Author’s response to reviews

Title: Safety, efficiency and health-related quality of telephone triage conducted by general practitioners, nurses, or physicians in out-of-hours primary care: a quasi-experimental study using the Assessment of Quality in Telephone Triage (AQTT) to assess audio-recorded telephone calls

Authors:

Dennis Graversen (d.graversen@ph.au.dk)
Morten Christensen (mbc@ph.au.dk)
Anette Pedersen (afp@ph.au.dk)
Anders Carlsen (andershelles@gmail.com)
Flemming Bro (fbro@ph.au.dk)
Helle Christensen (helle.collatz.christensen.02@regionh.dk)
Claus Høstrup Vestergaard (cv@ph.au.dk)
Linda Huibers (huibers@ph.au.dk)

Version: 1 Date: 23 Oct 2019

Author’s response to reviews:

Dear Editor Maja Racic,

We submit our revised manuscript entitled "Safety, efficiency and health-related quality of telephone triage conducted by general practitioners, nurses, or physicians in out-of-hours primary care: a quasi-experimental study using the Assessment of Quality in Telephone Triage (AQTT) to assess audio-recorded telephone calls" (FAMP-D-19-00246). Thank you for considering our manuscript for publication and all reviewers for the positive and constructive comments. The valuable remarks helped us to improve the manuscript.

We carefully revised our manuscript according to the reviewer’s comments. We include a point-by-point response detailing our considerations and reactions upon the comments and suggestions raised. We revised the manuscript accordingly, adding a clean copy and an annotated copy using Track Changes to highlight our changes.

We have attached Figure 2 in a separate file for better resolution.
We hope that you will find that the suggestions and criticism have been met in a satisfying way and that the manuscript has improved accordingly.

Yours faithfully, on behalf of the other authors,

Point-by-point comments

Reviewer reports:
(Reviewer 1): Overall, the topic of the manuscript is very good. There are some weaknesses in the methodology, analysis, and interpretation of the data.

- Inconsistent use of terms such as GPs, physicians, doctors; they all could mean the same but authors have used them in different ways. They have explained at one place but for readers, instead of physicians, they could use the term specialists.

• We acknowledge that this can be confusing, as GPs are both physicians and specialists. Thus, we have chosen to keep the definitions GPs and physicians, but have specified the definitions as follows. In the paragraph ‘design and settings’: “Physicians employed at MH-1813 have different medical specialties (e.g. internal medicine, pediatrics, anesthesiology, surgery), and varying experience (including junior physicians), with only a minority being a GP. We refer to this group as physicians in the rest of the article”. We have added a statement about GPs being specialists as well: “In 1992 a reform introduced large-scale GPCs, with GP specialists performing telephone triage”. Furthermore, the typo in Figure 1 has been changes as “doctors” has been changed to “physicians”.

- Ideally, the assessment process should include more than one assessor per call to gauge inter-rater reliability, especially given the evidently lower inter-rater agreement of the AQTT (Graversen et al., 2019). Further, Graversen et al. (2019) found a satisfactory inter-rater agreement of the AQTT only when categorizing outcomes of items into poor quality (rated "1" or "2") and sufficient quality (rated "not applicable", "3", "4" and "5"). The authors cited the article but took a different approach in categorizing the outcomes of health-related items by excluding "not applicable" and "sufficient" rating (rated "3") and categorizing outcomes into poor quality (rated "1" or "2") and good quality (rated "4" and "5"). The decision to alter the dichotomization of AQTT outcomes may affect some parts of the analysis, reporting, and finding.

• We have made some adjustments to clarify our categorisation to avoid confusion. Firstly, the aim of the categorisation good quality in Table 2 was to highlight calls with the best quality (i.e. rated “4” or “5”). However, as adding another categorisation can lead to confusion and this was not an aim of our study, we decided to delete this categorisation.

• When comparing the quality by analysing the RR in dichotomised outcomes of poor quality, we have compared poor quality (rated “1” or “2) vs. sufficient quality (rated “3”, “4” or “5”). It is correct that we handled the category “not applicable” different in the two articles. As described, the answering category “n/a” was used “if an item is correctly found not relevant or available information is insufficient for assessment”. Thus, the assessment “n/a” could both reflect a correct and sufficient quality (i.e. “correctly found not relevant”), but could also
potentially reflect a poor performance (i.e. “available information is insufficient for
evaluation”). Thus, in the present paper managing the category “n/a” as “sufficient quality”
could potentially overestimate the health-related quality. Consequently, we have chosen to
exclude “n/a” from analyses. Post-hoc sensitivity analyses of the inter-rater reliability with “n/a”
as “missing” rather than as “sufficient quality” did not change the ICC reliability estimates
considerably (for all but three items under &lt;0.05) and always towards a higher reliability. We
have added a statement concerning the interpretation of “n/a” to the limitations.

