PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	The Healthcare Complaints Analysis Tool: Reliability testing on a sample of Danish patient compensation claims
AUTHORS	Bogh, Soren; Kerring, Jonas; Jakobsen, Katrine; Hilsøe, Camilla; Mikkelsen, Kim; Birkeland, Søren

VERSION 1 – REVIEW

REVIEWER	Tom Reader London School of Economics
REVIEW RETURNED	05-Sep-2019

GENERAL COMMENTS	Overall, I thought this was a very solid paper, and have outlined my thoughts on it below. Introduction. The paper effectively sets out the field, neatly summarising the value of studying and learning from patient narratives, and the challenges faced by healthcare institutions to do this. Rather than just providing experience data (i.e., satisfaction), patient-generated information can supplement more traditional outcome data. It then focuses on testing a system for measuring safety and quality information in complaints - the Healthcare Complaints Analysis Tool - and does this within a Danish setting. I think there is potentially a lot of value in applying HCAT in a different setting to the UK: for example, showing it works in a different healthcare system, operates in a different language context, revealing complaint trends in different countries, and showing the transfer ability and comparability of data across multiple contexts. However, I don't think the paper made enough of this: I think the justification and benefit for testing HCAT in Denmark can be developed further, and encourage them to do so. Additionally, the paper focuses on compensation claims rather than complaints: I think this is a strength of the paper. It tests HCAT against other forms of patient narratives, and is a contribution. The authors might make more of this. I think some description of the type of domains/problems (e.g., safety, medical errors) that HCAT reveals would also be useful, as readers may be unfamiliar with this. Indeed, some further justification as to why HCAT was selected, and not another tool, might be helpful. Method and Results In total, 140 compensation claims were analysed, and this seems like an appropriate amount for a reliability study. Complaints were randomly selected, which is good. Complaints were coding electronically, with the English version of the coding system being matched up with Danish words in the letters. Good randomisation of complaint coding. Statistical analysis was well done, and looked

	at the reliability of HCAT. Similar to the original paper reporting on HCAT, "quality" had poorest reliability: what did the authors think were the reasons for this? Is it something inherent in coding quality, or a limitation of the coding system itself. Harm seemed to be coded less reliably in the original HCAT study: perhaps the authors could expand further on why this is the case in the discussion? Could it be a feature of the compensation claims? The paper looks at reliability at the sub-problem level (e.g., the level of diagnosis error): the original HCAT paper doesn't seem to report this, so this is a contribution as well. It might be worth flagging this, and commenting on whether the reduced lack of reliability at the problem level (e.g., for quality) is explained by a group of sub-problems. I thought the reliability data was well presented, and reported upon appropriately and critically. Discussion Very good discussion, comparing the Danish HCAT reliability analysis against the original study. Good critical points around, for example coding harm, however it would have been useful to have more explanation as to why things might have been difficult to code. Again, a key contribution of the paper is using HCAT to code patient narratives of healthcare problems provided through different sources (compensation claims), and it is worth emphasising this. Key critique of the article is that use of an English language version of HCAT: however this is recognised, and discussed. Did the authors, through their analysis, identify ways that HCAT could be improved? Overall this was a solid article, testing an established methodology for analysing healthcare complaints in the English health system to Denmark, testing the reliability at all levels of coding (which was not reported in the original HCAT manuscript), and showing HCAT can be used to reliably code patient narratives provided the compensation claims (i.e., not just complaints). Some more consideration of how this is all useful, and might add to quality and safety improvement would be beneficial for the article.

REVIEWER	Edward Gorgon University of the Philippines Manila, Philippines
REVIEW RETURNED	09-Sep-2019

GENERAL COMMENTS	Overall comments • Use of the appropriate reporting guideline for reliability and agreement studies (GRRAS) can enhance completeness and transparency, and aid interpretation of the study findings. • The paper addressed funding, patient-public involvement, and ethical approval, although did not address data sharing. • The manuscript requires English-language editing and proofreading to improve quality of communication. For example, a number of grammatical and typographical errors can be found across the manuscript. • The abstract was generally accurate except for some parts of Results – for example, a sweeping statement that “stage of care” (moderate point estimate for “operation or procedure”; moderate 95% CI lower limits for “examination and diagnosis” and “operation or procedure”) and “staff involved” (moderate 95% CI lower limits for “medical staff” and “staff unspecified”) has “satisfactory” reliability (when what constituted "satisfactory" was not well-justified in the text).

