The Healthcare Complaints Analysis Tool: development and reliability testing of a method for service monitoring and organisational learning

Alex Gillespie, Tom W Reader

ABSTRACT

Background Letters of complaint written by patients and their advocates reporting poor healthcare experiences represent an under-used data source. The lack of a method for extracting reliable data from these heterogeneous letters hinders their use for monitoring and learning. To address this gap, we report on the development and reliability testing of the Healthcare Complaints Analysis Tool (HCAT).

Methods HCAT was developed from a taxonomy of healthcare complaints reported in a previously published systematic review. It introduces the novel idea that complaints should be analysed in terms of severity. Recruiting three groups of educated lay participants (n=58, n=58, n=55), we refined the taxonomy through three iterations of discriminant content validity testing. We then supplemented this refined taxonomy with explicit coding procedures for seven problem categories (each with four levels of severity), stage of care and harm. These combined elements were further refined through iterative coding of a UK national sample of healthcare complaints (n= 25, n=80, n=137, n=839). To assess reliability and accuracy for the resultant tool, 14 educated lay participants coded a referent sample of 125 healthcare complaints.

Results The seven HCAT problem categories (quality, safety, environment, institutional processes, listening, communication, and respect and patient rights) were found to be conceptually distinct. On average, raters identified 1.94 problems (SD=0.26) per complaint letter. Coders exhibited substantial reliability in identifying problems at four levels of severity; moderate and substantial reliability in identifying stages of care (except for ‘discharge/transfer’ that was only fairly reliable) and substantial reliability in identifying overall harm.

Conclusions HCAT is not only the first reliable tool for coding complaints, it is the first tool to measure the severity of complaints. It facilitates service monitoring and organisational learning and it enables future research examining whether healthcare complaints are a leading indicator of poor service outcomes. HCAT is freely available to download and use.

INTRODUCTION

Improving the analysis of complaints by patients and families about poor healthcare experiences (herein termed ‘healthcare complaints’) is an urgent priority for service providers1–3 and researchers.4 5 It is increasingly recognised that patients can provide reliable data on a range of issues,6–12 and healthcare complaints have been shown to reveal problems in patient care (eg, medical errors, breaching clinical standards, poor communication) not captured through safety and quality monitoring systems (ie, incident reporting, case review and risk management).13–15 Patients are valuable sources of data for multiple reasons. First, patients and families, collectively, observe a huge amount of data points within healthcare settings;16 second, they have privileged access to information on continuity of care,17 18 communication failures,19 dignity issues20 and patient-centred care;21 third, once treatment is concluded, they are more free than staff to speak up;22 fourth, they are outside the organisation, thus providing an independent assessment that reflects the norms and expectations of society.23 Moreover, patients and their families filter the data, only writing complaints when a threshold of dissatisfaction has been crossed.24
Unlocking the potential of healthcare complaints requires more than encouraging and facilitating complaint reporting (eg, patients being unclear about how to complain, believing complaints to be ineffective or fearing negative consequences for their healthcare),\(^4\)\(^3\)\(^2\)\(^5\) It also requires systematic procedures for analysing the complaints, as is the case with adverse event data.\(^2\) It has even been suggested that patient complaints might actually precede, rather than follow, safety incidents, potentially acting as an early warning system.\(^5\)\(^2\)\(^6\)

However, any systematic investigation of such potential requires a reliable and valid tool for coding and analysing healthcare complaints. Existing tools lag far behind established methods for analysing adverse events and critical incidents.\(^2\)\(^7\)\(^\sim\)\(^3\)\(^1\)\(^3\) The present article answers recent calls to develop reliable method for analysing healthcare complaints.\(^4\)\(^3\)\(^5\)\(^3\)\(^1\)\(^\sim\)

A previous systematic review of 59 articles reporting healthcare complaint coding tools revealed critical limitations with the way healthcare complaints are analysed.\(^2\)\(^6\) First, there is no established taxonomy for categorising healthcare complaints. Existing taxonomies differ widely (eg, 40% do not code safety-related data), mix general issues with specific issues, fail to distinguish problems from stages of care and lack a theoretical basis. Second, there is minimal standardisation of the procedures (eg, coding guidelines, training), and no Healthcare Complaints Analysis Tool (HCAT) has been thoroughly tested for reliability (ie, that two coders will observe the same problems within a complaint). Third, analysis of healthcare complaints often overlooks secondary issues in favour of single issues. Finally, despite the varying severity of problems raised (eg, from parking charges to gross medical negligence), existing tools do not assess complaint severity.