• We adjusted our manuscript accordingly:
  o Statistical analysis: “For health-related specific items, we categorized into poor quality (rated
    “1” or ”2”) and sufficient quality (rated “3”, “4, or “5”). “Not applicable” was recoded into
    “missing”.”
  o Statistical analysis: “We calculated the relative risk (RR) in dichotomized outcome of having
    poor quality (vs. sufficient quality) on the health-related specific items and of clinically relevant
    undertriage or overtriage (vs. not clinically relevant undertriage or overtriage) for the three
    groups of triage professionals, using binomial regression.”
  o Legends of Table 2: “The RR for “poor quality” vs. “sufficient quality (i.e. rated “3”, “4” or
    “5”) was analysed using binomial regression model.”
  o Strengths and limitations: “The management of “n/a” was ambiguous as it could both reflect a
correct performance (i.e. “correctly found not relevant”), but could also potentially cover a poor
performance (i.e. “available information is insufficient for assessment”). In the testing of the
reliability of AQTT “n/a” was recoded into “3”, but for the purpose of this paper, we chose to
exclude “n/a”. Managing “n/a” as “sufficient quality” could overestimate the quality. A post-hoc
sensitivity analysis of the inter-rater ICC reliability excluding “n/a” did not change the reliability
considerably, and always towards a higher reliability”

I have concerns over the statistics used and reported wrong statistical calculations at a few
places. For example, Table 1 says the difference was calculated using chi-squared test and
student t-test. Both of them are incorrect in this situation. Chi-square is used for categorical data
while here the data is presented in Mean, t-test is used for 2 sets of comparison here the
comparison is between 3 so ANOVA would be more appropriate test. In addition, when applying
tests, we need to present data variance such as SD which is also missing here.

• We are aware that an ANOVA analyses would be an appropriate method to compare all three
triage professionals. However, we have chosen to conduct only pairwise comparisons of triage
conducted by GPs vs. nurses and GPs vs. physicians. Although, no gold standard exists
concerning the most optimal telephone triage professional, we preferred to use GPs as a
reference group. The rationale behind this was that GPs have the longest experience with
telephone triage. Furthermore, they have the most extensive knowledge of handling primary care
patients, which is the study population (i.e. patients calling out-of-hours primary care).

• Additionally, we have changed 95%CI into SD as suggested by the reviewer.
  o Statistical analysis: “We used descriptive analyses to describe patient and call characteristics,
stratified by triage professional group. We also conducted pairwise comparisons of calls
conducted by GPs versus nurses and calls conducted by GPs versus physicians, using chi-
squared test for categorical data and student's t-test comparing mean length of call.”
Statistical analysis: “All comparative analyses were conducted pairwise, using GP-led triage as reference group.”

Title of Table 1: “Baseline distribution of patient and call characteristics, stratified by triage professional group”

Legend to Table 1: “*Significant difference between nurses or physicians in pairwise comparison with GPs serving as reference group (p<0.05), using chi-squared test (categorical variables) and student’s t-test (continuous data).”

Length of call, min and sec. (SD in sec)

- In the methods section, authors say "We used descriptive analyses to describe patient and call characteristics stratified by groups of triage 15 professionals, using chi-squared test and student's t-test to analyse differences." This is incorrect as the use of statistical tests are for inferential statistics not descriptive.

• Our description of the analyses is indeed imprecise, as we described our population and we also compared the groups of triage professionals. We have adjusted the text as following:
  “We used descriptive analyses to describe patient and call characteristics stratified by triage professional group. We also conducted pairwise comparisons of calls conducted by GPs versus nurses, and calls conducted by GPs versus physicians, using chi-squared test for categorical data and student's t-test for mean length of call.”

- In Figure 2. the information is hardly visible. Also the term Doctor is given here which has not been used in the text. Are GPs not Doctors?

• Indeed, doctors should be “physicians”. Thank you for drawing our attention towards this.

- In Table 2. authors mentioned General Linear Regression Model which has not been mentioned anywhere else if the data met the assumptions of the model? they also mentioned the word binomial regression model in the methods which is different than linear regression.

We have adjusted this typo, changing it into “binomial regression model”.

I suggest keeping the information as descriptive in the manuscript. Considering a number of limitations of this data set, authors needs to be careful in writing the conclusion about nurses being better than physicians.

• We acknowledge that this study has some limitations. However, we argue that these limitations have been addressed in this revised manuscript, allowing the reader to interpret appropriate. Moreover, the main limitation (i.e. having one assessor per call and not having nurses assessing) is described in the discussion and additional analyses of the similar-to-me bias showed that our crude differences may be seen as the most conservative difference between GP-led and nurse-led triage, justifying our conclusions.
Reviewer 2:

A really interesting research and outcome. I agree with the suggested following studies.

• Thank you for reviewing and finding our research interesting.