Author’s response to reviews

Title: Severely malnourished children with a low weight-for-height have a higher mortality than those with a low mid-upper-arm-circumference: I. Empirical data demonstrates Simpson's paradox

Authors:
Emmanuel Grellety (Emmanuel.Grellety.Bosviel@ulb.ac.be)
Michael Golden (Mike@pollgorm.net)

Version: 2 Date: 14 Feb 2018

Author’s response to reviews:

Dear editor,

Thank you for reviewing our manuscripts. Below we include a point-by-point response to the comments and questions raised by the reviewers.

Best wishes

Emmanuel Grellety & Michael H. Golden

Reviewer #1:

Reviewer 1 simply made some general comments on paper 1 and did not comment on papers 2 and 3. For this reason our responses to this reviewer are added as an addendum to this document.
Reviewer #2:

This series of three papers presumably reflects a single paper that has been submitted in the past and returned to be redrafted.

R2 A-1: This is correct.

The papers represent a detailed analysis which seeks to rebut an ongoing debate in the literature around the most suitable approach to identifying children with severe acute malnutrition in order to provide most appropriate care. It hinges on the relative importance and use of the different WHO criteria, mid-upper arm circumference, weight for height, the presence of nutritional oedema, separately or together.

R2 A-2: This is correct: We are advocating that the current WHO recommendations are correct and should not be changed, as such it is not a rebuttal of current recommendations, but puts the onus on those that wish to change the current WHO recommendations to demonstrate that their simplified approach will do no harm. Our papers demonstrate that this is not the case.

Paper 1.

There are two important points addressed in the first paper. Firstly, the relative risk of mortality associated with each criterion on its own, and when it is combined with the other criteria. The authors have analysed a body of data to which they have access, some 76,887 aged from 6 to 60 months, of whom 3,558 died. The second point is statistical and seeks to clarify the basis for different interpretations which are considered in the literature stating greater benefit of MUAC over weight for height for a range of reasons.

R2 A-3: This is correct.
1. The analysis of the data show clearly that each criterion carries its own risk of mortality and any approach that selects one over another will fail to adequately identify all the children at significant risk. Further it shows that the risks are interactive, rather than simply additive. Simply combining the groups without a better understanding of the biological/pathological basis for these differences may not be appropriate.

R2 A-4: We totally agree with this statement and thank the reviewer. In particular, we agree that simply combining the groups as has been done with most other papers addressing this topic (paper II) “may not be appropriate”, for the reasons we have demonstrated. Editor: It would be helpful if this sort of sentiment could be included in an editorial in the Journal.

2. Simpson's paradox is an accepted statistical principle and although it may be a common cause for analytical concern it is not widely familiar to many carrying out clinical-related research. A clear concise articulation of the basis of the problem it identifies at an early stage of the manuscript would have been helpful for the usual reader in following the detailed information and arguments that follow. A simple statement along the lines that it is a special or extreme form of confounding which identifies how groups, (or criteria) might, or might not, be combined, for analysis would be of value in.

R2 A-5: Thank you. We have introduced a simple statement.

3. Page 4, line 40; should it be "underestimated" for "overestimated"?

R2 A-6: We thank the reviewer. However, our statement is correct; the reason why prevalence may have been overestimated is because of measurement error (Grellety E, Golden MH. The Effect of Random Error on Diagnostic Accuracy Illustrated With the Anthropometric Diagnosis of
Malnutrition. PLoS One. 2016;11(12):e0168585). Nevertheless, as stated, this error is of much less importance than the underestimation of the annual burden because the estimates are based upon point prevalence and not incidence – so that the annual burden can be about 4.8 (wide confidence interval) times as high as the prevalence would suggest. (Isanaka S et al. Improving Estimates of Numbers of Children With Severe Acute Malnutrition Using Cohort and Survey Data. Am J Epidemiol. 2011;173:932–40). This is a general introductory statement to emphasise the importance of the topic so we do not think that it needs to be referenced in the manuscript.

4. Page 7, line 130; this is a tortuous introduction to the second a priori hypothesis.

R2 A-7: Thank you. The primary hypotheses have been itemised to clarify the objectives. We have left the section on age as it is usually assumed that MUAC will primarily select younger children and WHZ older children and as younger children have an inherently higher risk of death the assumption has been that this is the main reason why MUAC has a better ROC curve for mortality – as we wished to explore the relationship between mortality risk and age group we introduced the reasons to question this inherent assumption.

5. Page 10 identifies a range of issues around judgements on data handling, which are welcome. However, it would be of value to have some sense of the magnitude of these issues, or frequency of occurrence.

R2 A-8: We were attempting to be completely transparent with our procedures – and did not indicate the magnitude of the issues – this was a mistake. The numbers of exclusions was fairly trivial – but as now mentioned in the “limitations” section we cannot be sure of the extent that the data was “cleaned” prior to obtaining it, although we were given multiple copies of most of the data and inspection of various “versions” did not reveal indication of prior cleaning. We have included a flow chart of our data handling to show the magnitude of the effect of our data handling.
6. In the Figures, it is not clear what the ALL column in red represents. Elsewhere in the text it is used with a meaning that does not fit here.

R2 A-9: Thank you. In the figures this represents ALL patients (IPF, OTP and SFC) as in table 2. This has been made clear in script and the legends to the figures.

Paper 2.

This paper offers a systematic review of the literature of all the other papers which relate the WHO criteria for SAM to outcome, but where the information is not sufficiently complete to contribute to Paper 1. In this way the authors seek to consider all the possible information that might argue against the thesis which is propounded in paper 1.

R2 A-10: This is correct. Thank you – precisely what we attempted to achieve.

The data and information are well organised and making it possible to follow the variable nature of the information presented and the changes in practice with time and geography. It is not possible to check all the data, but by and large the argument is persuasively made that none of this information differs in a way that would cause major concern for the interpretations offered in Paper 1.