To begin addressing these limitations, the previous systematic review\(^2\)\(^6\) aggregated the coding taxonomies from the 59 studies, revealing 729 uniquely worded codes, which were refined and conceptualised into seven categories and three broad domains (http://qualitysafety.bmj.com/content/23/8/678/F4.large.jpg). The overarching tripartite distinction between clinical, management and relational domains represents theory and practice on healthcare delivery. The ‘clinical domain’ refers to the behaviour of clinical staff and relates to the literature on human factors and safety.\(^3\)\(^3\)\(^\sim\)\(^3\)\(^5\)\(^\sim\)\(^1\)\(^3\)\(^1\) The ‘management domain’ refers to the behaviour of administrative, technical and facilities staff and relates to the literature on health service management.\(^3\)\(^6\)\(^\sim\)\(^3\)\(^8\) The ‘relationship domain’ refers to patients’ encounters with staff and relates to the literatures on patient perspectives,\(^3\)\(^5\) misunderstandings,\(^4\)\(^0\) empathy,\(^4\)\(^1\) and dignity.\(^2\)\(^0\) These domains also have an empirical basis in studies of patient–doctor interaction, where the discourses (or ‘voices’) of medicine, institutions and patients are evident,\(^4\)\(^1\)\(^\sim\)\(^2\)\(^2\)\(^0\) and clashes between the ‘system’ (clinical and management domains) and ‘lifeworld’ (relational domain) are observed.\(^4\)\(^3\)\(^\sim\)\(^4\)\(^5\) Although the taxonomy developed in the systematic review\(^2\)\(^6\) is comprehensive and theoretically informed, it remains a first step. It needs to be extended into a tool, similar to those used in adverse event research,\(^2\)\(^0\)\(^\sim\)\(^2\)\(^2\)\(^0\) that can reliably distinguish the types of problem reported, their severity and the stages of care at which they occur.

Our aim is to create a tool that supports healthcare organisations to listen\(^4\)\(^6\) to complaints, and to analyse and aggregate these data in order to improve service monitoring and organisational learning. Although healthcare complaints are heterogeneous\(^4\)\(^7\) and require detailed redress at an individual level,\(^4\)\(^8\) we demonstrate that complaints and associated severity levels can be reliably identified and aggregated. Although this process necessarily loses the voice of individual complainants, it can enable the collective voice of complainants to inform service monitoring and learning in healthcare institutions.

**METHOD**

Tool development often entails separate phases of development, refinement and testing.\(^4\)\(^9\)\(^\sim\)\(^5\)\(^0\) We developed and tested the HCAT through three phases (for which ethical approval was sought and obtained) with the following aims:

1. To test and refine the conceptual validity of the original taxonomy.
2. To develop the refined taxonomy into a comprehensive rating tool, with robust guidelines capable of distinguishing problems, their severity and stages of care.
3. To test the reliability and calibration of the tool.

**Phase 1: testing and refining discriminant content validity**

Discriminant content validity examines whether a measure (eg, questionnaire item) or code (eg, for categorising data) accurately reflects the construct in terms of content, and whether a number of measures or codes are clearly distinct in terms of content (ie, that they do not overlap).\(^5\)\(^1\) To assess whether the categories identified in the original systematic review\(^2\)\(^6\) conceptually subsumed the subcategories and whether these categories were distinct from each other, we followed a six-step discriminant content validity procedure.\(^5\)\(^1\) First, we listed definitions of the problem categories and their associated domains. Second, we listed the subcategories as the items to be sorted into the categories. Third, we recruited three groups (n=58, n=58, n=55) of non-expert, but educated lay participants from a university participant pool (comprising students from a range of degree programmes across London who were paid £5 for 30 min) to perform the sorting exercise. Fourth, participants sorted each of the subcategories into one of the seven problem categories and provided a confidence rating on a scale of 0–10. In addition, we asked participants to indicate whether the subcategory item being sorted was either a ‘problem’ or a ‘stage of
care’. Fifth, we analysed the data to examine the extent to which each subcategory item was sorted under their expected category and participants’ confidence. Finally, we used this procedure to revise the taxonomy through three rounds of testing.

**Phase 2: tool development through iterative application**

To broaden the refined taxonomy into a comprehensive tool, we first incorporated coding procedures established in the literature. To record background details, we used the codes most commonly reported in the healthcare complaint literature, namely: (1) who made the complaint (family member, patient or unspecified/other), (2) gender of the patient (female, male or unspecified/other) and (3) which staff the complaint refers to (administrative, medical, nursing or unspecified/other). To record the stage of care, we adopted the five basic stages of care coded within adverse event reports, namely: (1) admissions, (2) examination and diagnosis, (3) care on the ward, (4) operation and procedures and (5) discharge and transfers. To record harm, we used the UK National Reporting and Learning System’s risk matrix, which has a five-point scale ranging from minimal harm (1) to catastrophic harm (5).

Next, we aimed to (1) identify the range of severity for each category and identify ‘indicators’ that covered the diversity of complaints within each category, both in terms of content and severity; (2) evaluate the procedures for coding background details, stage of care and harm; and (3) establish clear guidelines for the coding process as explicit criteria have been linked to inter-rater reliability. We used an iterative qualitative approach (repeatedly applying HCAT to healthcare complaints) because it is suited for creating taxonomies (in our case indicators) that ensure a diversity of issues can be covered parsimoniously. Also, through experiencing the complexity of coding healthcare complaints, this iterative qualitative approach allowed for us to refine both the codes and the coding guidelines.

We used the Freedom of Information Act to obtain a randomised (ie, all personally identifying information removed) random sample of 52 healthcare conglomerates (2 trusts) during the period April 2011 to March 2012. We used the Freedom of Information Act to obtain a random sample of 52 healthcare conglomerates (2 trusts) during the period April 2011 to March 2012. This yielded a dataset of 1082 letters, about 1% of the 107 000 complaints received by NHS Trusts during the period. This sample reflects the population of UK healthcare complaints with a CI of 3 and a confidence level of 95%.

The authors then separately coded subsamples of the complaint letters using HCAT, subsequently meeting to discuss discrepancies. Once sufficient insight had been gained, HCAT was revised and another iteration of coding ensued. After four iterations (n= 25, n=80, n=137, n=839), the sample of complaints was exhausted, and we had reached saturation (ie, the fourth iteration resulted in minimal revisions).

**Phase 3: testing tool reliability and calibration**

To test the reliability and calibration of HCAT, we created a ‘referent standard’ of 125 healthcare complaints. This was a stratified subsample of the 1081 healthcare complaints described in the previous section. To construct the referent standard, the authors separately coded the letters and then agreed on the most appropriate ratings. Letters were included such that the referent standard comprised at least five occurrences of each problem at each severity level (ie, so it was possible to test the reliability of coding for all HCAT problems and severity levels). Because healthcare complaints often relate to multiple problem categories (and some are less common than others), it was impossible to have a completely balanced distribution (table 1). These letters were all type written (either letters or emails), digitally scanned, with length varying from 645 characters to 14 365 characters (mean 2680.58, SD 1897.03).

