Author’s response to reviews

Title: Use of Test Accuracy Study Design Labels in NICE’s Diagnostic Guidance

Authors:

Maria Olsen (m.olsen@amc.nl)
Zhivko Zhelev (Z.Zhelev@exeter.ac.uk)
Harriet Hunt (H.A.Hunt@exeter.ac.uk)
Jamie Peters (J.Peters@exeter.ac.uk)
Patrick Bossuyt (p.m.bossuyt@amc.uva.nl)
Christopher Hyde (C.J.Hyde@exeter.ac.uk)

Version: 1 Date: 23 May 2019

Author’s response to reviews:

Reviewer 1:

(…)

The manuscript raises a very important point, namely that there is inconsistency in the terms used to describe study design and study design features in the analysed reports, and that similar terms have been interpreted in very different ways.

Reply: We thank the reviewer for these supportive comments.

Although the paper has clearly highlighted and illustrated the lack of consistency in terminology by analysing NICE Diagnostic Guidance and evidence reports, it does not yet offer a way forward. The authors might already have considered how to develop a framework (or nomenclature) for improving terminology related to DTA study design, and it would be highly valuable to incorporate this in the manuscript, and take it beyond establishing the problem only.

Reply: We fully agree with the reviewer that the manuscript lacked a perspective on how to improve the current practice.

We have now added a supplementary table that emphasizes more on what terms that should be avoided, with corresponding alternative terms, to limit the confusion that we found in the
primary analysis (Supplementary table S4). We believe this can help improve the current communication of DTA studies. This is added to the results section (pg. 7, second paragraph) and in the Supplementary table (S4).

From previous exploratory work we learned that the development and evaluation of a complete framework with nomenclature is a complex objective. We plan to launch a separate project to do is, building on the descriptive, empirical study reported in this manuscript.

Reviewer 2:

This is an important paper, drawing attention to the muddled terminology in diagnostic studies. I have very few suggestions as I think it is a good paper as it is.

Reply: We thank the reviewer for the kind comments.

I would like it to go that bold step forward and suggest that the DTA research community could use table 2 as a framework for studies to identify and report on key features of study design.

I particularly like table 2 and offer the following suggestions. I think this table could be worked up into a framework for what should be clarified in describing a study, as it is your derivation of concepts that need to be described. So it would become the start of a way forward, more as part of a reporting guideline with recommended terminology? Maybe you will think this one step too far for this article, in which case I would still change its concept to define what needs to be described. At the moment I think it describes the muddle, rather than being a valuable contribution to the start of straightening it out?

Reply: We thank the reviewer for this encouragement; we fully agree that discussing “a way forward” is valuable, as also addressed by reviewer 1.

We are planning to do develop a more complete framework, with recommend terminology, using a Delphi-like process, in a separate project. However, to improve the current manuscript, and as an intermediate step, we have added a supplementary table (S4). Although, instead of suggesting what specific features that authors should describe, we have highlighted the specific terms that lead to confusion, and thus, should preferably be avoided and replaced with other existing, to limit the current confusion.

We believe that a further definition, addition/selection, and ranking of features and terms should be done, using more rigorous methods, and not just by this team of authors.
Table 2 title - Should it be something like key design features.

Reply: We have now rephrased the title to “Identified key design domains and features”.

However, to keep the focus on the primary objective (i.e. the range of labels), we have moved this table (previous Table 2) to supplementary (now as Supplementary table S3), also as we believe this is more coherent with the new Supplementary table S4 (highlighting confusing terminology).

Number of index test - ? change to different feature heading as this is not the same as comparative concept (direct or indirect) as a study could report on a number of tests, but not aim to compare them.

Reply: We agree that a study evaluating multiple index tests is not per se a “comparative” study and that this example may not be confusing. We initially aimed to illustrate that the “comparative” label/term was used to describe a study that included multiple index test. We have adjusted the feature heading to “Number of evaluated index test, incl. type of comparison” in the table to make this more clear.

Who receives the different tests - are you including the concept of randomisation to reference test - or only to different index tests? What does comparator refer to here - seems to overlap to "head to head" index tests.

Reply: In the previous Table 2 (now Supplementary table S3), randomization was not included as a part of this design feature, because it was identified separately. We acknowledge that this may not have been entirely clear. We have now clarified this in Supplementary table S3 by separating participant flow and test group allocation (that includes randomization).

We agree that the term “comparator” is unclear. However, for this example we think it still illustrates that participants are allocated to different tests. We have instead deleted “incl. head-to-head” to avoid confusion.

Order of tests - a few more words are needed to describe

Reply: This has now been rephrased to “The sequential order of tests performed”.

Adding a few words in Table 2 Eligibility heading. Suggestion "Eligibility: number of included groups"
Reply: This has been adjusted.

Eligibility - inclusion criteria - do you mean "Asymptomatic, symptomatic, high risk groups"

Reply: In the previous Table 2, we intended to illustrate that this feature (heading) was derived from studies that had used "Asymptomatic, symptomatic, high risk groups" as labels for describing parts of the study design related to ‘key inclusion criteria’. We have now reworded the feature heading to “Eligibility criteria: Key inclusion criteria” (Supplementary table S3).

Can we move away from the wording "consecutive" as in my experience there is no such study. I have used the following wording in a recent paper "recruited consecutive (i.e. unselected) eligible patients." Meaning they were as consecutive as possible, whilst not 100% consecutive as I do not believe this is ever possible. I don't really like my wording, as they were selected by the inclusion criteria - so maybe there is something better?

Reply: We recognize that truly consecutive sampling of participants may only be achievable in rare circumstances. However, we believe that despite that cases can be missed when aiming to recruit participants consecutively, the term and practice still clearly distinguish the different sampling types (consecutive versus random versus convenience sampling). Both a random and a consecutive plan can result in a group of unselected eligible patients. The term is also widely used and probably well understood term by the general DTA community 1,2.

Recruitment setting - maybe split up single centre and location - this is two important concepts in one heading.

Reply: We agree and have now separated these features to ‘Centres’ and ‘Location/Care’, respectively.

Data collection "conduct/time of the index and reference standard" - this one is tricky. In many cases it is about when the sample was taken for a biomarker, rather than when the sample was unfrozen and the test done. Maybe this just needs a footnote, that where stored samples are used, this is relative to the date of sample collection.

Reply: We fully agree that this feature is difficult. Yet, at this stage the feature and descriptions of ‘data collection’ varies and with no clear consensus on further details and respective terminology. For this reason, we prefer to keep this interpretation as simple as possible but have
added “planned and/or collected” to the right column, and also removed "conduct/time of the index and reference standard" to avoid the need for further explanation.

Analysis - can you make this clearer? What are you meaning by discordant case analysis (you pointed out the two definitions in table 1)? Post-hoc analysis of prospective data - sometimes this is OK if the data were collected with a protocol and is fit for purpose.

Reply: We agree that the post-hoc analysis may be appropriate. However, we do not intend to rank or deem what features are appropriate at this stage. We have now removed the example of “Discordant analysis” from the table to avoid this confusion (S3).

I really like the concept from Ida Sim about protocol driven data collection or non-protocol driven data (databases). Can this be included somewhere maybe at the top, as this is so critical to many of the biases.

Reply: We recognize that information regarding protocol driven or non-protocol driven is important and have added the term of “protocol driven” to the alternative, preferred terminology in Supplementary table S4.

In both Supplementary table S3 and S4, we have ordered the features according to the chronological study order (as they would appear in a protocol) because we believe it is too soon to deem what features are the most important, and this should be addressed in separate study. For this reason, we choose to keep the feature at this place in the table instead of placing it at the top.

References:
