithful to the original system. Please remove this Table and rather, refer to the original publication or replicate it as this summarization will likely cause confusion for users in the future. I have the same concern regarding Figure 1.

Thank you. We have amended figure 1 (according to the figure used in the SOP TRM training package) and have referred to the supplemental file 1 rather than table 1.

4. **Discussion is a fair reflection of the challenges faced in developing a system which is reliable and valid but not perfect. I also agree that modification for use in palliative care may be useful but such a process should have input from palliative care physicians and researchers.**

Thank you. We have amended the section in the discussion (lines 261-263) to state: "Another proposal includes using a separate classification tool for patients on palliative care trials (fig. 2). The development of such tool would be possible through collaboration with palliative care physicians and researchers."

Reviewer: 2

1. **"Validation" in the title and "evaluation" throughout the text are used interchangeably – It would be helpful to be specific about what the objective is – i.e. to evaluate criterion validity (i.e. how well the system works at predicting the correct outcome) or reliability (between raters/ time scales). If the former, it must be clear what the gold-standard is. The text states that the consultant results were considered gold-standard, but what about the two cases they disagreed on?**
   Thank you for identifying this inconsistency. We have amended the abstract objective lines 22-24 to state: "To evaluate the reliability of the newly developed consensus-based definition of TRM and explore the use of the cause-of-death attribution system outside the centre where it was initially validated (Toronto, Canada)." We have also amended the objectives in the main section (lines 83-85) to state "This study aimed to evaluate the reliability of the newly developed consensus-based definition of TRM and explore the use of the cause-of-death attribution system at a regional paediatric oncology centre in Leeds, England."

2. **In relation to the point about, it is unclear how the 10 TRM deaths were classified as such, given the disagreement between consultants. (E.g. second paragraph of results).**
   Thank you we have amended line 142 to state: "Ten deaths (33%) were identified as TRM by at least one reviewer."

3. **It is not appropriate to present an overall kappa statistic for all 60 reviews, since each review is counted twice for each reviewer (using 2 or 4 weeks).**
   Thank you for your comment, we counted this as 60 reviews as the assessors were supplied with 2 different anonymised and randomised sets of notes (using 2 or 4 weeks) and therefore could potentially have different outcomes assigned by reviewers between the two durations for the same clinical records. Table 1 demonstrates the differences in the kappa statistic for two and four weeks.

4. **The order in which the records were given to the reviewers should be stated – i.e. was only the 2 weeks of data presented first, followed by additional data from the previous 4 weeks? This is important as it could influence the intra-rater reliability, especially since all cases were reviewed on the same day.**

Thank you for your comment. We have amended this section (lines 94-97) to state: "Thirty patient records were included. Copies of the clinical records, with information from both 2 weeks prior to death, and with the information extending back to 4 weeks prior to death were anonymised. This resulted in 60 sets of anonymised case-notes (30 patients, each with 2 time periods) which were presented in a different random order for each assessor."

5. **It is difficult to tell how the results presented in Table 3 are derived – this could be improved by a clearer description of the statistical methods for each comparison, and by presenting the number classified as TRM deaths by each reviewer. It is not clear what is meant by "independent reviewers" – this could be made more clear by providing a footnote to indicate which kappa statistic is being used for which comparison.**
   Thank you for your comment. We have amended lines 110-114 of the methods section to state "Group consensus classification between and within the CRAs and Consultant group was evaluated using the Cohen's kappa statistic, and across all individuals using the Fleiss' kappa statistic. The strength of agreement was defined as slight (0.00-020), fair (0.21-0.40), moderate (0.41-0.60), good (0.61-0.80) and very good (0.81-1.00) [8]."
   We have added a footnote to table 3 (now table 1) to highlight whether the Fleiss' kappa statistic (comparison between all 4 reviewers) or the Cohen's kappa statistic (comparison between two groups- the CRAs/physicians or CRA vs physicians) in lines 183-184 stating:
   "*calculated using the Fleiss kappa statistic (between 4 reviewers)
   **calculated using the Cohen's kappa statistic (between 2 reviewers or 2 groups)"

6. **It is also unclear how the comparison between CRAs and consultants was performed, given there were disagreements within each group.**
   We are sorry this was unclear, and thank the reviewer for the opportunity to clarify. We have amended methods section lines 116-118 to state "A numerical code was used to combine agreement/disagreement between the individual consultants (TRM was recorded as "0" and non-TRM outcomes were recorded as "1" in Excel. When calculating inter-relater reliability between the CRAs and consultants if disagreement between individuals was recorded then the outcome was recorded as "2".

7. **The sample size calculation needs to be re-written, as it is unclear what is meant by the "null hypothesis".**
   Thank you for your comment. We appreciate this was poorly explained. We have amended the method section on lines 119-121 to state "A sample size of 27 deaths determined whether k was good (i.e., ≥0·61), with a power of 0·80, and two-sided α of 0·05 and assuming that treatment-related mortality accounted for 20% of deaths."
8. **Where percentages are presented, numerators should able be provided.**
   Thank you for your comment. We have included numerators with all percentages recorded

9. **Table 3 and Table 2 are in the wrong order.**
   Thank you for identifying this, it has been amended (now Table 1 and 2) and altered in the results section lines 149, 151, 153, 177 and 186.

10. **The discussion should include a strengths/limitations section.**
    Thank you for this suggestion. We have amended the discussion section lines 195-223 to state:
    "Strengths of the study
    This study is, to the best of our knowledge, the first revalidation of the standardised definition of treatment-related mortality and cause of death attribution system for paediatric cancer

patients [7]. It demonstrates that the system is reliable and established its validity in an alternative centre and health care system with different treatment protocols. It can be used after very limited training, with “very good” agreement between assessors irrespective of discipline (Fleiss kappa 0.92, 95% CI 0.83-0.98). The study confirms the observations of the development group and shows that information from two weeks prior to the death of a patient is sufficient to consistently attribute death to TRM or disease.

Limitations of the study
“Although consultants’ opinions are considered gold-standard, in this study we identified how even experienced clinicians may disagree on use of the algorithm. Consultants disagreed on the classification of death in two cases- this may have occurred due to the individual consultant’s clinical experience, or previous contact with the patients. Even though the cases were anonymised and randomised the physicians may have recognised the patient due to their potential clinical involvement in direct patient care. The differences identified highlight how the TRM classification tool is unlikely to ever have perfect agreement between reviewers irrespective of experience, and clinical, and scientific knowledge.

In this study reviewers attributed death to one primary probable, or possible, cause. Whilst developing the study protocol, we decided to limit the number of causes of death for simplicity. However, reviewers found it challenging to identify only one cause of death, and distinguish between probable and possible causes.
Since the development of this study, a standard operating procedure TRM web-based tool has been published (https://www.sungresearch.com/trm-training-manual/) and includes working examples. Use of this tool when delivering the training package should help clarify how to use the cause-of-death attribution system and minimise misunderstanding. Currently, the web-based tool is available in English, having the tool available in other languages could potentially reduce confusion and improve harmonisation across clinical trials.”

VERSION 2 – REVIEW

REVIEWER	Harron, Katie LSHTM, UK Competing interests: no competing interests
REVIEW RETURNED	05-Oct-2017

GENERAL COMMENTS	The authors have addressed all of my previous comments.

VERSION 2 – AUTHOR RESPONSE