To test the reliability of HCAT, 14 participants with MSc-level psychology education were recruited from the host department as ‘raters’ to apply HCAT to the referent standard. We chose educated non-expert raters because complaints are routinely coded by educated non-clinical experts, for example, hospital administrators. There are no fixed criteria on the number of raters required to assess the reliability of a coding framework, and a relatively large group of raters (n=14) was recruited in order to provide a robust test of reliability and better understand any variations in coding. Raters were trained during one of two 5 h training courses (each with seven raters). Training included an introduction to HCAT, applying HCAT to 10 healthcare complaints (three in a group setting and seven individually) and receiving feedback. Raters then had 20 h to work independently to code the 125 healthcare complaints. SPSS Statistics V.21 and AgreeStat V.3.2 were used to test reliability and calibration.

| Table 1 Distribution of Healthcare Complaints Analysis Tool problem severity across the referent standard |
|---|---|---|---|---|
| Not present | Low | Medium | High |
| Quality | 81 | 10 | 22 | 12 |
| Safety | 73 | 5 | 24 | 23 |
| Environment | 101 | 6 | 10 | 8 |
| Institutional processes | 86 | 10 | 18 | 11 |
| Listening | 99 | 5 | 11 | 10 |
| Communication | 96 | 7 | 14 | 8 |
| Respect and patient rights | 88 | 19 | 13 | 5 |
First, we used Gwet's AC1 statistic to test among raters the inter-rater reliability of coding for categories and their underlying severity ratings (not present (0), low (1), medium (2) and high (3)).\textsuperscript{60, 61} This test examines the reliability of scoring for two or more coders using a categorical rating scale, taking into account skewed datasets, where there are several categories and the distributions of one rating occurs at a much higher rate than another (ie, 0s in the current study because the majority of categories are not present in each letter). Furthermore, quadratic ratings were applied, in order that large discrepancies in ratings (ie, between 0 and 3) were treated as more significant in terms of indicating poor reliability than small discrepancies (ie, between 2 and 3).\textsuperscript{60} Gwet's AC1 test was also applied to test for inter-rater reliability in coding the stages of care complained about. Although Gwet's AC1 is the most appropriate test for the data, we also calculated Fleiss' $\kappa$ because this is more commonly used and provides a more conservative test (because it ignores the skewed distribution). Finally, because harm was rated as a continuous variable, an intraclass correlation (ICC) coefficient was used to test for reliability. To interpret the coefficients, the following commonly used guidelines\textsuperscript{60, 63} were followed: 0.01–0.20 = poor/slight agreement; 0.21–0.40 = fair agreement; 0.41–0.60 = moderate agreement; 0.61–0.80 = substantial agreement and 0.81–1.00 = excellent agreement.

Second, we tested whether the 14 raters applied HCAT to the problem categories in a manner consistent with the referent standard (ie, as coded by the authors). Gwet's AC1 (weighted) was calculated by comparing each rater's coding of problem categories and severity against the referent standard and then calculating an average Gwet's AC1 score. The average inter-rater reliability coefficient (ie, across all 14 raters) was then calculated for each problem category in order to provide an overall assessment of calibration. Again, Fleiss' $\kappa$ was also calculated in order to provide a more conservative test.

**RESULTS**

**Phase 1: discriminant content validity results**

The first test of discriminant content validity revealed large differences in the correct sorting of subcategories by participants (range 21%–97%, mean = 76.19%, SD = 19.35%). There was overlap between 'institutional issues' (bureaucracy, environment, finance and billing, service issues, staffing and resources) and 'timing and access' (access and admission, delays, discharge and referrals). The 'humaneness/caring' category was also problematic, with subcategory items often miscategorised as 'patient rights' or 'communication.' Finally, participants would often classify subcategory items as a 'stage of care'.

Accordingly, we revised the problematic categories and subcategories twice. During these revisions, we removed reference to stages of care (ie, subcategory items 'admissions', 'examinations' and 'discharge'), we merged 'humaneness/caring' into 'respect and patient rights' and in light of recent literature that emphasises the importance of listening,\textsuperscript{64, 65} we created a new category 'listening' (information moving from patients to staff) as distinct from 'communication' (information moving from staff to patients). Also, we reconceptualised the management domain as 'environements' and 'institutional processes', which proved easier for participants to distinguish. The third and final test of discriminant content validity yielded much improved results, with subcategory items being correctly sorted into the categories and domains on average 85.65% of the time (range, 58%–100%; SD, 10.89%).