	Study design • The research question was clear and the study design was appropriate for the research question. The research design has similarities to the design of the interrater reliability part of the original study that reported the development of HCAT by Gillespie et al (2016). • Acute medicine was selected as the practice area of interest although there was no description of why this was used. • Assessors were four academics (3 masters-level, 1 PhD-level). It is not clear why they were chosen as assessors – will their characteristics be similar to those who are intended to routinely use the HCAT in practice? Also, the web-based HCAT form that was used by the assessors was in English while the compensation claims were in Danish, which means that prospective users whose source data are not English require a specific level of bilingual ability in order to use the HCAT at similar levels of reliability. • Assessor training and adherence to the HCAT manual were described. Blinding of the assessors to each other's ratings was reported. Can the authors clarify too if the one assessor was blinded to the previous test administration results during the second test administration (intra-assessor reliability)? Data sources • Data were analysed and coded from compensation claims in this study versus patient complaints which were the originally targeted by Gillespie et al in developing the HCAT. Can the authors provide further justification for this decision apart from broadly saying that compensation claims "emphasise patient perspectives on healthcare quality"? The potential range of data that can be covered in compensation claims may have some overlap but may not necessarily be comparable to that in patient complaints. Statistical analysis • Linear regression was used to calculate the average number of problem categories per claim letter and average time spent per claim letter. How does this relate to the study aim which was to test reliability of the HCAT? • Level of harm data were treated as continuous variable and intraclass correlation coefficients were used. Which ICC model was used? Was the same standard/guideline by Landis & Koch for categorical data used in interpreting the ICCs? Presentation and interpretation of results • Gwet's AC1 statistic with 95% confidence interval was used to report reliability as in the original HCAT study by Gillespie et al (2016). In the original article, Gillespie et al also reported Fleiss kappas which were helpful. Reporting a combination of coefficients provides a better picture of reliability and agreement as opposed to single summary measures (Kottner et al 2011). Therefore, can the authors also report Fleiss kappas, to allow readers to make further comparisons with the only available reliability study on the HCAT? • It would be useful if the authors can clarify the basis for considering moderate agreement (coefficients = 0.41 to 0.60) as "satisfactory". It seems lenient to consider reliability and agreement coefficients lower than 0.60 or 0.70 as satisfactory. A clear justification would be helpful. • To demonstrate sufficient reliability and agreement, the lower limits of 95% confidence intervals deserve attention. For example, although the text reports that "stage of care" had substantial or

	excellent reliability, the point estimate for inter-assessor reliability of “operation or procedure” was only moderate at 0.55 and the lower limits for “examination and diagnosis” and “operation o procedure” for both intra-assessor and inter-assessor reliability extended down to “moderate” territory (0.45 – 0.59). • What is the practical relevance of the overall high reliability coefficients and some poor-to-moderate point estimates and 95% CI lower limits in relation to the purpose and consequences of data generated from the HCAT in real-world practice? This is not very clear in the Discussion of the study findings.

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1
Reviewer Name: Tom Reader, Institution and Country: London School of Economics
Overall, I thought this was a very solid paper, and have outlined my thoughts on it below.
Dear Tom Reader, Thank you for feedback and thoughts on our paper. We have made a point to point answer below.

a. Introduction.
The paper effectively sets out the field, neatly summarising the value of studying and learning from patient narratives, and the challenges faced by healthcare institutions to do this. Rather than just providing experience data (i.e., satisfaction), patient-generated information can supplement more traditional outcome data. It then focuses on testing a system for measuring safety and quality information in complaints - the Healthcare Complaints Analysis Tool - and does this within a Danish setting. I think there is potentially a lot of value in applying HCAT in a different setting to the UK: for example, showing it works in a different healthcare system, operates in a different language context, revealing complaint trends in different countries, and showing the transfer ability and comparability of data across multiple contexts. However, I don't think the paper made enough of this: I think the justification and benefit for testing HCAT in Denmark can be developed further, and encourage them to do so.
We have tried to elaborate this in the introduction
Until now, however, reliability testing has only been performed in the UK, and the usefulness of the HCAT needs to be further tested in healthcare systems with different organizational frameworks and different language settings.

