BMJ Open

The Healthcare Complaints Analysis Tool: Reliability testing on a Danish Patient Compensation claim material

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2019-033638
Article Type:	Research
Date Submitted by the Author:	15-Aug-2019
Complete List of Authors:	Bogh, Soren; Odense University Hospital, OPEN Kerring, Jonas; Department of Psychiatry Middelfart Jakobsen, Katrine; Region Zealand Hilsøe, Camilla; Region of Southern Denmark, Department for Quality and Research. Mikkelsen, Kim; Patient insurance association, denmark, Patient injury prevention and learning Birkeland, Søren; Odense University Hospital, OPEN
Keywords:	Reliability, Healthcare complaints, Patient safety, Health service research

Title

The Healthcare Complaints Analysis Tool: Reliability testing on a Danish Patient Compensation claim material

Authors

Søren Bie Bogh, Jonas Harder Kerring, Katrine Prisak Jakobsen, Camilla Hagemann Hilsøe, Kim Lyngby Mikkelsen, Søren Fryd Birkeland

Affiliations

Søren Bie Bogh, PhD., Researcher, soren.bie.bogh@rsyd.dk

OPEN, Odense Patient data Explorative Network, University of Southern Denmark and Odense University Hospital, DK-5000 Odense C, Denmark

Jonas Harder Kerring, Cand.Psych., jonas.harder.kerring@rsyd.dk. Department of Psychiatry Middelfart.

Katrine Prisak Jakobsen, cand.scient.san.publ., Consultant, katjak@regionsjaelland.dk

Region Zealand, DK-4180 Sorø, Denmark

Camilla Hagemann Hilsøe, Cand. Cur., Camilla.hilsoe@rsyd.dk. Department for Quality and Research. Region of Southern Denmark.

Kim Lyngby Mikkelsen, Kim.Lyngby.Mikkelsen@Patienterstatningen.dk The Patient Compensation Association, Denmark. Kalvebod Brygge 45, 1560 Copenhagen V.

Søren Fryd Birkeland, LLM PhD., Soren.Birkeland@rsyd.dk associate professor, OPEN, Odense Patient data Explorative Network, University of Southern Denmark and Odense University Hospital, DK-5000 Odense C, Denmark

Corresponding details:

Søren Bie Bogh, PhD., soren.bie.bogh@rsyd.dk

Running title: Categorizing claim letters

Word count: Abstract: 240 words, manuscript: 2890 words

Keywords**:** Reliability**,** Healthcare complaints, Patient safety**,** Acute care, Health service research

Abstract

Objective

We aimed to test the intra- and inter-assessor reliability of Healthcare Complaints Analysis Tool (HCAT) for categorizing the claim letters in a sample of Danish patient compensation claims.

Design, setting and participants

We used a random sample of 140 compensation cases completed by the Danish Patient Compensation Association. All compensation claims were filed in relation to the field of acute medicine at Danish hospitals from 2007 to 2018. Four assessors were trained according to the HCAT manual before each assessing the sample independently.

Main outcome measures

We calculated intra- and inter-assessor reliability at domain, problem category, and sub-category levels, and undertook the reliability testing of ratings based on the level of harm and the descriptive details provided by the claim cases.

Results

The HCAT was reliable for identifying problem categories, with reliability scores ranging from 0.55 to 0.99. Reliability was lower when coding the "severity" of the problem. Inter-assessor reliability was generally lower than intra-assessor reliability. The categories "quality" and "safety" were the least reliable. Reliability at the sub-category level was, in general, satisfactory, with only a few sub-categories found to be unusable as a consequence of poor reliability. Reliability was also satisfactory when coding "stage of care" and information concerning the "complainer" and the "staff involved." However, the coding of "harm level" was found to be unreliable (intra-reliability: 0.06; inter-reliability: 0.29).

Conclusion

Overall, HCAT was found to be a reliable tool for categorizing problem types in patient compensation claims.

Article Summary - Strengths and limitations of this study

- A key strength of this study is the testing of the HCAT instrument including its constituent main categories and subcategories in a large sample of complaint cases outside the setting where HCAT was developed
- Likewise, multiple trained raters were used showing high reliability estimates
- A main limitation of the study is that cases only related to injury compensation cases about emergency hospital care
- Additionally, due to skewed distribution of case scorings at the main and subcategory levels, our study cannot stand alone and must be followed by further studies in different health services settings

INTRODUCTION

Insight gleaned through first-person stories can and should be used to improve quality and safety in healthcare (1). In healthcare quality research, patient voices are often collected through purpose designed projects with a modest number of participants to assess patients' and their families' experiences within the healthcare system. This provides empirical evidence about the local context; however, the heterogeneity of methods used across projects makes learning difficult from a broader perspective. Nonetheless, there exist formalized collections of data pertaining to patient perspectives on healthcare in most countries, where systems have been established by national authorities and supervising organizations to address patients' and relatives' concerns. Such data from formal systems often benefit from being collected through standardized forms completed at the initiative of the patient or relatives. These forms are collected in great numbers by healthcare organizations, which, along with national authorities, may receive high volumes of patient complaints, and compensation claims. Information from these sources is collected at the patient's initiative, which often reflects an attempt to prevent an incident from happing again (2, 3). Such sources have been mentioned in the literature as essential indicators of problems in healthcare systems (4, 5), although challenges arise when attempting to use them for quality improvement (6). Patient complaints and reports of adverse events have been systematically collected for years, but, for the most part, they have not methodically been used as a sign of quality or for the purpose of healthcare improvement. However, these sources have great potential to complement the more technical measures of quality, such as process performance measures (e.g., early antiplatelet therapy initiated to manage stroke) and more traditional outcomes (e.g., mortality and length of stay). Overall, they may provide a more nuanced picture of quality and identify opportunities for improvement. To locate problems and trends in healthcare systems based on patient complaints, reports of adverse events, and claims of compensation, data must be aggregated, and analyzed in a systematic manner. Reader and colleagues conducted a systematic review of empirical research on patient complaints, with the overall aim of developing a taxonomy for guiding and standardizing the process of analyzing such complaints (7). This review was followed by the development of the "Healthcare Complaints Analysis Tool" (HCAT) (8), a standardized tool for systematically codifying and analyzing complaints to reliably assess healthcare problems and their severity. The HCAT has been applied in several countries (5, 9-14) and has been used to identify "blind spots" in healthcare systems (15). When aggregated, a statistical analysis of HCAT data may highlight problems and trends in healthcare. The HCAT taxonomy condenses data using a three-level hierarchy of "domains," "problem categories," and 36 sub-categories. Further, the taxonomy includes data on severity, stage of care, level of harm, the person making a complaint, the gender of the patient, and the staff groups to which the complaint refers. The first studies testing HCAT's reliability have already been published, while studies testing the reliability of sub-categories are still in progress (8). However, until now, reliability testing has only been performed in English settings, and it is

unknown whether HCAT would be useful outside the English language and in healthcare systems with different organizations, authorities, and supervising bodies.

We therefore scored and analyzed patient compensation claims aiming to test the reliability of the HCAT taxonomy and to clarify its potential for quality improvement in a Danish setting. Reliability coefficients are presented for the three domains, the seven problem categories, and the 36 sub-categories. See Figure 1 for an outline of the taxonomy.

METHODS AND MATERIAL

Data source and coding form

In Denmark, the compensation and disciplinary systems are separated. In our research, we only considered compensation claims handled by the Danish Patient Compensation Association (DPCA). In our reliability study, we included a random sample of 140 cases completed by the DPCA from 2007 to 2018. The healthcare that was the subject of each complaint must been provided at a Danish hospital within the field of acute medicine, according to DPCA classifications. According to Danish law (the Act on Complaints and Compensations 995/2018), a patient can receive compensation for health expenses, lost earnings, pain and suffering, permanent injury, loss of ability to work, and funeral expenses, if their injury could have been avoided by an experienced specialist acting differently, and/or if their complication is more rare and serious than the condition for which they have been treated. Compensation claims are managed in the DPCA through the obtaining of all relevant written information (including medical charts, radiographic material, and anesthetic charts), followed by the requisitioning of statements from specialized medical consultants. Finally, a decision is made by the DPCA. In our study, only patient claim letters, drafted by a patient, or relative, were used for analyses, thereby emphasizing patient perspectives on healthcare quality as set forth by the HCAT. Our sample included both accepted and rejected claims. Assessors (see below) reviewed claims letters using DPCA's electronic case management system at the DPCA's office facility in Odense, Denmark. Based on the HCAT manual, a web-based coding form was developed using Research Electronic Data Capture (REDCap). The REDCap form was designed to cover all areas addressed by the HCAT. Additionally, our web-based coding form allowed assessors to make written notes if claims did not fit into the predesigned problem categories. These written notes were intended to inform a future national adaption of the HCAT taxonomy, if necessary. As instructed in the HCAT manual, assessors read the full claim letter, and then completed the web-based form. To be as close as possible to the original HCAT form, the web-based form was in English. As a result, assessors were required to identify Danish keywords while reading the Danish letter and attribute them to English categories.

Patient and Public Involvement

The public and patients were not involved in the design, conception of this study or in the interpretation of the results. Final study results will be shared with stakeholders.

Assessors

Our team of assessors (KP, JK, CH, and SB) consisted of four academics, one student enrolled in a Master of Science (MSc) in Nursing, one PhD-educated general practitioner, one Master in Psychology, and one Master in Public Health Sciences.

The four assessors independently familiarized themselves with the HCAT. This included an introduction to HCAT manual (16) and undertaking an online course developed by the inventors of HCAT (8). In a joint session, HCAT was applied to ten consecutive compensation claims. The first three claims were reviewed, and HCAT was applied by the group as a whole. In the seven remaining cases, HCAT was applied individually, followed by a feedback-and-discussion session. Assessors were trained to adhere as closely as possible to the HCAT manual. Their coding should thus be empirically based, and free, as far as possible, from individual clinical judgments. Afterwards, the assessors independently coded 140 randomly selected healthcare compensation claims. To calculate intra-assessor reliability, one assessor scored all cases twice at intervals of six weeks between the first and second assessments. The claim order was randomized, and the assessors were blinded to each other's ratings.

Statistics

Linear regression was used to calculate the average number of problem categories per claim letter and the average time spent per claim letter. Regressions used robust standard error at case level to account for the heterogeneity in cases. Gwet's AC1 statistic was used to test intra- and inter-assessor reliability, both at coding the relevant category (0, 1) and when using the severity ratings (0, 1, 2, 3) (17). The severity ratings were only applied at the problem-category level and were analyzed using quadratic weights to assign large discrepancies more weight than small ones. Gwet's AC1 test was also applied to the reliability testing of stages of care and to the descriptive detail about each compensation case. The level of harm was coded on a scale from 1 (negligible) to 5 (catastrophic), and treated as a continuous variable; in this regard, intra-class correlation coefficients were used to test reliability. Cases were excluded when three or four assessors found a case to be inapplicable (e.g., from the lack of a patient claim letter or for complaints not pertaining to acute medicine).

Our interpretation of reliability follows the commonly used guideline: values 0.01 to 0.20 denote poor/slight agreement; 0.21 to 0.40, fair agreement; 0.41 to 0.60, moderate agreement; 0.61 to 0.80, substantial agreement; and 0.81 to 1.00, excellent agreement (18). In our statistical analyses, we used Stata, version 15 (StataCorp LP, College Station, Texas).

Ethical approval

The Danish Data Protection Agency and the DPCA approved our handling of the data (please see project approval 17/18411). According to Danish law, no approval from an ethics committee was required for this study (Act on Research Ethics Review of Health Research Projects (Act 1083, dated 15/09/2017; Para 14)).

Table 1 The distribution of the individual assessors' coding of the problem categories

	Assessor 1		Assessor 2	Assessor 3	Assessor 4
	Round A	Round B			
Quality	110	87	89	128	135
Safety	99	109	86	117	109
Environment	2	1	5	14	1
Institutional processes	7	11	25	26	32
Listening	8	13	19	18	31
Communication	3	6	11	5	13
Respect	4	2	2	1	4
Not applicable	13	7	18	15	12
Total	246	236	255	324	337

Results

Six cases were found to be "not applicable" by three or four of the assessors and were therefore excluded, leaving 134 cases for analysis. The distribution of the individual assessors' coding of the problem categories is shown in Table 1. On average, assessors applied 1.97 (95% confidence interval (CI) 1.85–2.09) categories and spent 4.63 (95% CI 4.19–5.09) minutes per claim letter. A steep learning curve was found regarding time spent per case assessment. Coding of the last 20 cases took on average one minute less than the coding of the first 20 cases. The results of the reliability analysis for the problem domain, the problem categories, the stages of care, the levels of harm, the person making the complaint, the gender of the patient, and the staff group are shown in Table 2. Gwet's AC1 test revealed that the HCAT was reliable in the identity problem domain, with excellent intra-assessor reliability, and substantial-to-excellent inter-assessor reliability. The ability of the HCAT to reliable identify problem categories was ranked from fair to excellent for both intra- and inter- assessor reliability. The "quality" category had the poorest intra-assessor reliability (0.55), while the "safety" category proved to have the lowest inter-assessor reliability (0.61). The reliability of the coding for "stage of care" was found to be substantial, or excellent, but for "operation and procedure," the coding showed the lowest levels of intra- and inter-reliability (0.62 and 0.74 respectively). The reliability levels of the coding for "gender," for the "complainer," and for "staff" complained about were excellent. Only when a claim was related to medical staff, or when the complainer was unspecified, was the inter-reliability found to be substantial, at 0.65 and 0.66 respectively. Both intra- and inter-assessor reliability were poor when coding

"level of harm" (0.06 and 0.29 respectively).

When coding the categories linked to the "severity" of the problem, intra-assessor reliability was ranked from fair to excellent and inter-assessor reliability from moderate to excellent. The reliability estimates for severity coding are shown in Table 3. The intra-assessor reliability for "quality" was fair (0.38), but the remaining six problem categories were found to have excellent intra-assessor reliability. The inter-assessor reliability for coding "quality" was found to be moderate (0.74), but was substantial or excellent for the remaining problem categories.

As shown in table 4, the intra-reliability was excellent for the majority of sub-categories used. Among the "quality" sub-categories, the "outcome and side effects" sub-category had significantly poorer reliability (-0.08) than the remaining sub-categories, for which the level of reliability was substantial or higher. Twenty-two of the 36 sub-categories were used in the intra-reliability testing, and 27 were utilized in the inter-reliability testing. Inter-reliability was excellent for the majority of the sub-categories. The sub-category "outcomes and side effects" had the poorest reliability, together with "examination & monitoring" (0.33 and 0.41 respectively).

Table 2 Intra-reliability and inter-reliability (n = 4) using 134 healthcare compensation claim letters from the DPCA