**Phase 2: creating the HCAT**

Applying HCAT to actual letters of healthcare complaint revealed that reliable coding at the subcategory level was difficult. However, while the raters often disagreed at the subcategory level, they agreed at the category level. Accordingly, the decision was made to focus on the reliability of the three domains and seven categories, with the subcategories shaping the severity indicators for each category. This decision to focus on the macro structure of HCAT is consistent with the overall aim of HCAT to identify macro trends rather than to identify and resolve individual complaints.

To develop severity indicators for each category, we iteratively applied the refined taxonomy to four samples ($n=25$, $n=80$, $n=137$, $n=839$) of healthcare complaints. These sample sizes were determined by the necessity to change some aspects of the tool. The increasing sample sizes reveal that fewer changes were required as the iterative refinement of the tool progressed. Rather than applying an abstract scale of severity, we identified vivid indicators of severity, appropriate to each problem category and subcategory, which should be used to guide coding. Figure 1 reports the final HCAT problem categories and illustrative severity indicators.

The coding procedures for background details, stage of care and harm proved relatively unproblematic to apply. The only modifications necessary included adding an 'unspecified or other' category for stage of care and a harm category '0' for when no information on harm was available.

Resolving disagreements about how to apply HCAT to a specific healthcare complaint led us to the development of a set of guidelines for coding healthcare complaints (box 1). The final version of the HCAT, with all the severity indicators and guidelines, is freely available to download (see online supplementary file). Figure 2 demonstrates applying HCAT to illustrative excerpts.

**Phase 3: reliability and calibration of results**

The results of the reliability analysis are reported in table 2. On average, raters applied 1.94 codes per
letter (SD, 0.26). The Gwet’s AC1 coefficients reveal that the problem categories, each with four levels of severity, were reliably coded (ie, with substantial agreement or better). Safety showed least reliability (0.69), and respect and patient rights showed most reliability (0.91). Additional analysis using Fleiss’ $\kappa$ (which takes no account of the skewed data) found moderate to substantial reliability for all problem categories and severity ratings (0.48 (listening)–0.61 (safety, respect and patient rights)). The most significant discrepancies between Gwet’s AC1 and Fleiss’ $\kappa$ occur on the items with the largest skew (ie, listening), thus underscoring the problem with Fleiss’ $\kappa$ and our rationale for privileging Gwet’s AC1. For stages of care, one showed substantial agreement (care on the ward), three showed moderate agreement (admissions, examination and diagnosis, operation or procedure) and one had only fair agreement (discharge/transfer). Demographic data were coded at substantial reliability or higher. The ICC coefficient also demonstrated harm to be coded reliably (ICC, 0.68; 95% CI 0.62 to 0.75).

The results of the calibration analysis are reported in table 3. Gwet’s AC1 scores show raters, on average, to have substantial and excellent reliability against the referent standard. Fleiss’ $\kappa$ scores show substantial agreement (0.62–0.67). Further analysis revealed some raters to be better calibrated (across all categories) against the referent standard than others.

Finally, exploratory analysis indicated that the length of letter (in terms of characters per letter) was negatively associated with reliability in coding for listening ($r=0.266, p<0.01$), communication ($r=0.211, p<0.05$) and environment ($r=0.202, p<0.05$). It was not associated with reliability in coding for respect
and patient rights, institutional processes, safety or quality. Furthermore, there was no relationship between the number of codes applied per letter and the length of the letter.