R2 A-11: We thank the reviewer for this comment and are pleased that he/she finds the argument persuasive. We have up-dated the paper in light of the other reviewers comments and we think have made it more persuasive with the meta-analyses and comments on each paper. The important thing is to show that the literature upon which policy has been changed is not sufficiently robust to allow the proponents to put many thousands of lives at risk.
1. Abstract line 20. Here and elsewhere this term is used and it is difficult to see how criteria can be additive for mortality (it is only possible to die once). If possible an alternate expression would be preferable. Presumably the risk of death is increased by having more than one risk factor. Is this additive?

R2 A-12: The risk is augmented if more than one diagnostic criterion is present. We have used the term “additive” to apply to the risk of death and not to the actual death itself. We have made the necessary changes to clarify this, and added a comment that some of the data is compatible with a synergistic rather than a simple additive interpretation.

2. Table 1. It is not clear what "Both sdy" and "Both Com" mean.

R2 A-13: The heading has been changed and also added to the footnote to the table. We apologise for the omission.

Paper 3.

This is a paper which takes available data to carry out a modelling simulation. This draws on the data and interpretations of Paper 1 and Paper 2. The consequence of using different criteria on the number of children who are not likely to die through not receiving the appropriate attention is assessed. Against the criteria for characterising SAM (MUAC and SAM), the two factors of importance are variations in case load and case fatality rate. The model allows an assessment of the practical implications of strong recommendations that have been made to only use MUAC as the anthropometric criterion for identifying children with SAM. If the logic of these papers is to be follow this has serious ramifications which are increasingly being discussed in the literature.

R2 A-14: This is correct, thank you.
1. Abstract, lines 10, 17, 18, 20, 24. Here and elsewhere there is generous use of "relative". At times this has clear technical implications, at others it has lack of clarity or allows ambiguity. Some care here would make it easier for the reader to follow.

R2 A-15: Thank you, we have revised the text and most of the “relatives” have disappeared.

2. The discussion of this Paper makes important points, but it is overlong and at times tends towards the polemical. The reader would benefit if this were concise and to the point.

R2 A-16: We have shortened and focused the discussion omitting some repetition and tangential points. We have also inserted a calculation of the actual numbers of children globally and in India that would be excluded to put the problem into context for the reader. It is now much more irenic but we have retained the section that addresses the reasons the MUAC-only lobby to have managed to persuade some governments to officially change policy. We consider that there are important lessons to be learned from this experience for policy makers. EM was recently in Bangladesh doing a survey of mortality and nutritional status of the refugees from Myanmar, and was forbidden to measure weight and height in the nutritional survey by the Government of Bangladesh. CDC did a similar survey and found that the SAM rate was 7.5% by WHZ and 0.7% by MUAC. This is a massive difference and illustrates our concern about only measuring MUAC – this policy will leave thousands of severely malnourished treatment in the refugee camps without any treatment. Perhaps an editorial to accompany our paper could highlight this example (contact "Bilukha, Oleg (CDC/CGH/DGDDER)” obb0@cdc.gov for details of this survey). We feel that it is important to suggest why this policy has been adopted by many people and countries. The pro-MUAC lobby have countered by suggesting that the cut-off point for MUAC should be increased to capture the WHZ children that the current cut-off point misses. We feel it is important to address this alternative change in policy in the discussion pending publication of a formal analysis of the consequences of such a change (manuscript in preparation).
3. There are no references for this paper.

R2 A-17: This was a problem with the submission – when we recognised that the references were missing from paper III we tried to resubmit – but the system used by Nutrition did not allow this, we wrote to the editor, and to the Editor in Chief but did not get a reply which allowed us to resubmit.

Taken as a whole these are important papers which contain a body of detail that at times appears repetitive eg Paper 3. The work needs to be accessible to an audience that would value the principles being enunciated without being overburdened by all the detail offered to justify the journey. This may be challenging but is important if the main message which is being made is to command the support it seeks.

R2 A-18: Thank you. We have tried to remove as much of the repetition as possible –shortened and focused this paper. Unlike papers I and II which focus exclusively on the assessment of mortality rates for the different presentations of SAM, in paper III we address the practical consequences by examining the mortality of all those who would be excluded from treatment. Thus, this paper’s thrust is quite different from the first two papers. We have deleted a large section of the discussion which addressed points made elsewhere.

Although the three papers address the problems of a MUAC-only program and the assertions that have been made to justify this position, each paper is made to be “stand-alone” at the editor’s request (the original submission combined all papers). This accounts for some of the repetition.

Reviewer #3:
Paper I:

This is an important paper as it confirms that children with a low WHZ are at increased risk of death, whereas many nutrition programs around the globe screen malnutrition with MUAC only.

R3 A-1: Thank you.

Major Comments

- The authors mention "The data from all the IPFs, OTPs and SFCs were combined to give three separate datasets as individual facilities did not contain sufficient deaths to allow for meaningful statistical analysis". Simply pooling results from various studies/settings is inadequate, because such approach does not account for the clustering of data (and one can expect high intra-correlation within clusters given the differences in countries, years, service organisation, refeeding strategies, etc…).

R3 A-2: These were not studies per se. They are program data where the data from separate consecutive admissions were recorded by operational staff or the authors, from the individual patient records. Nevertheless, the reviewer’s comment has merit. We have now included a flow chart to indicate the number of different IPF, OTP and SFC sites that were included. Each of the three datasets is quite separate and no individual child is counted in both datasets; a small number moved from one facility to another and these were excluded from the sending program. This is made clear in the flow chart and manuscript. In view of the large number of sites (and standardisation of the guidelines), we consider “clustering” to be much less of a problem than it could be.

There are two options here. If the authors have data aggregated per country at their disposal, they should present a meta-analysis (this can be done for both case fatality rates and relative risks) where results of individual studies are weighted according to sample size and number of events. A
forest plot would be a very visual way of showing the variations in CFR/RR across country. Heterogeneity of the case fatality rate across studies (even within strata of mode of treatment) should be investigated (possibly by meta-regression). If the authors have individual participant data at their disposal, which seems to be the case, then multilevel statistical models (random effect: country) with adjustment for covariates (age, sex, …) should be used. Such statistical models would also allow formally testing interactions (e.g. interaction between mode of treatment and the relative risk of death).