	Intra-reliability				Inter-reliability			
HCAT problem categories	Agreement	Gwet's AC	95% CI		Agreement	Gwet's AC	95% CI	
Clinical problems	0.95	0.94	0.89	0.99	0.90	0.88	0.84	0.93
Quality	0.73	0.55	0.40	0.70	0.74	0.64	0.56	0.73
Safety	0.84	0.76	0.65	0.87	0.75	0.61	0.51	0.70
Management problems	0.94	0.93	0.88	0.98	0.78	0.68	0.60	0.77
Environment	0.99	0.99	0.98	1.00	0.93	0.93	0.89	0.96
Institutional processes	0.94	0.93	0.88	0.98	0.80	0.72	0.64	0.81
Relationship problems	0.93	0.91	0.84	0.97	0.80	0.71	0.62	0.80
Listening	0.95	0.94	0.89	0.99	0.86	0.81	0.75	0.88
Communication	0.98	0.98	0.95	1.00	0.93	0.92	0.87	0.96
Respect and patients' rights	0.97	0.97	0.94	1.00	0.97	0.97	0.94	0.99
Stage of care								
Admissions	0.97	0.97	0.94	1.00	0.94	0.93	0.90	0.97
Examination and diagnosis	0.83	0.70	0.58	0.82	0.80	0.68	0.59	0.77
Care on ward	0.99	0.99	0.98	1.00	0.94	0.93	0.89	0.97
Operation or procedure	0.78	0.62	0.48	0.76	0.74	0.55	0.45	0.66
Discharge/transfers	0.99	0.98	0.96	1.00	0.89	0.88	0.83	0.93
Other stage	0.96	0.96	0.92	1.00	0.88	0.86	0.81	0.92
Complainer								
Family member	0.98	0.97	0.94	1.00	0.95	0.94	0.90	0.98
Patient	0.90	0.85	0.77	0.94	0.85	0.77	0.70	0.85
Complainer unspecified	0.92	0.90	0.84	0.96	0.84	0.80	0.74	0.87
Patient gender								
Female	0.96	0.93	0.86	0.99	0.94	0.87	0.81	0.93
Male	0.96	0.91	0.84	0.98	0.93	0.85	0.79	0.92
Gender unspecified	0.96	0.96	0.92	1.00	0.94	0.93	0.90	0.97
Complained about								

Administrative staff	0.96	0.96	0.92	1.00	0.99	0.99	0.98	1.00
Medical staff	0.88	0.81	0.71	0.90	0.79	0.65	0.56	0.75
Nursing staff	0.99	0.99	0.98	1.00	0.97	0.96	0.94	0.99
Staff unspecified	0.88	0.80	0.70	0.90	0.80	0.66	0.57	0.76
Harm level†		0.06	-0.19	0.28		0.29	0.08	0.47

HCAT: Healthcare Complaints Analysis Tool; CI: Confidence interval; † Interclass correlations coefficient

Table 3 Case severity: Domain and problem intra- and inter-assessor reliability (n = 4) using 134 healthcare claim letters

	Intra-reliability				Inter-reliability			
HCAT problem categories	Agreement	Gwet's AC1	95% CI		Agreement	Gwet's AC	95% CI	
Clinical problems	0.94	0.86	0.80	0.92	0.87	0.75	0.68	0.82
Quality	0.78	0.38	0.21	0.55	0.74	0.48	0.37	0.59
Safety	0.92	0.83	0.77	0.90	0.86	0.71	0.63	0.78
Management problems	0.99	0.99	0.99	1.00	0.98	0.97	0.96	0.98
Environment	1.00	1.00	1.00	1.00	0.96	0.96	0.93	0.98
Institutional processes	0.96	0.96	0.93	0.99	0.87	0.82	0.77	0.88
Relationship problems	0.99	0.99	0.98	1.00	0.98	0.97	0.96	0.98
Listening	0.97	0.97	0.95	0.99	0.92	0.89	0.85	0.93
Communication	0.99	0.99	0.98	1.00	0.97	0.96	0.94	0.98
Respect and patients' rights	1.00	1.00	0.99	1.00	0.99	0.99	0.98	1.00

HCAT: Healthcare Complaints Analysis Tool; CI: Confidence interval

Table 4 Sub-category intra- and inter-assessor reliability (n = 4) using 134 healthcare claim letters

	Intra-reliability				Inter-reliability			
	Agreement	Gwet's AC	95% CI		Agreement	Gwet's AC	95% CI	
Quality								
Neglect—hygiene and personal care		No ratings				No ratings		
Neglect—nourishment and hydration		No ratings				No ratings		
Neglect—general	0.75	0.66	0.54	0.78	0.84	0.81	0.75	0.87
Rough handling and discomfort	0.99	0.99	0.98	1.00	0.99	0.99	0.97	1.00
Examination and monitoring	0.85	0.74	0.63	0.85	0.69	0.41	0.30	0.51
Making and following care plans	0.99	0.98	0.96	1.00	0.98	0.98	0.96	1.00
Outcomes and side effects	0.45	-0.08	-0.26	0.09	0.59	0.33	0.22	0.43
Other		No ratings			1.00	1.00	0.99	1.00
Safety								
Error—diagnosis	0.83	0.67	0.54	0.80	0.76	0.53	0.44	0.63
Error—medication	0.99	0.99	0.98	1.00	0.99	0.99	0.97	1.00
Error—general	0.85	0.82	0.73	0.90	0.74	0.63	0.54	0.72
Failure to respond		No ratings				No ratings		
Clinician skills	0.88	0.84	0.76	0.92	0.86	0.83	0.77	0.89
Teamwork	0.98	0.98	0.95	1.00	0.98	0.98	0.95	1.00
Other		No ratings			0.97	0.97	0.95	0.99
Environment								
Accommodation	0.99	0.99	0.98	1.00	1.00	1.00	0.99	1.00
Preparedness		No ratings				No ratings		
Ward cleanliness		No ratings				No ratings		
Equipment	1.00	1.00	1.00	1.00	0.98	0.98	0.97	1.00
Staffing		No ratings			0.95	0.95	0.92	0.98
Security		No ratings				No ratings		
Other		No ratings			1.00	1.00	0.99	1.00
Institutional Processes								
Delay—access	0.96	0.96	0.93	1.00	0.90	0.88	0.84	0.93

Delay—procedure	0.94	0.94	0.89	0.98	0.87	0.85	0.80	0.90
Delay—general	0.99	0.99	0.98	1.00	0.98	0.98	0.96	1.00
Bureaucracy	0.99	0.98	0.96	1.00	0.96	0.95	0.93	0.98
Visiting		No ratings			0.99	0.99	0.98	1.00
Documentation	0.99	0.98	0.96	1.00	0.99	0.99	0.98	1.00
Other		No ratings			0.99	0.99	0.98	1.00
Listening								
Ignoring patients	0.95	0.94	0.90	0.99	0.90	0.89	0.85	0.94
Dismissing patients	0.92	0.91	0.85	0.97	0.88	0.86	0.80	0.92
Token listening		No ratings			0.99	0.99	0.98	1.00
Other		No ratings				No ratings		
Communication								
Delayed communication	1.00	1.00	1.00	1.00	0.99	0.99	0.98	1.00
Incorrect communication		No ratings			0.95	0.95	0.92	0.98
Absent communication	0.98	0.98	0.95	1.00	0.96	0.95	0.92	0.98
Other		No ratings				No ratings		
Respect and patient rights								
Disrespect	0.97	0.97	0.94	1.00	0.97	0.97	0.95	0.99
Confidentiality		No ratings				No ratings		
Rights		No ratings			0.99	0.99	0.98	1.00
Consent		No ratings				No ratings		
Privacy		No ratings				No ratings		
Other		No ratings				No ratings		

Discussion

Even though the HCAT was developed, tested, and refined in an English setting, it is based upon a systematic review of the international literature, which we deem to be the most comprehensive to date (7). This was the main reason for our interest in the HCAT. In this study, we estimated both the intra- and inter-reliability of the HCAT to gain clarity regarding its usefulness in settings outside the UK. The four assessors achieved an overall satisfactory level of reliability when using the HCAT on patient claims for compensation. As expected, intra-assessor reliability was superior to its inter-assessor counterpart in most problem categories. The HCAT was highly reliable when identifying problem categories, but its reliability was lower when coding the *severity* of problems. However, the reliability of each sub-category was satisfactory in most cases. Likewise, the HCAT showed satisfactory reliability when information about the complainer and the staff who were the subject of the complaint was coded. It was very difficult to code the level of harm incurred, which resulted in low reliability scores. Overall, the HCAT seemed relatively time-effective to use; its application took, on average, less than five minutes per compensation claim, and assessors quickly became familiar with the tool.