DISCUSSION

The present article has reported on the development and testing of a tool for analysing healthcare complaints. The aim is to facilitate organisational listening, to respond to the ethical imperative to listen to grievances and to improve the effectiveness of healthcare delivery by incorporating the voice of patients. Many complainants aim to contribute information that will improve healthcare delivery, yet to date there has been no reliable tool for aggregating this voice of patients in order to support system-level monitoring and learning. The lack of a reliable tool for distinguishing problem types and severity has been an obstacle. HCAT provides a reliable additional data stream for monitoring healthcare safety and quality.

Second, HCAT can contribute to understanding the relational side of patient experience. Nearly, one third of healthcare complaints relate to the relationship domain, and a better understanding of these problems, and how they relate to clinical and management practice, is essential for improving patient satisfaction and perceptions of health services. These softer aspects of care have proved difficult to monitor, and again, HCAT can provide a reliable additional data stream.

Third, HCAT can contribute to the management of healthcare. Concretely, HCAT could be integrated into existing complaint coding processes such that the HCAT severity ratings can then be extracted and passed onto managers, external monitors and researchers. HCAT could be used as an alternative metric of success in meeting standards (eg, on hospital hygiene, waiting times, patient satisfaction). It could also be used longitudinally as a means to assess clinical, management or relationship interventions. Additionally, HCAT could be used to benchmark units or regions. Accumulating normative data would allow for healthcare organisations to be compared for deviations (eg, poor or excellent complaint profiles), and

**Figure 2** Applying Healthcare Complaints Analysis Tool to letters of complaint (excerpts are illustrative, not actual). GP, general practitioner.
this would facilitate interorganisational learning (eg, sharing practice).73

Across these three domains, HCAT can bring into decision-making the distinctive voice of patients, providing an external perspective (eg, in comparison with staff and incidents reports) on the culture of healthcare organisations. For example, where safety culture is poor (and thus incident reporting likely to be low), the analysis of complaints can provide a benchmark that is independent of that poor culture.

Finally, one of the main innovations of HCAT is the ability to reliably code severity within each complaint category. To date, analysis of healthcare complaints has been limited to frequency of problem occurrence (regardless of severity). This effectively penalises institutions that actively solicit complaints to improve quality; it might be that the optimum complaint profile is a high percentage of low-severity complaints, as this would demonstrate that the institution facilitates complaints and has managed to protect against severe failures.

**Future research**

Having a reliable tool for analysing healthcare complaints paves the way for empirically examining recent suggestions that healthcare complaints might be a leading indicator of outcome variables.45 There is already evidence that complaints predict individual outcomes;34 the next question is whether a pattern of complaints can predict organisation-level outcomes. For example: Do severe clinical complaints correlate with hospital-level mortality or safety incidents? Might complaints about management correlate with waiting times? Do relationship complaints correlate with patient satisfaction? If any such relationships are found, then the question will become whether healthcare complaints are leading or lagging indicators.