R3 A-3: We thank you for this critical and very important comment. We have attempted to address is issue with a meta-analysis. Using sub-groups to show differences when we group by region, type of treatment facility and oedema status. As the results for marasmus from East Africa were unexpected, it shows the wisdom for this suggestion.

However, we were not able to use country level data because there were insufficient events for the critical outcomes in many countries (and none of the West African countries even when combined had enough deaths for these M-muac and M-whz – the critical variables). there were sufficient deaths in the oedematous groups to include West Africa, but the contribution was small. As explained for another reviewer, we consider comparison of MUAC and WHZ related deaths in the oedematous children to also be a different but biologically valid comparison of the differences between these two anthropometric indices.

We found that by far the highest CFR was in the children satisfying both criteria, but inclusion of these children, as we have shown leads to mathematical coupling and so all M-both children were excluded from the main analysis, this reduced the number of deaths available very considerably – and is the probably reason why all other authors have included this group in their analyses (paper II).

- Not enough information is provided on the characteristics of the refeeding programs per country: year, admission criteria, refeeding protocol, co-morbidities, etc…Such information would be useful to interpret the results, e.g. the variations in fatality rates across settings. If the authors have such information at their disposal, they should insert a table with study characteristics.
The feeding protocols for the three types IPF, OTP and SFC are each different but are standardised across all countries and follow the guidelines quoted closely as virtually all the children were being cared for by several international NGOs. This is the main reason for comparing the three modes of treatment as the same results would indicate that the mode of treatment is not a major variable (note that the meta-analysis shows that the modes of treatment by region give broadly the same result for each region, with East Africa being different from the other regions). We do not have information, unfortunately, on other confounding variables such as HIV or breast-feeding status, pneumonia, diarrhoea etc. for sufficient numbers of children to make an unbiased estimate of the effect that this has on the data presented. However, such factors are not considered in admission decisions – but do determine whether they are treated in IPF or OTP. This has now been discussed in the limitations section of the paper.

- The proportion of defaulters is substantial in some settings. If they have that information at their disposal, the authors should provide the characteristics of defaulters (how much did they differ from remainders?). If not, this should be discussed as a study limitation.

R3 A-6: Thank you. This is an important point which we failed to make in the initial analysis. We have addressed this is the discussion section on limitations, and presented a table showing the proportion of defaulters by diagnostic group to allow for judgements to be made concerning the importance of this potential verification bias.

- Why are both relative risks and odds ratios reported? The authors should opt for only one measure as they convey the same information.

R3 A-7: Thank you. We have removed second figure from the paper and the tables. We have however used the OR for the meta-analysis.
It is a merit of the study to present results stratified by the presence of oedema. However, as nutritional oedema is considered a stand-alone criterion of SAM, the relative risk of K-muac, K-whz and K-both should be computed in comparison with Kwash, not with M-muac.

R3 A-8: Thank you this is a very useful comment. We have included the RR of K-muac, K-whz and K-both using kwash as the reference in table 2 with the confidence intervals.

The reason why we also compared the oedematous groups with M-muac is to demonstrate that oedema is a major confounder relative to MUAC and WHZ without oedema.

One could also wonder if the stratification by presence of oedema should be reported at all in this paper, as 1. the objective of the paper is about comparing mortality risk associated with a low MUAC vs. a low WHZ; 2. WHZ is not an accurate indicator of malnutrition in the presence of oedema; 3. nutritional oedema is considered a stand-alone criterion of SAM. The authors could consider concentrating on oedema-free children in this paper, which would help focusing the message and the discussion.

R3 A-9: In addition to the reason above, there three further reasons why we wish to retain the data on oedematous children.

First, the importance of including oedematous children’s analysis is in the interpretation of the papers reviewed in paper II. Most other authors have not excluded oedematous children from their analysis, this section demonstrates that oedema is a major confounder and that it is particularly relevant as the effect is different in children with a low MUAC and a low WHZ.

However, the reviewer is correct in stating that oedema will increase WHZ. In most cases by about 3% (Golden MH The clinical assessment of oedema implications for feeding the malnourished child. Eur J Clin Nutr 1989 43 581-2, Grellety Y Management of severe malnutrition in Africa PhD thesis U Aberdeen 2000).
Second, in consideration of the biological and physiological differences in children with a low MUAC or WHZ (if there is a real difference, which has not been demonstrated before), we can treat oedema as a confounder and examine whether there is a difference between K-muac and K-whz that mirrors a difference between M-muac and M-whz. This comparison is not of practical importance for deciding which children will be admitted or not, but was included to highlight possible physiological differences between these groups of children – i.e. does having a low MUAC have a different meaning and interpretation from having a low WHZ? If so then they are not proxies for each other.

Third, it was a completely unexpected finding that having a low MUAC did not augment the risk of death of children who had oedema, in contrast to children with oedema and a low WHZ. This is a completely new and practically important observation, which highlights the potential for there to be a meaningful difference biologically between these two anthropometric measures. It also addresses an important practical controversy. Some observers consider that children with mild or moderate oedema can be safely treated as outpatients, whist may clinicians and countries consider this to be dangerous and such children should be initially treated as in-patients. The two groups of health workers have generally been identifying children by different anthropometric criteria which appears to have informed their judgements – and our data supports both points of view and helps to address this controversy.

-The discussion section contains very interesting points. However, this section needs to be better structured (for example the Simpson paradox is explained at length in one section, and then referred to again in the section on ROC curves). On line 427, the authors state "Apart from confounding, co-morbidity, bias and the stochastic nature of prognostic models there is another major problem with the analysis of the data for WHZ and MUAC: that of "mathematical coupling". The discussion could be structured along those elements.