Interpretation of findings and comparison with existing literature

Our finding that the "quality and safety" categories had the lowest reliability corresponds with the findings of Gillespie and Reader who developed the HCAT (8). In their study, substantial reliability was achieved in this domain, still leaving room for improvement. Across categories, our study achieved reliability comparable to that found by Gillespie and Reader, and in some areas, our reliability estimates were even

greater. Other research has reported reliability coefficients for the HCAT and found reliability at the problem-category level ranging from 0.75 to 0.98 (11) and overall HCAT reliability coefficients ranging from 0.81 (9) to 0.92 (5). No other study has reported on reliability at the sub-category level, and more extensive and robust studies are needed to establish the reliability of the HCAT at this level. In our analyses, low intra-assessor reliability coefficients tended to appear together with low inter-reliability, indicating a possible problem in the definition or the pre-training of our assessors in regard to these categories specifically.

We experienced skewed distribution among the seven problem categories (see Table 1). The study sample was randomly selected from patient compensation claims, and we expect that the prevalence of problems in the study reflect the true prevalence of problems in patients' claims for compensation across the field of acute medicine. While most agreement statistics are only valid with a prevalence of around 50%, Gwet's AC1 statistic is valid with both high and low prevalences (17). The HCAT tool focuses on the identification of macro trends, which could be difficult to analyze if up to 95% of all claims fall into the quality problem category. This emphasizes the need for reliable sub-categories. Our findings at the sub-category level generally pointed toward satisfactory reliability, but with significant fluctuations in some sub-categories.

Strengths and Limitations

The HCAT is developed for analyzing and coding "complaints." Despite this, we included patient compensation claims exclusively. Therefore, we anticipated that the range of categories could be insufficient. However, such inadequacy was detected only at the sub-category level.

In our coding of claim letters, we followed the HCAT manual as rigorously as possible. Regarding the pre-training of the assessors, we followed the tutorial process described by Gillespie and Readers reliability study (8). Retrospectively, reliability measures could probably have been improved by spending more time becoming familiar with—and agreeing on—the classes of Danish words that indicate specific problem categories. Likewise, reliability estimates may have benefited from the improved preparation of assessors on how to rate the level of severity. Finally, it remains unclear, whether a complete translation of the HCAT into Danish might have resulted in even higher reliability—as this was our first study using the HCAT we aimed to test the reliability of the original version of the tool. A more balanced material would likely have strengthened our study, delivering reliability scores for all sub-categories. However, it is uncertain whether all sub-categories can be found in compensation cases.

Conclusion

By way of conclusion, in our study of the application the HCAT in Danish healthcare, and within a complaint system different from the English one, we found it to perform successfully. The HCAT proves to be a reliable tool for distinguishing problem types in patient compensation claim letters, thereby potentially making it appealing for future use in quality research and quality improvement.

Funding

This work was supported by Region of Southern Denmark [17/33854].

Conflicts of interests

None

Author contributions

SB and SBB devised the project with advice from CHH and KLM. SBB was principal investigator and together with SB responsible for study design, protocol, data collection, and analysis with input from CH, KLM, JHK, and KPJ. SB, CHH, JHK, and KPJ scored compensation cases. All authors contributed with interpretation of results and significant revisions to drafts of the manuscript, and have read and approved the final manuscript.

Reference

1. Vincent CA, Coulter A. Patient safety: what about the patient? Quality & safety in health care. 2002;11(1):76-80.
2. Friele RD, Sluijs EM. Patient expectations of fair complaint handling in hospitals: empirical data. BMC Health Serv Res. 2006;6:106.
3. Bouwman R, Bomhoff M, Robben P, Friele R. Patients' perspectives on the role of their complaints in the regulatory process. Health Expect. 2016;19(2):483-96.
4. Pichert JW, Hickson G, Moore I. Using Patient Complaints to Promote Patient Safety. In: Henriksen K, Battles JB, Keyes MA, Grady ML, editors. Advances in Patient Safety: New Directions and Alternative Approaches (Vol 2: Culture and Redesign). Advances in Patient Safety. Rockville (MD)2008.
5. Harrison R, Walton M, Healy J, Smith-Merry J, Hobbs C. Patient complaints about hospital services: applying a complaint taxonomy to analyse and respond to complaints. Int J Qual Health Care. 2016;28(2):240-5.
6. de Vos MS, Hamming JF, Marang-van de Mheen PJ. The problem with using patient complaints for improvement. BMJ Qual Saf. 2018;27(9):758-62.
7. Reader TW, Gillespie A, Roberts J. Patient complaints in healthcare systems: a systematic review and coding taxonomy. BMJ Qual Saf. 2014;23(8):678-89.
8. Gillespie A, Reader TW. The Healthcare Complaints Analysis Tool: development and reliability testing of a method for service monitoring and organisational learning. BMJ Qual Saf. 2016;25(12):937-46.
9. Mattarozzi K, Sfrisi F, Caniglia F, De Palma A, Martoni M. What patients' complaints and praise tell the health practitioner: implications for health care quality. A qualitative research study. Int J Qual Health Care. 2017;29(1):83-9.
10. Bouwman R, Bomhoff M, Robben P, Friele R. Classifying Patients' Complaints for Regulatory Purposes: A Pilot Study. J Patient Saf. 2016.
11. Mack JW, Jacobson J, Frank D, Cronin AM, Horvath K, Allen V, et al. Evaluation of Patient and Family Outpatient Complaints as a Strategy to Prioritize Efforts to Improve Cancer Care Delivery. Joint Commission journal on quality and patient safety / Joint Commission Resources. 2017;43(10):498-507.
12. Jerng JS, Huang SF, Yu HY, Chan YC, Liang HJ, Liang HW, et al. Comparison of complaints to the intensive care units and those to the general wards: an analysis using the Healthcare Complaint Analysis Tool in an academic medical center in Taiwan. Crit Care. 2018;22(1):335.
13. de Vos MS, Hamming JF, Chua-Hendriks JJC, Marang-van de Mheen PJ. Connecting perspectives on quality and safety: patient-level linkage of incident, adverse event and complaint data. BMJ Qual Saf. 2019;28(3):180-9.
14. Wallace E, Cronin S, Murphy N, Cheraghi-Sohi S, MacSweeney K, Bates M, et al. Characterising patient complaints in out-of-hours general practice: a retrospective cohort study in Ireland. The British journal of general practice : the journal of the Royal College of General Practitioners. 2018;68(677):e860-e8.

15. Gillespie A, Reader TW. Patient-Centered Insights: Using Health Care Complaints to Reveal Hot Spots and Blind Spots in Quality and Safety. Milbank Q. 2018;96(3):530-67.
16. Gillespie A, Reader TW. Healthcare Complaints Analysis Tool London, United Kingdom: The London School of Economics and political science; 2015 [1-20]. Available from: https://qualitysafety.bmj.com/content/suppl/2016/01/05/bmjqs-2015-004596.DC1/bmjqs-2015-004596supp_new.pdf.
17. Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement. Br J Math Stat Psychol. 2008;61(Pt 1):29-48.
18. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159-74.

Domain	Problem category	Sub-category
Clinical problems Issues relating to the quality and safety of the clinical and nursing care provided by healthcare staff	**Quality:** Clinical standards of healthcare staff behavior	Neglect—hygiene and personal care; neglect—nourishment and hydration; neglect—general; rough handling and discomfort; examination and monitoring; making and following care plans; outcomes and side effects
	Safety: Errors, incidents, and staff competencies	Error—diagnosis; error—medication; error—general; failure to respond; clinician skills; teamwork
Management problems Issues relating to the environment and the organizations within which healthcare is provided	**Environment:** Problems with the facilities, services, clinical equipment, and staffing levels	Accommodation; preparedness; ward cleanliness; equipment; staffing; security
	Institutional processes: Problems with bureaucracy, waiting times, and accessing care	Delay—access; delay—procedure; delay—general; bureaucracy; visiting; documentation
Relationship problems Issues relating to the behavior of any specific member(s) of staff toward the patient or their family/friends	**Listening:** Information from patients disregarded or not acknowledged by healthcare staff	Ignoring patients; dismissing patients; token listening
	Communication: Absent or incorrect communication from healthcare staff to patients	Delayed communication; incorrect communication; absent communication
	Respect and patient rights: Disrespect or violation of patient rights by staff	Disrespect; confidentiality; rights; consent; privacy

Figure 1 Domains, problem categories, and sub-categories of the HCAT taxonomy (cf. Gillespie A. and Reader TW (8))