---

**Table 2** Reliability of raters (n=14) coding 125 healthcare complaints

<table>
<thead>
<tr>
<th>HCAT problem categories</th>
<th>Gwet’s AC1</th>
<th>95% CI</th>
<th>Fleiss’ κ</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quality</td>
<td>0.72</td>
<td>0.65 to 0.80</td>
<td>0.50</td>
<td>0.41 to 0.58</td>
</tr>
<tr>
<td>Safety</td>
<td>0.69</td>
<td>0.61 to 0.76</td>
<td>0.61</td>
<td>0.54 to 0.69</td>
</tr>
<tr>
<td>Environment</td>
<td>0.85</td>
<td>0.88 to 0.94</td>
<td>0.60</td>
<td>0.51 to 0.70</td>
</tr>
<tr>
<td>Institutional processes</td>
<td>0.81</td>
<td>0.75 to 0.86</td>
<td>0.58</td>
<td>0.49 to 0.66</td>
</tr>
<tr>
<td>Listening</td>
<td>0.86</td>
<td>0.82 to 0.91</td>
<td>0.48</td>
<td>0.52 to 0.70</td>
</tr>
<tr>
<td>Communication</td>
<td>0.81</td>
<td>0.76 to 0.86</td>
<td>0.52</td>
<td>0.44 to 0.61</td>
</tr>
<tr>
<td>Respect and patient rights</td>
<td>0.91</td>
<td>0.88 to 0.95</td>
<td>0.61</td>
<td>0.52 to 0.70</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Stages of care</th>
<th>Gwet’s AC1</th>
<th>95% CI</th>
<th>Fleiss’ κ</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Admissions</td>
<td>0.45</td>
<td>0.47 to 0.67</td>
<td>0.45</td>
<td>0.35 to 0.55</td>
</tr>
<tr>
<td>Examination and diagnosis</td>
<td>0.57</td>
<td>0.49 to 0.65</td>
<td>0.57</td>
<td>0.50 to 0.65</td>
</tr>
<tr>
<td>Operation or procedure</td>
<td>0.58</td>
<td>0.47 to 0.68</td>
<td>0.57</td>
<td>0.47 to 0.67</td>
</tr>
<tr>
<td>Care on the ward</td>
<td>0.66</td>
<td>0.47 to 0.67</td>
<td>0.66</td>
<td>0.47 to 0.67</td>
</tr>
<tr>
<td>Discharge/transfer</td>
<td>0.38</td>
<td>0.25 to 0.50</td>
<td>0.45</td>
<td>0.35 to 0.55</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Complainer</th>
<th>Gwet’s AC1</th>
<th>95% CI</th>
<th>Fleiss’ κ</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient</td>
<td>0.90</td>
<td>0.86 to 0.94</td>
<td>0.90</td>
<td>0.86 to 0.94</td>
</tr>
<tr>
<td>Family member</td>
<td>0.89</td>
<td>0.84 to 0.94</td>
<td>0.86</td>
<td>0.81 to 0.92</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Patient gender</th>
<th>Gwet’s AC1</th>
<th>95% CI</th>
<th>Fleiss’ κ</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>0.92</td>
<td>0.88 to 0.96</td>
<td>0.85</td>
<td>0.79 to 0.92</td>
</tr>
<tr>
<td>Female</td>
<td>0.89</td>
<td>0.85 to 0.94</td>
<td>0.88</td>
<td>0.84 to 0.93</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Complained about</th>
<th>Gwet’s AC1</th>
<th>95% CI</th>
<th>Fleiss’ κ</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medical staff</td>
<td>0.63</td>
<td>0.60 to 0.70</td>
<td>0.63</td>
<td>0.56 to 0.69</td>
</tr>
<tr>
<td>Nursing staff</td>
<td>0.64</td>
<td>0.57 to 0.70</td>
<td>0.64</td>
<td>0.56 to 0.70</td>
</tr>
<tr>
<td>Administrative staff</td>
<td>0.62</td>
<td>0.54 to 0.70</td>
<td>0.62</td>
<td>0.54 to 0.70</td>
</tr>
</tbody>
</table>

p<0.001 for all tests.

**Table 3** Average calibration of raters (n=14) against the referent standard

<table>
<thead>
<tr>
<th>HCAT problem categories</th>
<th>Average Gwet’s AC1</th>
<th>Range</th>
<th>Fleiss’ κ</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quality</td>
<td>0.79</td>
<td>0.59 to 0.88</td>
<td>0.62</td>
<td>0.45 to 0.77</td>
</tr>
<tr>
<td>Safety</td>
<td>0.76</td>
<td>0.69 to 0.83</td>
<td>0.68</td>
<td>0.49 to 0.78</td>
</tr>
<tr>
<td>Environment</td>
<td>0.89</td>
<td>0.73 to 0.94</td>
<td>0.67</td>
<td>0.49 to 0.78</td>
</tr>
<tr>
<td>Institutional processes</td>
<td>0.84</td>
<td>0.73 to 0.89</td>
<td>0.63</td>
<td>0.58 to 0.72</td>
</tr>
<tr>
<td>Listening</td>
<td>0.89</td>
<td>0.82 to 0.94</td>
<td>0.62</td>
<td>0.52 to 0.77</td>
</tr>
<tr>
<td>Communication</td>
<td>0.86</td>
<td>0.72 to 0.93</td>
<td>0.62</td>
<td>0.41 to 0.76</td>
</tr>
<tr>
<td>Respect and patient rights</td>
<td>0.91</td>
<td>0.87 to 0.94</td>
<td>0.65</td>
<td>0.51 to 0.72</td>
</tr>
</tbody>
</table>

p<0.001 for all tests.