-On line 439, the authors state "but also confounding due to the presence of oedema, HIV, convulsions, measles and other biases that affect children with a low MUAC and WHZ differently." The authors should discuss how these confounding factors may have affected their own results.
R3 A-10: Thank you, we agree and have re-structured the discussion, without omitting critical information. It is now in a much more structured and logical order. We have also added problems with confounding of our own data to the limitations section of the discussion.

Minor comments

- In the methods section, the authors report on data management (age, oedema, sex not recorded). It could be useful, as an indicator of data quality, to report if lacking information affected some particular datasets.

R3 A-11: Thank you, this was an omission. We have added a flow chart to indicate the level of data cleaning we undertook. And also discussed this in the limitations section of the manuscript.

- Table 3a-3b (significances levels) should be integrated with table 2

R3 A-12: Table 2 is already very wide and the comparisons do not fit easily into the rows of the table. We initially tried to incorporate the significance levels but failed to find a satisfactory solution to formatting the data unless a further 5 columns are added to the table. The table would then be far to wide to fit onto a page of the article.

- Some typos remaining, e.g. "complimentary" (l415), "is" (instead of "in"; l483)

R3 A-13: Thank you these have been rectified.

Paper II:
On the basis of below elements, I consider that the paper is not (yet) in a sufficient stage of preparation. The authors might consider turning the document into a concept paper or a position paper given their stated objective, or to transfer the main elements in the discussion of paper I.

Major comments

- The paper does not apply the PRISMA statement recommendations for reporting systematic review (http://www.equator-network.org/reporting-guidelines/prisma/), which makes a correct review of it quite impossible. The authors are invited to fulfill the criteria of the PRISMA statement.

R3 A-14: We thank the reviewer for this comment. This comment has led to a major revision of the paper. We have fulfilled the PRISMA recommendations and completed the PRISMA check list. We would point out that this is not a “normal” systematic review – because we have ignored the analyses done by the authors of the original papers because of the flaws in most of the papers and have gone back to the actual data and re-analysed the papers. Most systematic reviews combine the author’s analyses rather than reject them.

In view of the flaws and heterogeneity we had considered that a full meta-analysis would not be appropriate. However, prompted (appropriately) by this reviewer, we have tried to circumvent this problem by doing meta-analysis on selected sub-groups of papers to estimate the sources of bias, error and confounding etc. This allowed us to eliminate several papers, although we acknowledge that the analysis still retains papers of doubtful value. We have assessed each papers bias potential and used a quality score in our attempt at amalgamation. However, the aim of this section was to show whether there were any substantial differences between our empirical data and the results from the published literature. We think that we have now achieved this. We hope that the paper now meets this referee’s approval.

- The objective of the review is unclear "to examine the evidence in the published literature that is relied upon to assert that a MUAC-only program is ethical". The authors should redefine their
objective. Is it to pinpoint the reasons why MUAC is presented as a better predictor of death in community programs, in contrast to what they observed in therapeutic program in their paper I?

- The rationale to include altogether very different types of study is not straightforward.

R3 A-15: The objective has been re-written. We aimed to be as comprehensive as possible and gather together all the published data that address the problem of the relative mortality of MUAC and WHZ children. We do not wish to open ourselves to the criticism that we have been selective in our appraisal of the extant data by simply ignoring papers which others rely upon to promote MUAC-only programs. Nearly all these papers have already been quoted by the advocates of MUAC-only programs as supporting their campaign – it is important to review all of these papers that are used (misused) and point out the problems with them, as well as re-analysing the raw numbers as far as possible.

As the authors themselves state "the data from individual datasets were combined, despite the fact that different diagnostic criteria were used in many of the studies, and similarly analysed. [...]The heterogeneity of the standards used, the admixture of oedematous cases and the failure to account for confounding, such as TB, HIV, and non-nutritional conditions, makes amalgamation of data from the different studies problematic and the absolute numbers computed and compared should be considered to be approximate at best".

R3 A-16: We totally agree that aggregation of the data is problematic and for that reason we have removed that section of the analysis entirely. We were not happy at all with this way of examining the data but needed to do some sort of synthesis – and initially we did not think that a meaningful meta-analysis was possible. This has now been rectified by sub-grouping.

The authors could have focused their review on studies with current standards and allowing disintegrating MUAC-only, WHZ-only and MUAC-WHZ, in children without oedema, and exclude other studies as inappropriate for this study.
R3 A-17: The problem is that the studies that have not disaggregated muac-only, whz-only and both are the very papers that are principally relied upon by the proponents of muac-only programs. As said above, to eliminate them from this review would then open us to the criticism of ignoring community based programs (see review I’s comments) etc. and exercising selection bias (see quoted comments from Briend in response to reviewer 1). One of the main critics of using WHZ (Mark Myatt) even posted on EN-net in response to a systematic review, commissioned by WHO, that the selection of studies incorporated in the review “amounted to fraud”. These are the people who are driving global policy on MUAC and WHZ!

- Although referring to published literature, the data presented in the companion paper I of this series and from an unpublished presentation were also included in the analysis. The former accounts for a substantial proportion of the information included. It is therefore not surprising that the authors state that "conclusions drawn from our empirical data (paper I) are supported by the published reports…”.

R3 A-18: We agree with this statement and the data in paper I have been removed. Thank you this is a very useful comment.

- Results and discussion are mixed together. There is no presentation of overall results. The discussion is organized in themas and only the results of some studies are picked up and discussed.

R3 A-19: We have separated results and discussion. We have described each of the papers in terms of the variables presented and have written a commentary on each paper (Additional file 2). We have scored the papers for potential biases (additional file 1). These quality scores are used in the final meta-analysis from which the conclusions are drawn. And a judgement made as to the most reliable set of results is given at the start of the discussion.
The statement: "In general the mortality rate was higher in those children fulfilling the WHO2006 WHZ criterion than the MUAC criterion alone" (l 18-19) is not substantiated by the results presented in table 3. Moreover, as in paper I, to reach the total the authors simply added the numbers from individual studies instead of making a meta-analysis (which accounts for the respective weights of the individual studies).