65x57mm (300 x 300 DPI)

BMJ Open

The Healthcare Complaints Analysis Tool: Reliability testing on a sample of Danish patient compensation claims

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2019-033638.R1
Article Type:	Original research
Date Submitted by the Author:	14-Oct-2019
Complete List of Authors:	Bogh, Soren; Odense University Hospital, OPEN Kerring, Jonas; Region of Southern Denmark, Department of Psychiatry Middelfart Middelfart Jakobsen, Katrine; Region Zealand, Department of Health Strategic Planning Hilsøe, Camilla; Region of Southern Denmark, Department for Quality and Research. Mikkelsen, Kim; Patient insurance association, denmark, Patient injury prevention and learning Birkeland, Søren; Odense University Hospital, OPEN
Primary Subject Heading:	Health services research
Secondary Subject Heading:	Emergency medicine, Health policy
Keywords:	Reliability, Healthcare complaints, Patient safety, Health service research

Title

The Healthcare Complaints Analysis Tool: Reliability testing on a sample of Danish patient compensation claims

Authors

Søren Bie Bogh, Jonas Harder Kerring, Katrine Prisak Jakobsen, Camilla Hagemann Hilsøe, Kim Lyngby Mikkelsen, Søren Fryd Birkeland

Affiliations

Søren Bie Bogh, PhD., Researcher, soren.bie.bogh@rsyd.dk

OPEN, Odense Patient data Explorative Network, University of Southern Denmark and Odense University Hospital, DK-5000 Odense C, Denmark

Jonas Harder Kerring, Cand.Psych., jonas.harder.kerring@rsyd.dk. Department of Psychiatry Middelfart.

Katrine Prisak Jakobsen, cand.scient.san.publ., Consultant, katjak@regionsjaelland.dk

Region Zealand, DK-4180 Sorø, Denmark

Camilla Hagemann Hilsøe, Cand. Cur., Camilla.hilsoe@rsyd.dk. Department for Quality and Research. Region of Southern Denmark.

Kim Lyngby Mikkelsen, Kim.Lyngby.Mikkelsen@Patienterstatningen.dk The Patient Compensation Association, Denmark. Kalvebod Brygge 45, 1560 Copenhagen V.

Søren Fryd Birkeland, LLM PhD., Soren.Birkeland@rsyd.dk associate professor, OPEN, Odense Patient data Explorative Network, University of Southern Denmark and Odense University Hospital, DK-5000 Odense C, Denmark

Corresponding details:

Søren Bie Bogh, PhD., soren.bie.bogh@rsyd.dk

Running title: Categorizing claim letters

Word count: Abstract: 240 words, manuscript: 2890 words

Keywords**:** Reliability**,** Healthcare complaints, Patient safety**,** Acute care, Health service research

Abstract

Objective

The study aim was to test the intra- and inter-assessor reliability of the Healthcare Complaints Analysis Tool (HCAT) for categorizing the information in the claim letters in a sample of Danish patient compensation claims.

Design, setting and participants

We used a random sample of 140 compensation cases completed by the Danish Patient Compensation Association that were filed in the field of acute medicine at Danish hospitals from 2007 to 2018. Four assessors were trained in using the HCAT manual before assessing the claim letters independently.

Main outcome measures

Intra- and inter-assessor reliability was tested at domain, problem category, and sub-category levels of the HCAT. We also investigated the reliability of ratings on the level of harm and of the descriptive details contained in the claim letters.

Results

The HCAT was reliable for identifying problem categories, with reliability scores ranging from 0.55 to 0.99. Reliability was lower when coding the "severity" of the problem. Inter-assessor reliability was generally lower than intra-assessor reliability. The categories of "quality" and "safety" were the least reliable of the seven HCAT problem categories. Reliability at the sub-category level was generally satisfactory, with only a few sub-categories having poor reliability. Reliability was at least moderate when coding the stage of care, the complainant, and the staff group involved. However, the coding of "level of harm" was found to be unreliable (intra-reliability 0.06; inter-reliability 0.29).

Conclusion

Overall, HCAT was found to be a reliable tool for categorizing problem types in patient compensation claims.

Article Summary - Strengths and limitations of this study

- The present study focuses on injury compensation claims related to emergency hospital care
- A key strength of the study is the testing of the whole HCAT instrument (the domains, categories, and sub-categories) in a large sample of complaint cases outside the setting where HCAT was developed
- Multiple, trained raters showed high inter-assessor reliability
- Due to skewed coverage of the HCAT domains and sub-category levels, our study cannot stand alone and must be followed by further studies in different healthcare settings

INTRODUCTION

Knowledge gained from first-person patient stories can and should be used to improve quality and safety in healthcare (1). In quality improvement research, patients' perspectives are often collected through custom-made projects with a modest number of participants to explore the experiences of patients and their families within the healthcare system. Such studies provide empirical evidence about the local context, but the heterogeneity of methods used makes learning difficult from a broader perspective.

Most countries have formalized collection of data on patient perspectives, where systems have been established by national authorities and supervising organizations to address patients' and relatives' concerns. These data often benefit from the use of standardized forms completed at the initiative of the patient or relatives. Healthcare organizations and national authorities may receive high volumes of patient complaints and compensation claims, and a main goal is to prevent an incident from happening again (2, 3). Such data sources are essential indicators of problems in healthcare systems (4, 5), but challenges arise when attempting to use them for quality improvement (6).

Although patient complaints and reports of adverse events have been systematically collected for many years, they have typically not been systematically used to assess or improve the quality of healthcare. These sources have great potential to complement other measures of quality such as process performance (e.g. initiation of antiplatelet therapy in the management of stroke) and outcomes (e.g. mortality and length of stay). Patient complaints and reports of adverse events may provide a more nuanced picture of quality and could help to identify potential areas for improvement.

If patient complaints, reports of adverse events, and compensation claims are to be used for quality improvement, these data must be aggregated in some way and then analyzed in a systematic manner. Reader and colleagues conducted a systematic review of empirical research on patient complaints, aiming to develop a taxonomy for guiding and standardizing the analysis of such complaints (7). This review was followed by the development of the Healthcare Complaints Analysis Tool (HCAT) (8), a standardized tool for systematically codifying and analyzing complaints to reliably assess healthcare problems and their severity. 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. The HCAT has been applied in several countries (5, 9-14) and has been used to identify "blind spots" in healthcare systems (15). The HCAT taxonomy condenses data using a three-level hierarchy of "domains," "problem categories," and 36 sub-categories (Figure 1). Further, the taxonomy includes data on severity, stage of care, level of harm, the person making a complaint, the gender of the patient, and the staff groups to which the complaint refers. The first study on the reliability of HCAT have already been published, while studies testing the reliability of sub-categories are in

progress (8). 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.

The aim of the current study was to test the reliability of the HCAT taxonomy by scoring and analyzing patient compensation claims and to clarify the potential of HCAT for quality improvement in a Danish healthcare setting. Reliability coefficients are presented for the three domains, seven problem categories, and 36 sub-categories of the HCAT.

METHODS AND MATERIAL

Data source and coding form

The compensation and disciplinary systems are separated in the Danish system, and this study only includes compensation claims handled by the Danish Patient Compensation Association (DPCA). 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 (8, 15)

According to Danish law (the Act on Complaints and Compensations 995/2018), a patient can receive compensation for health expenses, lost earnings, pain and suffering, permanent injury, loss of ability to work, and funeral expenses if their injury could have been avoided by an experienced specialist acting differently, and/or if a complication was rarer and more serious than expected for the condition treated. Compensation claims are managed in the DPCA by obtaining all relevant written information (including medical charts, radiographic material, and anesthetic charts) and requisitioning statements from medical and/or surgical specialists. The DPCA then decides whether or not to award compensation to the patient.

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 (16) 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. We included only patient claim letters that were drafted by the patient or a relative, thereby emphasizing patients' perspectives on the quality of healthcare. Our sample included both accepted and rejected claims.