Additionally, the paper focuses on compensation claims rather than complaints: I think this is a strength of the paper. It tests HCAT against other forms of patient narratives, and is a contribution. The authors might make more of this.
We have tried to elaborate this in the discussion

We focused on compensation claims rather than complaints and thereby tested HCAT against different forms of patient narratives than previously used. We see this as a strength of our study. As our sample represents a narrower spectrum of patient narratives, we anticipated that fewer problem categories would be utilized, but this only seemed to be the case at the sub-category level.

I think some description of the type of domains/problems (e.g., safety, medical errors) that HCAT reveals would also be useful, as readers may be unfamiliar with this.
Thanks for very useful comment. We have clarified this in the text and referred to figure 1.

Indeed, some further justification as to why HCAT was selected, and not another tool, might be helpful.
We have added a sentence in the introduction explaining the rationale of our choice

The HCAT taxonomy is, to our knowledge, the first tool to be based on a thorough review of the literature and developed with a rigorous and transparent method.

b. Method and Results
In total, 140 compensation claims were analysed, and this seems like an appropriate amount for a reliability study. Complaints were randomly selected, which is good. Complaints were coding electronically, with the English version of the coding system being matched up with Danish words in the letters. Good randomisation of complaint coding. Statistical analysis was well done, and looked at the reliability of HCAT. Similar to the original paper reporting on HCAT, "quality" had poorest reliability: what did the authors think were the reasons for this? Is it something inherent in coding quality, or a limitation of the coding system itself.
We have elaborated this in the discussion

The low reliability in the “quality” category might be because judging quality issues is more subjective than, for example, rating complaints about arrogant behavior, which are often directly stated in the letter of complaint. Further, some of the sub-categories in the quality and safety categories can be difficult to distinguish from each other. For example, the ‘Neglect – general’ sub-category under the ‘Quality’ problem (e.g. “Infected wound not attended to”) in some instances may tend to largely overlap with the ‘Error – general’ sub-category under ‘safety”. Such ambiguities about the definition of sub-categories reduced the inter-assessor reliability.

Harm seemed to be coded less reliably in the original HCAT study: perhaps the authors could expand further on why this is the case in the discussion? Could it be a feature of the compensation claims?
First, a minor error in the data management of the harm scored has been corrected, resulting in adjustment of the harm reliability estimates. However, the changes do not change the fact that we showed substantial lower reliability than the original HCAT study. The pre-training did not leave any signs of this, so we were kinda surprised over the low reliability. In hindsight, we might have tended to focus on coding the reliability categories and less on coding the harm.
It is probably too early to conclude that the low estimates are a feature to compensation claims, but it is definitely a possibility that these are harder to code with high reliability.

We have tried to expand this in the discussion.

In contrast to the original HCAT reliability study, the level of harm was less reliably scored in our study. This may be due to insufficient training and calibration of raters as the training may have focused more on achieving high agreement on problem categories. However, establishing the extent of harm is a major challenge in compensation claims relative to complaints about disciplinary responsibility with DPCA decisions about damages in practice also being regularly appealed (22).

c. The paper looks at reliability at the sub-problem level (e.g., the level of diagnosis error): the original HCAT paper doesn't seem to report this, so this is a contribution as well. It might be worth flagging this, and commenting on whether the reduced lack of reliability at the problem level (e.g., for quality) is explained by a group of sub-problems.
When looking at quality sub-categories, the ‘Outcome and side effects’ sub-category ratings appear to be only scarcely reliable. However, we do not think that this lack of reliability at the sub-category level can explain the lack of reliability at the problem level. A reason, however, might be the difficulty to

distinguish the definition of some sub-categories from different problem categories. See our response to your comment above.

I thought the reliability data was well presented, and reported upon appropriately and critically. Thank you

d. Discussion
Very good discussion, comparing the Danish HCAT reliability analysis against the original study. Good critical points around, for example coding harm, however it would have been useful to have more explanation as to why things might have been difficult to code.
Please see our answer at point b. We think we have explained why harm was difficult to code.

Again, a key contribution of the paper is using HCAT to code patient narratives of healthcare problems provided through different sources (compensation claims), and it is worth emphasising this. That's a good point, we have mentioned it under strengths in the discussion section.