R3 A-20: We agree with this comment. It was higher, but not significantly, in the set of papers that used the WHO standards, excluded oedema and separated children with both deficits. However, in none of the analyses were there pooled significant differences. So our final conclusion is that WHZ and MUAC are not statistically different in terms of the extant published literature.

Minor comments

-Referring the reader to studies highlighted with different colors in the tables makes the reading extremely complex

R3 A-21: We have removed all the colours from the table and referred to the paper numbers and authors.

Paper III:

Major comments

-The paper convey an important message for policy-making: both the mortality risk associated with the each nutritional indicators and the case-load associated with each indicator must be considered. This said, this is a general rule (the authors could refer to the concept of Population Attributable Risk Fraction), although it might be useful to emphasize it given the terms of the current debate around MUAC vs. WHZ.
R3 A-22: We thank the reviewer for this critical point. Although we recognise that it is a general rule to examine statistics such as attributable fraction this has not been the case with any of the proponents in this debate, which has focused only on the difference in CFRs. As this reviewer points out, this is inadequate to allow for a policy change. We totally agree and are in accord with this reviewer. We have also redrafted the discussion to include a short comment on Population Attributable Risk Fraction. Although our data does not allow this statistic to be generated, having the ratio of SAM categories in the case-load and the relevant CFRs would allow this statistic to be generated at country level with available data.

-There are some apparent contradictions in the methods section. First, the authors state "The ratio of S-muac: S-whz: S-Both reported in the patient data differed significantly from that found in representative community surveys of malnourished children. For that reason the ratios of the number of children in the S-muac, S-whz and S-Both categories reported in papers I and II were not used in any calculation." (l 142-145). They also state in paper I that "The ascertainment bias indicates that the data do not reflect the children with SAM in the community and disproportionally describes the experience of more severely affected children than are generally found during a community survey" (1489-491). Notwithstanding, the authors used the mortality rate of those children in their computation. Second, they present datasets from the literature as "flawed in various ways", but still used the figures from these datasets.

R3 A-23: The comment about not using the patient data only pertain to the ratios of S-muac, S-whz and S-both – i.e. the case-load, and not to the relative mortality rates. We have re-drafted this section of the methods to make this clear. Thank you for pointing out that our original statements were open to misinterpretation. It is important to realise that the absolute magnitude of the mortality rates does not affect the proportion of children that are excluded with a MUAC-only policy, only the ratios between the CFRs and case-load categories. In terms of the literature data, we have now included a meta-analysis of those few papers which used the WHO diagnostic recommendations in our calculations and used only the results from this to inform the ratio of
CRFs. To our knowledge, these are the only estimates of the separate mortality risks for S-muac, S-whz and S-both, which make our computations possible.

-The authors state that "The mortality risk associated with a low MUAC is additive to the risk associated with a low WHZ when children have both deficits" (I 299). In view of results in paper I, the effect seems rather multiplicative.

R3 A-24: Semantics can be difficult. We are not using the term in a strict mathematical sense. The risk of death it is much higher in most studies when a child has both deficits and is very approximately the sum of the risk of death of having one or the other deficit. In some of the data the relative risks seem to by synergistic and we have very cautiously added a sentence to this effect. But we are not at all confident that they are synergistic, which use of the term multiplicative would imply, and this has biological implications that we are not prepared to support at the moment. With the present amount of information available the precise mathematical relationship is not clear.

-The discussion section is rather a plea to use both WHZ and MUAC for diagnosing SAM (which is a fully correct conclusion based on results of paper I), and could be easily integrated in the discussion section of paper I.

R3 A-25: As stated with reviewer 2 we initially submitted this study as a single long and complicated paper and were asked to divide it into three separate papers. However, this paper is different in that it deals with the practical implications of only measuring MUAC and not simply on a comparison of CFRs. The only comment Reviewer 1 makes whilst commenting upon paper I is that comparison of mortality of single deficit children is not related to the practice on the ground because a measure of MUAC will also pick up the children with both deficits. This paper addresses this practical issue. Thus the thrust of this paper and the first two papers are quite distinct.
Minor comments

-What means "inadequate # deaths" in table 4?

R3 A-26: Thank you: The tables have been redrafted.

Reviewer #4:

Manuscript I

The reviewer would like to see a flowchart of the study design. The reviewer understands that data used in this manuscripts come from three sources (IPFs, OTPs, and SFCs). However it will be helpful to know the details.

For examples, the number of children who were re-classified as SAM and abstracted from the SFC database (Line 166), or the number of children who were excluded (e.g., did not meet at least one of the criteria for SAM (Lines 177-179), errors of recording (Lines 179-181), children from SFC who were recorded as being transferred to an OTP or IPF if the data from the receiving treatment facility were available (Lines 202-204), children from OTP who were recorded as being transferred to an IPF if the corresponding IPF data were available (Lines 204-205)), or the number of children who were transferred to other facilities from IPF (Line 206). These are just examples, please provide all necessary details so the reviewer can have a clear picture of the study design.

R4 A-1: Thank you. A flow chart has been added to address each of these points. We apologise for not adding a flow chart to the original submission as it would have clarified these issues much more easily than in the script.

Lines 166-167: Please confirm that all three datasets (IPFs, OTPs, and SFCs) are independent to each other.
R4 A-2: This has been emphasised in the script, and the flow chart makes it clear that when children moved from one mode of treatment to be included in a second dataset, they were excluded from the initial dataset.

Line 187: Please add the number of children who were assigned an age according to their height.

R4 A-3: This is now given in the script and flow chart.

Line 191-192: Please add the number of children who were assigned a sex at random.

R4 A-4: It is now given in the script and flow chart.