Four of the authors acted as assessors (see below) and reviewed the claims letters using DPCA's electronic case management system at the DPCA office in Odense, Denmark. Based on the HCAT manual, a web-based coding form was developed using Research Electronic Data Capture (REDCap). This was designed to cover all areas addressed by the HCAT and also allowed assessors to note cases where the claims did not fit

into the predesigned problem categories. These notes were intended to inform a possible national adoption of the HCAT taxonomy. As instructed in the HCAT manual, assessors read the full claim letter and then completed the web-based form. To be as close as possible to the original HCAT form, the web-based form was in English. As a result, assessors were required to identify Danish keywords while reading the Danish letter and attribute them to English categories.

Patient and Public Involvement

The public and patients were not involved in the conception or design of this study, nor in the interpretation of the results. Final study results will be shared with stakeholders.

Assessors

Our team of assessors (KP, JK, CH, and SB) consisted of four academics: a student enrolled in a Master of Science (MSc) in Nursing, a PhD-educated general practitioner, a researcher with a Master in Psychology, and a researcher with a Master in Public Health Sciences. The assessors were chosen with the expectation that their qualifications would represent potential future users of the HCAT for quality improvements.

The four assessors independently familiarized themselves with the HCAT. This included an introduction to the HCAT manual (17) and an online course developed by the inventors of HCAT (8). In a joint session, HCAT was applied to ten consecutive compensation claims. The first three claims were reviewed and analyzed using the HCAT by the group as a whole. For the seven remaining cases, HCAT was applied individually, followed by feedback and discussion within the group. Assessors were trained to adhere as closely as possible to the HCAT manual. Their coding should thus be empirically-based and as far as possible free from individual clinical judgments.

After this training, the assessors independently coded the 140 healthcare compensation claims selected for study. 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.

Statistics

Linear regression was used to calculate the average number of problem categories per claim letter and the average time spent per claim letter. Regressions used robust standard error at case level to account for the heterogeneity in cases. Gwet's AC1 statistic was used to test intra- and inter-assessor reliability, for both coding the relevant category (0, 1) and using the severity ratings (0, 1, 2, 3) (18). The severity ratings were only applied at the problem-category level and were analyzed using quadratic weights to assign large discrepancies more weight than small ones. Gwet's AC1 test was also applied to the reliability testing of

stages of care and to the descriptive detail about each compensation case. 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. Cases were excluded when three or four assessors found a case to be inapplicable (e.g. absence of a patient claim letter or complaints not pertaining to acute medicine).

Our interpretation of reliability followed the commonly used guideline: values 0.01 to 0.20 denote poor/slight agreement; 0.21 to 0.40, fair agreement; 0.41 to 0.60, moderate agreement; 0.61 to 0.80, substantial agreement; and 0.81 to 1.00, excellent agreement (19). We used Stata, version 15 (StataCorp LP, College Station, Texas) for the statistical analyses.

Ethical approval

The Danish Data Protection Agency and the DPCA approved our handling of the data (project approval 17/18411). According to Danish law, no approval from an ethics committee was required for this study (Act on Research Ethics Review of Health Research Projects (Act 1083, dated 15/09/2017; Para 14)).

Table 1 The distribution of the individual assessors' coding of the problem categories

	Assessor 1		Assessor 2	Assessor 3	Assessor 4
	Round A	Round B			
Quality	110	87	89	128	135
Safety	99	109	86	117	109
Environment	2	1	5	14	1
Institutional processes	7	11	25	26	32
Listening	8	13	19	18	31
Communication	3	6	11	5	13
Respect	4	2	2	1	4
Not applicable	13	7	18	15	12
Total	246	236	255	324	337

Results

Six cases were found to be "not applicable" by three or four of the assessors and were excluded, leaving 134 cases for analysis. Table 1 shows the distribution of the individual assessors' coding of the problem categories. On average, assessors applied 1.97 (95% confidence interval (CI) 1.85–2.09) HCAT categories and spent 4.63 (95% CI 4.19–5.09) minutes per claim letter. We observed a steep learning curve regarding time spent per case, where coding of the last 20 cases took on average one minute less than coding of the first 20 cases.

Table 2 shows results of the reliability analysis for the three HCAT domains (Clinical, Management, and Relationship problems) and the seven problem categories under these, and for information on stage of care, the person making the complaint, the gender of the patient, the staff group involved, and the level of harm. Gwet's AC1 test revealed that the HCAT was reliable in identifying the problem domain, with excellent intra-assessor reliability and substantial to excellent inter-assessor reliability. The ability of the HCAT to reliably identify problem categories was fair to excellent for both intra- and inter- assessor reliability. The category of "quality" (Clinical standards of healthcare and behavior) had the poorest intra-assessor reliability (0.55), while the category of "safety" (Errors, incidents, and staff competencies) had the lowest inter-assessor reliability (0.61). The reliability for coding the overall "stage of care" category was excellent, but the coding of "operation and procedure" showed the lowest levels of intra- and inter-reliability (0.62 and 0.74 respectively). The reliability for coding of complainant, patient gender, and involved staff group was excellent, although reliability was only substantial when a claim was related to medical staff (0.65) or the complainant was unspecified (0.66). Both intra- and inter-assessor reliability were poor when coding "level of harm" (0.4 and 0.19 respectively).

The reliability estimates for severity coding (Table 3) showed that for the three domains, intra-assessor reliability was excellent and inter-assessor reliability was substantial to excellent. While six problem categories had excellent intra-assessor reliability and substantial to excellent inter-assessor reliability, the problem category "quality" had only fair intra-assessor reliability (0.38) and moderate inter-assessor reliability (0.74).

As shown in Table 4, inter-assessor reliability and intra-assessor reliability were substantial to excellent for most of the 36 HCAT sub-categories. The sub-category "outcome and side effects" (under the "quality" category) had significantly poorer intra-assessor reliability (-0.08) than the other sub-categories as well as the poorest inter-assessor reliability (0.33). The sub-category "examination & monitoring" (under the "quality category) also had poor inter-assessor reliability (0.41). Twenty-two of the 36 sub-categories were used in the intra-assessor reliability testing, and 27 were utilized in the inter-assessor reliability testing.

Table 2 Intra-reliability and inter-reliability (n = 4) using 134 healthcare compensation claim letters from the DPCA

	Intra-reliability				Inter-reliability			
HCAT problem categories	Agreement	Gwet's AC	95% CI		Agreement	Gwet's AC	95% CI	
Clinical problems	0.95	0.94	0.89	0.99	0.90	0.88	0.84	0.93
Quality	0.73	0.55	0.40	0.70	0.74	0.64	0.56	0.73
Safety	0.84	0.76	0.65	0.87	0.75	0.61	0.51	0.70
Management problems	0.94	0.93	0.88	0.98	0.78	0.68	0.60	0.77
Environment	0.99	0.99	0.98	1.00	0.93	0.93	0.89	0.96
Institutional processes	0.94	0.93	0.88	0.98	0.80	0.72	0.64	0.81

Relationship problems	0.93	0.91	0.84	0.97	0.80	0.71	0.62	0.80
Listening	0.95	0.94	0.89	0.99	0.86	0.81	0.75	0.88
Communication	0.98	0.98	0.95	1.00	0.93	0.92	0.87	0.96
Respect and patients' rights	0.97	0.97	0.94	1.00	0.97	0.97	0.94	0.99
Stage of care								
Admissions	0.97	0.97	0.94	1.00	0.94	0.93	0.90	0.97
Examination and diagnosis	0.83	0.70	0.58	0.82	0.80	0.68	0.59	0.77
Care on ward	0.99	0.99	0.98	1.00	0.94	0.93	0.89	0.97
Operation or procedure	0.78	0.62	0.48	0.76	0.74	0.55	0.45	0.66
Discharge/transfers	0.99	0.98	0.96	1.00	0.89	0.88	0.83	0.93
Other stage	0.96	0.96	0.92	1.00	0.88	0.86	0.81	0.92
Complainant								
Family member	0.98	0.97	0.94	1.00	0.95	0.94	0.90	0.98
Patient	0.90	0.85	0.77	0.94	0.85	0.77	0.70	0.85
Complainant unspecified	0.92	0.90	0.84	0.96	0.84	0.80	0.74	0.87
Patient gender								
Female	0.96	0.93	0.86	0.99	0.94	0.87	0.81	0.93
Male	0.96	0.91	0.84	0.98	0.93	0.85	0.79	0.92
Gender unspecified	0.96	0.96	0.92	1.00	0.94	0.93	0.90	0.97
Complained about								
Administrative staff	0.96	0.96	0.92	1.00	0.99	0.99	0.98	1.00
Medical staff	0.88	0.81	0.71	0.90	0.79	0.65	0.56	0.75
Nursing staff	0.99	0.99	0.98	1.00	0.97	0.96	0.94	0.99
Staff unspecified	0.88	0.80	0.70	0.90	0.80	0.66	0.57	0.76
Harm level†		0.40	0.01	0.58		0.19	0.09	0.29