HCAT, Healthcare Complaints Analysis Tool.
Limitations
One limitation of the current research is that the inter-rater reliability, despite being moderate to substantial, has room for improvement. For example, the reliability of applying the listening, communication, and environment categories was moderately associated with length of letter, indicating the need to improve how these categories are applied to longer and potentially more complex letters. This highlights the challenge of attempting to analyse and learn from complex and diverse written experiences. Healthcare complaints report interpretations of patient experiences and HCAT, in turn, interprets and codifies these experiences. This complexity results in inevitable variability in how complaints are understood and coded, especially for the relationship problems, such as listening and communication (which showed the weakest reliability using Fleiss’ Kappa). In order to improve reliability, future research might have healthcare professionals code the letters (eg, for comparing clinical vs non-clinical rater groups). Also, given that HCAT has only been tested on complaints from the UK, further research is needed to assess its application in other national contexts.

A second limitation is that HCAT was not tested for reliability at the subcategory level; instead, we focused on the seven overarching problem categories. To make HCAT a tool that can be applied universally, we have had to reduce the specificity of the problems that it aims to reliably identify. The rationale is that it is more useful to measure severity reliably for these seven categories than have unreliable and unscaled measurements of fine-grained problems. Nonetheless, the problem categories are underpinned by more specific subcategory codes (on which the indicators are based) that could be used by healthcare institutions while retaining the basic structure of HCAT (three domains and seven categories). This would ensure that data would be comparable across institutions.

A final limitation is that the data used in the present analysis, despite coming from a range of healthcare institutions, do not include general practice (GP) complaints (because these are not handled by the NHS Trusts in the UK). Accordingly, using HCAT for GP care, a specialist unit or a specific cultural context might require some adaptation. In such cases, we recommend preserving the HCAT structure of three domains and seven categories, which we hope will prove to be broadly applicable, and instead adding appropriate severity indicators within the relevant categories.

Conclusion
Historically, healthcare complaints have been viewed as particular to a patient or member of staff. Increasingly, however, there have been calls to better use the information communicated to healthcare services through complaints. HCAT addresses these calls to identify and record the value and insight in patient reported experiences. Specifically, HCAT provides a reliable and theoretically robust framework through which healthcare complaints can be monitored, learnt from and examined in relation to healthcare outcomes.

Acknowledgements The authors would like to acknowledge Jane Roberts, Kevin Corti and Mark Noort for their help with data collection and analysis.

Contributors AG and TR contributed equally to the conceptualisation, design, acquisition, analysis, interpretation and write up of the data. Both authors contributed equally to all aspects of developing the Healthcare Complaints Analysis Tool and writing the article. Both authors approve the submitted version of the article.

Funding London School of Economics and Political Science.

Competing interests None declared.

Ethics approval London School of Economics and Political Science Research Ethics Committee.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement The data from the discriminant content validity study and the reliability study are available. Please contact the authors for details.

Open Access This is an Open Access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/4.0/

REFERENCES


Bradley EH, Curry LA, Devers KJ. Qualitative data analysis for health services research: developing taxonomy, themes, and theory. Health Serv Res 2007;42:1758–72.


Hallgren KA. Computing inter-rater reliability for observational data: an overview and tutorial. Tutor Quant Methods Psychol 2012;8:23–34.


Christiaans-Dingelhoff I, Smits M, Zwaan L, et al. To what extent are adverse events found in patient records reported by patients and healthcare professionals via complaints, claims and incident reports? BMC Health Serv Res 2011;11:49.