Line 230: The reviewer would like to know the reason why the authors used the Marascuilo procedure to adjust for multiple comparisons as opposed to using False Discovery Rate (FDR)-controlling procedures? In addition, the assumption of the Chi-square test is that the expected value in each cell is greater than 5. Please confirm that this assumption was always met each time Chi-square tests were conducted.

R4 A-5: Yes, none of the Chi-squared tests presented had less than 5 expected values. We have added Fisher’s exact test to the tables where appropriate.

We understand that FDR procedures were introduced following innovations, such as genomics, where a relatively small number of subjects had a large number of variables so that when Family Wise Error Rate (FWER) procedures such as Dunett’s or Bonferroni corrections led to almost no results being declared significant. As such they are much less conservative, if more powerful, than
FWER. Thus, we consider that FDR is not appropriate for our data where we have a relatively large number of subjects and a small number of variables.

Thus a FWER is more appropriate. We have used the Marascuilo procedure because we are familiar with this technique, have software to calculate the results and find it an intuitive, useful and robust FWER procedure. I wonder if the reviewer has a criticism of the veracity of the results produced using this procedure. However, we can find no papers in the statistical literature that either criticises or refines this procedure.

Table 2: The totals number of children in "all patients aged 6 to 60m" for the 7 diagnostic categories do not match with the total number of children in Table S1. For example, in Table 2, there are 7191 children in M-muac; whereas in Table S1, there are 8860 in M-muac. Please provide an explanation as to why the n are different in both tables.

R4 A-6: Thank you very much for noticing this error! And we apologise for not picking up on this before submission. The heading labelled M-muac was actually M-muac plus K-muac! The table has been amended and the numbers of patients admitted with the other categories has been added.

Tables 3a and 3b: In order to fully interpret the results in these tables, it is important to complement the p-value results with the following:

- Effect sizes for chi-square test (e.g., Cramer's V)

R4 A-7: Cramer’s V has been added to the tables.

- 95% confidence intervals (some -but not all- 95% confidence intervals were presented in Figures 1, 2 and 3)
R4 A-8: Confidence intervals are now included for all variables.

- Numeric values for p-values ≥0.05 (as opposed to 'ns')

R4 A-9: This makes the table much more difficult to read. And as none of the other reviewers have commented upon this we have decided to continue to omit the actual p values of non-significant comparisons, particularly, as mentioned above, that we have done multiple comparisons and this makes type one errors much more likely.

Table 5: Minor typo in last footnote: did the authors meant "Criterion Y identifies more deaths than criterion X, but when the children with both criteria are included criterion X appears to have higher case fatality rate" instead of "Criterion Y identifies more deaths than criterion X, but when the children with both criteria are included criterion Y appears to have higher case fatality rate" (as currently written in manuscript).

R4 A-10: Thank you – the footnotes have been changed.

Table S1: Only 4 diagnostic categories (M-muac, M-whz, M-Both, Kwash) are presented in Table S1. The reviewer would like to know why the other 3 diagnostic categories (K-muac, K-whz, K-Both) are not presented.

R4 A-11: Table S1 has been amended to include the missing diagnostic categories (and the totals corrected as they were not correct)

Figure 3: The reviewer does not understand what the Y-axis in Figure 3 represents. For example, according to Table 2, the RR for M-whz (all patients, 6 to 60m) is 2.6 and the odd ratio should be
2.61; however in figure 3, it seems the red bar (All) for M-whz is showing 4? Please explain what the Y-axis represents in Figure 3.

R4 A-12: As requested by reviewer 2 this figure has been withdrawn.

Manuscript II

Lines 115-118: The reviewer is confused as to whether the groups S-muac, S-whz, S-Both are mutually exclusive or not?

R4 A-13: Yes they are mutually exclusive – this has been made clear.

Line 123: The reviewer would like to know the reason why the authors used the Marascuilo procedure to adjust for multiple comparisons as opposed to using False Discovery Rate (FDR)-controlling procedures?

R4 A-14: See the previous answer to this same query. However, as referred to by reviewer 3 the whole analysis by combing these studies is itself problematic because of the reasons given and the analysis has been omitted, and meta-analyses on groups of studies that have common characteristics analysed group-wise.

Lines 123-124: The reviewer would like to know whether it was possible to conduct Fisher's Exact tests in the case where there were fewer than 5 children in any expected category.
R4 A-15: Yes, Fisher’s exact tests have been done (where the total numbers have been less than 500 as computation is almost impossible with large sample sizes and very low numbers of events), and Cramer’s V added to all comparisons.

Tables 2 and 3: In order to fully interpret the results in these tables, it is important to complement the p-value results with the following:

- Effect sizes for chi-square test (e.g., Cramer's V)
- 95% confidence intervals (some -but not all- 95% confidence intervals were presented in Figures 1, 2 and 3)
- Numeric values for p-values ≥0.05 (as opposed to 'ns')

R4 A-16: This has been done. Thank you.

Manuscript III

Methods: the reviewer suggests adding the formula for the calculation of the proportion of the total expected deaths of SAM children either in the text (Method section) or as a footnote under all tables.

R4 A-17: Thank you. Although the formulae are quite simple, we have added them to the methods section.

Line 231: The reviewer does not understand where the percentage 92% in "92% of deaths will occur in S-whz children who will remain unidentified in the community" comes from. Please explain.
R4 A-18: We have changed to denominator to percent of children excluded and made this consistent throughout the manuscript. This denominator is more easily understood. The 92% was a typo for which we apologise.

All tables: The reviewer suggests changing the wording as follows: Red: <25%, dark pink: 25% - <50%, Light pink: 50% - <75%, Green: 75% - <90%, Blue: ≥90%.

R4 A-19: Thank you: References to colours have been removed from the text entirely – and left in the footnotes to the tables.

ADDENDUM:

Reviewer #1:

R1 A-0 General response to Reviewer 1: This reviewer has only made a few general and not very helpful comments on our first paper. S/he has not commented/read papers 2 and 3. All the comments are rebutted, and in particular the comment that we have made mistakes in our quotation of papers claiming that children with WHZ < -3Z are healthy is wrong and upsetting.