HCAT: Healthcare Complaints Analysis Tool; CI: Confidence interval; † Intra-class correlations coefficient

Table 3 Case severity: Domain and problem intra- and inter-assessor reliability (n = 4) using 134 healthcare claim letters

	Intra-reliability				Inter-reliability			
HCAT problem categories	Agreement	Gwet's AC1	95% CI		Agreement	Gwet's AC	95% CI	
Clinical problems	0.94	0.86	0.80	0.92	0.87	0.75	0.68	0.82
Quality	0.78	0.38	0.21	0.55	0.74	0.48	0.37	0.59
Safety	0.92	0.83	0.77	0.90	0.86	0.71	0.63	0.78
Management problems	0.99	0.99	0.99	1.00	0.98	0.97	0.96	0.98
Environment	1.00	1.00	1.00	1.00	0.96	0.96	0.93	0.98
Institutional processes	0.96	0.96	0.93	0.99	0.87	0.82	0.77	0.88
Relationship problems	0.99	0.99	0.98	1.00	0.98	0.97	0.96	0.98
Listening	0.97	0.97	0.95	0.99	0.92	0.89	0.85	0.93
Communication	0.99	0.99	0.98	1.00	0.97	0.96	0.94	0.98
Respect and patients' rights	1.00	1.00	0.99	1.00	0.99	0.99	0.98	1.00

HCAT: Healthcare Complaints Analysis Tool; CI: Confidence interval

Table 4 Sub-category intra- and inter-assessor reliability (n = 4) using 134 healthcare claim letters

	Intra-reliability				Inter-reliability			
	Agreement	Gwet's AC	95% CI		Agreement	Gwet's AC	95% CI	
Quality								
Neglect—hygiene and personal care		No ratings				No ratings		
Neglect—nourishment and hydration		No ratings				No ratings		
Neglect—general	0.75	0.66	0.54	0.78	0.84	0.81	0.75	0.87
Rough handling and discomfort	0.99	0.99	0.98	1.00	0.99	0.99	0.97	1.00

Examination and monitoring	0.85	0.74	0.63	0.85	0.69	0.41	0.30	0.51
Making and following care plans	0.99	0.98	0.96	1.00	0.98	0.98	0.96	1.00
Outcomes and side effects	0.45	-0.08	-0.26	0.09	0.59	0.33	0.22	0.43
Other		No ratings			1.00	1.00	0.99	1.00
Safety								
Error—diagnosis	0.83	0.67	0.54	0.80	0.76	0.53	0.44	0.63
Error—medication	0.99	0.99	0.98	1.00	0.99	0.99	0.97	1.00
Error—general	0.85	0.82	0.73	0.90	0.74	0.63	0.54	0.72
Failure to respond		No ratings				No ratings		
Clinician skills	0.88	0.84	0.76	0.92	0.86	0.83	0.77	0.89
Teamwork	0.98	0.98	0.95	1.00	0.98	0.98	0.95	1.00
Other		No ratings			0.97	0.97	0.95	0.99
Environment								
Accommodation	0.99	0.99	0.98	1.00	1.00	1.00	0.99	1.00
Preparedness		No ratings				No ratings		
Ward cleanliness		No ratings				No ratings		
Equipment	1.00	1.00	1.00	1.00	0.98	0.98	0.97	1.00
Staffing		No ratings			0.95	0.95	0.92	0.98
Security		No ratings				No ratings		
Other		No ratings			1.00	1.00	0.99	1.00
Institutional Processes								
Delay—access	0.96	0.96	0.93	1.00	0.90	0.88	0.84	0.93
Delay—procedure	0.94	0.94	0.89	0.98	0.87	0.85	0.80	0.90
Delay—general	0.99	0.99	0.98	1.00	0.98	0.98	0.96	1.00
Bureaucracy	0.99	0.98	0.96	1.00	0.96	0.95	0.93	0.98
Visiting		No ratings			0.99	0.99	0.98	1.00
Documentation	0.99	0.98	0.96	1.00	0.99	0.99	0.98	1.00
Other		No ratings			0.99	0.99	0.98	1.00
Listening								
Ignoring patients	0.95	0.94	0.90	0.99	0.90	0.89	0.85	0.94
Dismissing patients	0.92	0.91	0.85	0.97	0.88	0.86	0.80	0.92
Token listening		No ratings			0.99	0.99	0.98	1.00
Other		No ratings				No ratings		
Communication								
Delayed communication	1.00	1.00	1.00	1.00	0.99	0.99	0.98	1.00
Incorrect communication		No ratings			0.95	0.95	0.92	0.98
Absent communication	0.98	0.98	0.95	1.00	0.96	0.95	0.92	0.98
Other		No ratings				No ratings		
Respect and patient rights								
Disrespect	0.97	0.97	0.94	1.00	0.97	0.97	0.95	0.99
Confidentiality		No ratings				No ratings		
Rights		No ratings			0.99	0.99	0.98	1.00
Consent		No ratings				No ratings		
Privacy		No ratings				No ratings		
Other		No ratings				No ratings		

Discussion

Even though the HCAT was developed, tested, and refined in a UK setting, it is based upon a systematic review of the international literature, which we deem to be the most comprehensive to date (7). This was the main reason for our interest in the HCAT. In this study, we estimated both the intra-assessor and inter-assessor reliability of the HCAT to investigate its usefulness in settings outside the UK. The four assessors achieved an overall satisfactory level of reliability when using the HCAT on patient claims for

compensation. As expected, intra-assessor reliability was superior to its inter-assessor counterpart in most problem categories. The HCAT was highly reliable when identifying problem categories, but its reliability was lower when coding the *severity* of problems. However, the reliability of each sub-category was still satisfactory in most cases. The HCAT showed satisfactory reliability when coding information about the complainant and the staff group involved in the complaint, but it was very difficult to code the level of harm incurred, which resulted in low reliability scores. Overall, the HCAT seemed relatively time-effective to use; its application took, on average, less than five minutes per compensation claim, and assessors quickly became familiar with the tool.

Interpretation of findings and comparison with existing literature

Our finding that the "quality" and "safety" categories had the lowest reliability corresponds with the findings of Gillespie and Reader who developed the HCAT (8). In their study, substantial reliability was achieved in this domain, but there was still room for improvement. 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.

Our study achieved reliability estimates for the HCAT problem categories that were comparable to, and in some cases higher than, the estimated reported by Gillespie and Reader (8). It should be noted, however, that the confidence intervals of some of the reliability estimates extended below the level of substantial agreement. Other studies investigating the reliability of the HCAT have reported reliability estimates of 0.75 to 0.98 at the problem-category level (11) and reliability coefficients of 0.81 (9) to 0.92 (5) for the HCAT as a whole. No other study has reported on reliability at the sub-category level, and more extensive and robust studies are needed to establish the reliability of the HCAT at this level.

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 (20). In our analyses, low intra-assessor reliability coefficients tended to appear together with low inter-assessor reliability, indicating a possible problem in the definition of these categories or in the training of our assessors. 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 practical use.