Key critique of the article is that use of an English language version of HCAT: however this is recognised, and discussed. Did the authors, through their analysis, identify ways that HCAT could be improved?
Again, we found that some of the sub-categories in the quality and safety categories were difficult to distinguish from each other. This is probably an area with potential for improvement.

e. Overall this was a solid article, testing an established methodology for analysing healthcare complaints in the English health system to Denmark, testing the reliability at all levels of coding (which was not reported in the original HCAT manuscript), and showing HCAT can be used to reliably code patient narratives provided the compensation claims (i.e., not just complaints). Some more consideration of how this is all useful, and might add to quality and safety improvement would be beneficial for the article.
We have elaborated on this in the new paragraph in the conclusion

Our study findings provide support for HCAT as a tool for systematizing patient complaints although the applicability and usefulness of the tool needs to be assessed further. Future studies could explore the value of continuous use of HCAT at management level to indicate areas for improvement, detect sites with poor staff-patient communication, and investigate how organizational changes affect patient experiences. Our study confirms at least moderate reliability throughout the HCAT taxonomy, except for the rating of level of harm, stage of care, and a number of subcategories.

3. Reviewer: 2
Reviewer Name: Edward Gorgon
Dear Edward Gordon. Thank you for your thorough review of our paper. We have made at point by point answer below.

a. Please leave your comments for the authors below Overall comments • Use of the appropriate reporting guideline for reliability and agreement studies (GRRAS) can enhance completeness and transparency, and aid interpretation of the study findings. The reviewer must be acknowledged for drawing attention to Kottner's paper from 2011. We have tried to follow the instructions from Kottner and added a comment on sample size considerations.

We included a random sample of 140 cases completed by the DPCA from 2007 to 2018. Based on previous literature, this sample size should be sufficient (16, 17).

• The paper addressed funding, patient-public involvement, and ethical approval, although did not address data sharing.
We have now addressed data sharing at the end of our paper
• The manuscript requires English-language editing and proofreading to improve quality of communication. For example, a number of grammatical and typographical errors can be found across the manuscript.
Thanks for the comment. We have had the manuscript proofread.
• The abstract was generally accurate except for some parts of Results – for example, a sweeping statement that “stage of care” (moderate point estimate for “operation or procedure”; moderate 95% CI lower limits for “examination and diagnosis” and “operation or procedure”) and “staff involved” (moderate 95% CI lower limits for “medical staff” and “staff unspecified”) has “satisfactory” reliability (when what constituted "satisfactory" was not well-justified in the text).
Thanks for this comment. We have changed the wording in the abstract.

Reliability was at least moderate when coding the stage of care, the complainant, and the staff group involved.

b. Study design
• The research question was clear and the study design was appropriate for the research question. The research design has similarities to the design of the interrater reliability part of the original study that reported the development of HCAT by Gillespie et al (2016).
• Acute medicine was selected as the practice area of interest although there was no description of why this was used.
We have elaborated this in the method section.

To be included in our analysis, the complaint behind the compensation claim must have been provided at a Danish hospital and classified by the DPCA as being within the field of acute medicine. This field is crucial to modern health services (18) and in many instances is the patient’s first contact with the secondary health system. Acute care has been continuously reorganized to meet patient expectations.

• Assessors were four academics (3 masters-level, 1 PhD-level). It is not clear why they were chosen as assessors – will their characteristics be similar to those who are intended to routinely use the HCAT in practice?
Also, the web-based HCAT form that was used by the assessors was in English while the compensation claims were in Danish, which means that prospective users whose source data are not English require a specific level of bilingual ability in order to use the HCAT at similar levels of reliability.
We expect that their qualifications will represent prospective users of HCAT, which is now added to the method section

The assessors were chosen with the expectation that their qualifications would represent potential future users of the HCAT for quality improvements.

• Assessor training and adherence to the HCAT manual were described. Blinding of the assessors to each other’s ratings was reported. Can the authors clarify too if the one assessor was blinded to the previous test administration results during the second test administration (intra-assessor reliability)?
Good point, thanks for that. She was blinded from her previous scores. This has now been added to the method section

To calculate intra-assessor reliability, one assessor scored all cases twice, with six weeks between the first and second assessments (and blinded to the scores). The order in which claims were reviewed was randomized between assessors, who were also blinded to each other's ratings.

c. Data sources
• Data were analysed and coded from compensation claims in this study versus patient complaints which were the originally targeted by Gillespie et al in developing the HCAT. Can the authors provide further justification for this decision apart from broadly saying that compensation claims "emphasise patient perspectives on healthcare quality"? The potential range of data that can be covered in compensation claims may have some overlap but may not necessarily be comparable to that in patient complaints.
Thanks for the very important comment. We have tried to address this at the beginning of the Strengths and limitations section.