Major comments

1. Sample of pre-selected malnourished children on unknown criteria

The sample included in this analysis of from inpatient malnutrition treatment facilities (IPFs), Out-patient treatment programs (OTPs) and supplementary feeding centres (SFCs) include participants who are pre-selected from community-based screening into the treatment programme. As such the
participants are all cases (malnourished) and such data is not adequate to make policy recommendations for various reasons

a. These populations do not reflect a general population.

R1 A-1.a: This criticism is rejected.

The general population is of no relevance to whether WHZ or MUAC, fulfilling the criteria for SAM, have a higher or lower mortality rate. The relevant population is the entire population of children that would be eligible for diagnosis and treatment of SAM in the population. We have indeed estimated this population from surveys of randomly selected children in many communities and present these results in paper III (not commented upon by reviewer).

By analogy, if one is looking at metabolic syndrome mortality diagnosed by hypertension or high cholesterol, one would examine the outcome of cohorts of adults with a metabolic syndrome who had either (or both) markers. One would not select members of the general population at random, or worry that the cohorts do not reflect the general population. The argument put forward by this reviewer is spurious in our opinion.

b. When screening and treatment of the sample is predominantly based on one of these criteria (WFH/MUAC) this biases the sample. For example, children pre-selected by WFH should not be evaluated for MUAC and vise versa. This is particularly a concern for comparisons of the association of WHZ and MUAC with mortality as is the question in this study.

R1 A-1.b: We have addressed this point in our paper. The proportion of children with each of the criteria alone, or both, in the cohorts does not reflect the relative proportions of SAM children in most populations as those satisfying both criteria are indeed over-represented. This is acknowledged in the paper and the potential effect of such an ascertainment bias are fully discussed in Grellety E, Golden MH. Response to Briend et al "Low Mid-Upper-Arm-Circumference identifies children with a high risk of death and should be the priority target for treatment". BMC Nutrition 2016, 2-63: 1-12, referenced in our paper. This argument is rebutted.
In our opinion it is disingenuous to suggest that the MUAC of those admitted with WHZ is irrelevant and vice-versa. The children with MUAC-only are presumed to be a representative sample of those with MUAC-only in the community; similarly, for WHZ-only, and those satisfying both criteria. The value of stratifying the patients into these groups is so that each individual group is more representative of that group in the community than the majority of children admitted with SAM. Indeed, it is precisely because the admitted children do not have the same proportion of children with both as seen in the community that many studies of admitted children do have an ascertainment bias when analysed. By excluding the children with both criteria from our comparisons of MUAC and WHZ we have either eliminated or at least attenuated such a bias.

In my opinion, these data cannot therefore be used to answer the question at hand and cannot be used to refute an argument that has been built on general population/community based data.

R1 A-1.c: We disagree with this reviewer’s personal opinion – see comment above on general population. The basis for the argument the reviewer presents is fully explored in paper III and indeed many of the papers reviewed that are widely quoted have a gross ascertainment bias. The data derived from community based studies are found to be insufficient to support the claim made by the reviewer - (note, paper II has not been commented upon by this reviewer).

2. Anthropometric classification is not of practical importance (line 213 to 221)

R1 A-2: We do not understand this comment. Anthropometric criteria (and the presence of oedema) are the whole basis for making the diagnosis of SAM, how could these not be relevant? There are two major criteria – MUAC <115mm and WHZ <-3Z which, by and large, identify different children although some children satisfy both criteria. The criteria that are used in the field are clearly of practical importance as they are the sole basis upon which children are either selected for, or refused, treatment.
a. The definition of a MUAC only criteria as used in this study needs to be stated in the introduction section.

R1 A-2.a: The definitions of MUAC-only, WHZ-only and those with both criteria are clearly defined in the methods section; this in our opinion is the appropriate place to introduce these definitions.

b. If the objective of the analysis is to evaluate currently proposed criteria, the 7 anthropometric classifications defined on page 11 and used as the basis of this analysis would need to be revised to make more practical sense. In policy and in current practice, the MUAC only criteria is defined giving no consideration to WFHZ (criteria one). In the same way, a WFHZ only criteria would not consider MUAC (criteria 2). As presented, the defined criteria are odd, not applicable in practice and results of which cannot be compared to studies that have defined MUAC/WFH only criteria differently.

In my opinion, authors should revise the defined anthropometric criteria used in this analysis to reflect current anthropometric criteria in policy and practice.

R1 A-2.b: We completely refute the suggestion. We have ONLY used the standard WHO criteria for the definition of SAM, which are EITHER/OR a MUAC of <115mm OR a WHZ <-3Z of the WHO standards.

The classification is for analytical purposes. And as mentioned above, the division of the patients into these groups overcomes potential ascertainment bias. In paper III we use various CFRs to compute the children who would be excluded from treatment using the criteria that are applied in practice (i.e. a MUAC only policy will admit both M-muac AND M-both children). The reviewer does not appreciate that when children with both criteria are included in each group (MUAC and WHZ) then the data and analysis are flawed because of extreme confounding rendering previous analyses and the conclusions drawn from the papers problematic. This reviewer has not commented upon papers II or III.
3. Kwashiorkor is not an unresolved criterion both in policy and practice. I would suggest that authors consider removing analysis on kwashiorkor completely from this analysis

R1 A-3: This is answered in detail in response to the other reviewer’s comments.

4. In the discussion section authors have dedicated 2 pages from line 405 to 451 to discuss merits and demerits of ROC curves. This section is a potential diversion from the main objective of this analysis…. to examine relative mortality rates of children who have SAM……. This section focuses on a different question on robustness of statistical methods used in trying to resolve this MUAC/WFH controversy. I would recommend raising this question in this manuscript but giving it a bit more attention and addressed it fully in a separate publication on statistical methods

R1 A-4: The relevance of this topic is that ROC curve analysis has been the only method used in all other publications to compare the risk of death of SAM children. The shortcomings of this approach, when confounding is not considered in the analysis, have not been properly appreciated by the nutritional audience relying upon such data to guide policy. These shortcomings need to be understood by the non-statistical reader to properly evaluate the existing literature and why our findings differ substantially from other papers addressing this subject (see paper II which the reviewer has not addressed).