We observed a skewed distribution among the seven HCAT problem categories, where most complaints were coded under the "quality" or "safety" categories. The study sample was a random selection of patient compensation claims, and we expect that the observed prevalence of problems reflects the true prevalence in patients' claims for compensation across the field of acute medicine. While most agreement statistics are only valid with a prevalence of around 50%, Gwet's AC1 statistic is valid with both high and low prevalences (18, 21). The HCAT tool focuses on the identification of macro trends, which could be difficult to analyze if up to 95% of all claims fall into the "quality" problem category. This emphasizes the need for reliable sub-categories. Our findings at the sub-category level pointed towards satisfactory reliability but with significant fluctuations in some sub-categories.

Strengths and Limitations

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.

We followed the HCAT manual as rigorously as possible when we coded the claim letters, and all the assessors followed the tutorial process described by Gillespie and Reader (8). In retrospect, we might have improved the reliability estimates by spending more time becoming familiar with—and agreeing on—the classes of Danish words that indicate specific problem categories. Likewise, the reliability estimates may have benefited from greater discussion among the assessors about how to rate the level of severity. Finally, it remains unclear whether a complete translation of the HCAT into Danish might have resulted in even higher reliability—as this was our first study using the HCAT, we aimed to test the reliability of the original version of the tool. It remains uncertain whether all sub-categories of problems can be found in compensation cases.

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.

In conclusion, we found that the HCAT performed successfully in a Danish healthcare setting with a different complaint system to the UK. The HCAT was shown to be a reliable tool for distinguishing problem types in patient compensation claim letters and thus has potential for future use in quality research and improvement.

Funding

This work was supported by the Region of Southern Denmark [17/33854].

Conflicts of interests

None

Data sharing

Data from DPCA is only approved for reporting and sharing at an aggregate level and can therefore not be shared. The codebook and statistical codes (in Danish) are available from the corresponding author.

Author contributions

SFB and SBB devised the project with advice from CHH and KLM. SBB was principal investigator and together with SFB was responsible for study design, protocol, data collection, and data analysis with input from CH, KLM, JHK, and KPJ. The compensation cases were scored by SFB, CHH, JHK, and KPJ. All authors contributed to interpreting the results and revising the manuscript and have read and approved the final manuscript.

Figure Legend

Figure 1 Domains, problem categories, and sub-categories of the HCAT taxonomy (cf. Gillespie A. and Reader TW (8))

References

1. Vincent CA, Coulter A. Patient safety: what about the patient? Quality & safety in health care. 2002;11(1):76-80.
2. Friele RD, Sluijs EM. Patient expectations of fair complaint handling in hospitals: empirical data. BMC Health Serv Res. 2006;6:106.
3. Bouwman R, Bomhoff M, Robben P, Friele R. Patients' perspectives on the role of their complaints in the regulatory process. Health Expect. 2016;19(2):483-96.
4. Pichert JW, Hickson G, Moore I. Using Patient Complaints to Promote Patient Safety. In: Henriksen K, Battles JB, Keyes MA, Grady ML, editors. Advances in Patient Safety: New Directions and Alternative Approaches (Vol 2: Culture and Redesign). Advances in Patient Safety. Rockville (MD)2008.
5. Harrison R, Walton M, Healy J, Smith-Merry J, Hobbs C. Patient complaints about hospital services: applying a complaint taxonomy to analyse and respond to complaints. Int J Qual Health Care. 2016;28(2):240-5.
6. de Vos MS, Hamming JF, Marang-van de Mheen PJ. The problem with using patient complaints for improvement. BMJ Qual Saf. 2018;27(9):758-62.
7. Reader TW, Gillespie A, Roberts J. Patient complaints in healthcare systems: a systematic review and coding taxonomy. BMJ Qual Saf. 2014;23(8):678-89.
8. Gillespie A, Reader TW. The Healthcare Complaints Analysis Tool: development and reliability testing of a method for service monitoring and organisational learning. BMJ Qual Saf. 2016;25(12):937-46.
9. Mattarozzi K, Sfrisi F, Caniglia F, De Palma A, Martoni M. What patients' complaints and praise tell the health practitioner: implications for health care quality. A qualitative research study. Int J Qual Health Care. 2017;29(1):83-9.
10. Bouwman R, Bomhoff M, Robben P, Friele R. Classifying Patients' Complaints for Regulatory Purposes: A Pilot Study. J Patient Saf. 2016.
11. Mack JW, Jacobson J, Frank D, Cronin AM, Horvath K, Allen V, et al. Evaluation of Patient and Family Outpatient Complaints as a Strategy to Prioritize Efforts to Improve Cancer Care Delivery. Joint Commission journal on quality and patient safety / Joint Commission Resources. 2017;43(10):498-507.
12. Jerng JS, Huang SF, Yu HY, Chan YC, Liang HJ, Liang HW, et al. Comparison of complaints to the intensive care units and those to the general wards: an analysis using the Healthcare Complaint Analysis Tool in an academic medical center in Taiwan. Crit Care. 2018;22(1):335.
13. de Vos MS, Hamming JF, Chua-Hendriks JJC, Marang-van de Mheen PJ. Connecting perspectives on quality and safety: patient-level linkage of incident, adverse event and complaint data. BMJ Qual Saf. 2019;28(3):180-9.
14. Wallace E, Cronin S, Murphy N, Cheraghi-Sohi S, MacSweeney K, Bates M, et al. Characterising patient complaints in out-of-hours general practice: a retrospective cohort study in Ireland. The British journal of general practice : the journal of the Royal College of General Practitioners. 2018;68(677):e860-e8.

15. Gillespie A, Reader TW. Patient-Centered Insights: Using Health Care Complaints to Reveal Hot Spots and Blind Spots in Quality and Safety. Milbank Q. 2018;96(3):530-67.
16. Schull MJ, Morrison LJ, Vermeulen M, Redelmeier DA. Emergency department overcrowding and ambulance transport delays for patients with chest pain. CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne. 2003;168(3):277-83.
17. Gillespie A, Reader TW. Healthcare Complaints Analysis Tool London, United Kingdom: The London School of Economics and political science; 2015 [1-20]. Available from: https://qualitysafety.bmj.com/content/suppl/2016/01/05/bmjqs-2015-004596.DC1/bmjqs-2015-004596supp_new.pdf.
18. Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement. Br J Math Stat Psychol. 2008;61(Pt 1):29-48.
19. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159-74.
20. STPK. [Årsberetning 2018 - Ankenævnet for Patienterstatningen]. Aarhus: Styrelsen for Patientklager; 2018.
21. Wongpakaran N, Wongpakaran T, Wedding D, Gwet KL. A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. BMC Med Res Methodol. 2013;13:61.

Domain	Problem category	Sub-category
Clinical problems Issues relating to the quality and safety of the clinical and nursing care provided by healthcare staff	**Quality:** Clinical standards of healthcare staff behavior	Neglect—hygiene and personal care; neglect—nourishment and hydration; neglect—general; rough handling and discomfort; examination and monitoring; making and following care plans; outcomes and side effects
	Safety: Errors, incidents, and staff competencies	Error—diagnosis; error—medication; error—general; failure to respond; clinician skills; teamwork
Management problems Issues relating to the environment and the organizations within which healthcare is provided	**Environment:** Problems with the facilities, services, clinical equipment, and staffing levels	Accommodation; preparedness; ward cleanliness; equipment; staffing; security
	Institutional processes: Problems with bureaucracy, waiting times, and accessing care	Delay—access; delay—procedure; delay—general; bureaucracy; visiting; documentation
Relationship problems Issues relating to the behavior of any specific member(s) of staff toward the patient or their family/friends	**Listening:** Information from patients disregarded or not acknowledged by healthcare staff	Ignoring patients; dismissing patients; token listening
	Communication: Absent or incorrect communication from healthcare staff to patients	Delayed communication; incorrect communication; absent communication
	Respect and patient rights: Disrespect or violation of patient rights by staff	Disrespect; confidentiality; rights; consent; privacy

Figure 1 Domains, problem categories, and sub-categories of the HCAT taxonomy (cf. Gillespie A. and Reader TW (8))

65x57mm (300 x 300 DPI)