We see this as a strength of our study. As our sample represents a narrower spectrum of patient narratives, we anticipated that fewer problem categories would be utilized, but this only seemed to be the case at the sub-category level.

d. Statistical analysis
• Linear regression was used to calculate the average number of problem categories per claim letter and average time spent per claim letter. How does this relate to the study aim which was to test reliability of the HCAT?
This was chosen to further describe the features of the HCAT rating. We expect this information to be valuable for other users, as it provides an insight into the practical use of HCAT. Please let us know if you think it should be eradicated from the manuscript anyway.

• Level of harm data were treated as continuous variable and intraclass correlation coefficients were used. Which ICC model was used?
We used a two-way random-effect model. We have added information about this in the method section.

The level of harm was coded on a scale from 1 (negligible) to 5 (catastrophic) and was treated as a continuous variable; intra-class correlation coefficients (two-way random-effect model) were thus used to test reliability

Was the same standard/guideline by Landis & Koch for categorical data used in interpreting the ICCs?
Yes, the interpretation of the ICC were done according to Landis & Koch classification

e. Presentation and interpretation of results • Gwet's AC1 statistic with 95% confidence interval was used to report reliability as in the original HCAT study by Gillespie et al (2016). In the original article, Gillespie et al also reported Fleiss kappas which were helpful. Reporting a combination of coefficients provides a better picture of reliability and agreement as opposed to single summary measures (Kottner et al 2011). Therefore, can the authors also report Fleiss kappas, to allow readers to make further comparisons with the only available reliability study on the HCAT?
We have not chosen to report Fleiss kappas because it is found to be more affected by prevalence and report more unstable inter-rater reliability than the Gwet's AC1 (Wongpakaran 2013). Our paper already presents many results, and we do not believe that twice as many (and less credible) results will contribute in a meaningful way.

• It would be useful if the authors can clarify the basis for considering moderate agreement (coefficients = 0.41 to 0.60) as “satisfactory”. It seems lenient to consider reliability and agreement coefficients lower than 0.60 or 0.70 as satisfactory. A clear justification would be helpful.
We agree that coefficients lower than 0.60 are not satisfactory. The description of HCAT as satisfactory was only meant to general terms. We have been through the manuscript and revised where it was not sufficiently clarified.

• To demonstrate sufficient reliability and agreement, the lower limits of 95% confidence intervals deserve attention. For example, although the text reports that “stage of care” had substantial or excellent reliability, the point estimate for inter-assessor reliability of “operation or procedure” was only moderate at 0.55 and the lower limits for “examination and diagnosis” and “operation o procedure” for both intra-assessor and inter-assessor reliability extended down to “moderate” territory (0.45 – 0.59).
That’s a really good point. We have incorporated this in the discussion section.

It should be noted, however, that the confidence intervals of some of the reliability estimates extended below the level of substantial agreement.

• What is the practical relevance of the overall high reliability coefficients and some poor-to-moderate point estimates and 95% CI lower limits in relation to the purpose and consequences of data generated from the HCAT in real-world practice? This is not very clear in the Discussion of the study findings.
We have elaborated on this in the discussion

The overall high reliability coefficients with few poor-to-moderate reliability coefficients stresses the need to make ongoing calibration and pre-training before it is put into use.

VERSION 2 – REVIEW

REVIEWER	Tom Reader London School of Economics UK
REVIEW RETURNED	30-Oct-2019

GENERAL COMMENTS	I thought the authors did a good job of revising the paper. They engaged with the issue of sub-problem reliability on HCAT, and also the lack of reliability for harm. The contextualisation of the paper is improved, and the strengths of the study are better accounted for. I do not have any further concerns: well done.

REVIEWER	Edward Gorgon University of the Philippines Manila, Philippines The University of Sydney, Australia
REVIEW RETURNED	27-Oct-2019

GENERAL COMMENTS	The authors have addressed the reviewers' comments as required. I have no further comments.