Minor comments

1. In several sections within the manuscript statements presented as facts are either not referenced or references are contrary to statement. I have isolated some examples below but would urge the authors to review all references carefully

R1 A-5: This is simply not the case – see below for rebuttal.
Introduction

The sentence beginning on line 79 to 81; Children admitted by MUAC show good response to treatment…………, the second section of this sentence is generalized but not referenced.

R1 A-6: This comment is covered by the reference given: the conditionals given are the criteria for outpatient management, which is the mode of treatment used to assess the response to treatment in reference #23. Further references are not required.

The sentence beginning on line 108-110; "A major Assertion justifying the point of view is that children with a low WHZ are relatively healthy (44 - 49) and therefore are not in need of treatment". This statement is not supported by the reference that are provided. I have checked through the following references and could not find evidence to support this statement.


R1 A-7: This accusation is upsetting and quite incorrect. All the references have been carefully read and appropriately quoted. This referee has not reviewed the references given sufficiently closely. The quotations from the cited references that state what we have reported are given as follows:

“For instance, “long legs”, which increases the proportion of children with a low WHZ is found mainly in wealthier families and has been associated with good health”

Ref 44. Hammond W, Badawi AE, Deconinck H.
“lower WHZ (or higher prevalence of wasting), though linear growth reflects good health”

Ref 46. Garenne M, Maire B, Fontaine O, Briend A.
“WHZ is influenced by body proportion, and in particular leg length … This is potentially a concern, as children with long legs, who usually are in better health, are more easily classified as malnourished with WHZ”

Ref 47 Briend, A.
“WFH is lower in children with long legs…this is a concern as children with long legs are usually healthier. Adding WFH to models has only little effect on the estimation of the risk of death. Therefore, there is no clear rationale for admitting these children into therapeutic feeding programmes.”
Ref 48 Taren D, de Pee S.
“MUAC is still more commonly used since increased leg length [whz] is mostly associated with healthy children”

Ref 49 Deconinck H
“Longer legs typically translate into lower WHZ (or higher prevalence of wasting), though linear growth reflects good health”.

Ref 45 has been substituted with this more appropriate ref from EN-Net which include, inter alia, the quotes below: http://www.en-net.org/question/1922.aspx, and http://www.en-net.org/question/1945.aspx (old references 51 and 52)

Myatt M. 7/9/15: “You can expect MUAC-/WHZ+ older children not to have elevated mortality risk” “WHZ is generally useless... as it will fill your program with older and healthy children who are not at greatly elevated (above baseline) mortality risk.”; “We will see reduced WHZ in the healthiest and wealthiest. What this means is that we can expect, for some children - particularly the older children, lower mortality risk with lower WHZ. In your specific setting you can expect MUAC- / WHZ+ older children not to have elevated mortality risk.” 16/9/15 “Low WHZ kids with high MUACs are unlikely to be at high mortality risk and we can assume that most they will survive and grow up”; 18/9/15 “WHZ will select heathy, older, and non-stunted children that do not need CMAM [treatment] services.” 3/10/15 “Clinical data is not very relevant. Robertfroid et al. (2015) review [a WHO commissioned review of WHZ vs MUAC] was highly selective in the evidence presented to the extent that borders on fraud.”

Briend A. 2/10/15: “I fully support Mark [Myatt’s] statement … The lead author [Roberfroid] himself has even just published an article highlighting that MUAC < 115 mm does not identify
correctly children with WFH <-3 (deja vu hundreds of times !!!; just wondering how this can still be published)… a dogmatic attitude in favour of low WFH children eliminates from treatment many high risk children and results in many unnecessary deaths.

2. In the introduction paragraph beginning on line 88 to 100, results of an analysis are presented (REF 42). Authors report the mean overlap found in their analysis and base their subsequent argument on it. Would be useful for readers to get an idea of how these figures compare to what others have found (above, below, similar?)

R1 A-8: The paper referenced (ref 42) gives a comprehensive account of data from 47 different countries. It reinforces what other authors have found in small scale studies, we do not consider it necessary or relevant to repeat a complete review here. However, we have included references 24 to 41 to emphasise that there is general agreement concerning this discordance. We consider this to be adequate, as the paper does not focus on discordance in diagnosis but on relative mortality rates. The data are used in paper III which this referee has not commented upon.

3. Sentence starting on line 93 to 94; "There was no satisfactory explanations for this phenomenon". This sentence is speculative. Can the authors provide the reader with what the current explanations are for this phenomenon and then present their argument on how these are not satisfactory.

R1 A-9: These are set out explicitly in the references given – there is little point in repeating these arguments as this is tangential to the objectives of this paper; see [42] for discussion.

4. Lines 185 to 187, the author describes how height measurement was used to define age in participants whose age is unknown.
R1 A-10: See response to the other reviewers. This procedure was used in a trivial number of patients (none of whom died).

a. Please provide a reference for validation of this method.

R1 A-10.a: Not considered necessary. This is standard practice in situations where age is unknown by/ unobtainable from the caregiver. Please see for example "Hall A, Oirere M, Thurstans S, Ndumi A, Sibson V. The practical challenges of evaluating a blanket emergency feeding programme in northern Kenya. PLoS One 2011;6:e26854". There are many more including several NGO guidelines.

b. Explain how stunting was accounted for in applying this method. Shorter children allocated a younger age may just be stunted and hence misclassified.

R1 A-10.b: As explained in the paper this was only used a) for a small number of children and b) to categorise the children into very broad age groups. As age is not critical to the analysis (and reported age is itself very inaccurate) we adopted this approach rather than bias the sample by eliminating data unnecessarily.