SCANDINAVIAN JOURNAL OF PRIMARY HEALTH CARE
2019, VOL. 37, NO. 1, 18–29
https://doi.org/10.1080/02813432.2019.1568712

RESEARCH ARTICLE

OPEN ACCESS

Quality of out-of-hours telephone triage by general practitioners and nurses: development and testing of the AQTT – an assessment tool measuring communication, patient safety and efficiency

D. S. Graversen, A. F. Pedersen, A. H. Carlsen, F. Bro, L. Huibers and M. B. Christensen

Research Unit for General Practice, Aarhus, Denmark & Department of Public Health, Aarhus University, Aarhus C, Denmark

ABSTRACT

Objective: To develop a valid and reliable assessment tool able to measure quality of communication, patient safety and efficiency in out-of-hours (OOH) telephone triage conducted by both general practitioners (GP) and nurses.

Design: The Dutch KERNset tool was translated into Danish and supplemented with items from other existing tools. Face validity, content validity and applicability in OOH telephone triage (OOH-TT) were secured through a two-round Delphi process involving relevant stakeholders. Forty-eight OOH patient contacts were assessed by 24 assessors in test-retest and inter-rater designs.

Setting: OOH-TT services in Denmark conducted by GPs, nurses or doctors with varying medical specialisation.

Patients: Audio-recorded OOH patient contacts.

Main outcome measures: Test-retest and inter-rater reliability were analysed using ICC$_{agreement}$, Fleiss' kappa and percent agreement.

Results: Major adaptations during the Delphi process were made. The 24-item assessment tool (Assessment of Quality in Telephone Triage – AQTT) measured communicative quality, health-related quality and four overall quality aspects. The test-retest ICC$_{agreement}$ reliability was good for the overall quality of communication (0.85), health-related quality (0.83), patient safety (0.81) and efficiency (0.77) and satisfactory when assessing specific aspects. Inter-rater reliability revealed reduced reliability in ICC$_{agreement}$ and in Fleiss' kappa. Percent agreement revealed satisfactory agreements when differentiating between 'poor' and 'sufficient' quality).

Conclusion: The AQTT demonstrated high face, content and construct validity, satisfactory test-retest reliability, reduced inter-rater reliability, but satisfactory percent agreement when differentiating between 'poor' and 'sufficient' quality. The AQTT was found feasible and clinically relevant for assessing the quality of GP- and nurse-led OOH-TT.

KEYPOINTS

Comparative knowledge is sparse regarding quality of out-of-hours telephone triage conducted by general practitioners and nurses.

- The assessment tool (AQTT) enables assessment of quality in OOH telephone triage conducted by nurses and general practitioners
- AQTT is feasible and clinically relevant for assessment of communication, patient safety and efficiency.
- AQTT can be used to identify areas for improvement in telephone triage

ARTICLE HISTORY
Received 16 May 2018
Accepted 18 October 2018

KEYWORDS
After hours care; telephone; triage; outcome assessment (health care); communication; patient safety; general practitioners

Introduction

Organisation of out-of-hours (OOH) primary care services is a health-policy issue in many countries [1–5]. Since a considerable proportion of contacts to OOH concerns minor health problems [4,6,7], appropriate OOH telephone triage (OOH-TT) seems a critical step in managing the increasing workload in OOH services [4].

In Denmark, GPs conduct OOH-TT, but one of five health-administrative regions decided in 2014 to use nurses in their OOH-TT. Patient safety and cost-

CONTACT Dennis Schou Graversen d.graversen@ph.au.dk Research Unit for General Practice, Aarhus, Denmark & Department of Public Health, Aarhus University, Bartholins Allé 2, 8000 Aarhus C, Denmark

effectiveness of telephone triage, whether by GP or nurse, remain debated issues among professionals, politicians and in the public [4,8,9].

It has been suggested that triage nurses and GPs have different approaches to decision-making [10] and information gathering [11–13]. Often, nurses use Computerised Decision Support System CDSS to guide their decision-making process [12]. However, CDSS are not always used as intended [14,15] and fixation on one of the presented symptoms might limit information gathering [16]. Nurses rely on CDSS when confronted with a clinical problem outside their expertise, potentially imposing communicative challenges [14]. Comparative studies of CDSS-guided nurse and GP-led telephone triage are sparse, but Murdoch et al. found nurses to ask three times as many questions as GPs [11]. Nurses more frequently delivered declarative statements requesting confirmation of presupposed absence of symptoms, whereas GPs tended to ask more interrogative questions [11].

Tools have been published assessing quality of physician-patient communication, but many tools are not suited for telephone triage or do not cover health-related quality [17,18]. Recently, a Dutch research group developed and validated the KERNset, which assesses the quality of communication, medical content and decisions in OOH-TT conducted by nurses using CDSS [19]. However, to our knowledge, a validated tool to assess the quality of both GP- and nurse-led OOH-TT does not exist.

The aim of this study was to develop and validate an assessment tool that assesses the quality of OOH-TT conducted by GPs, other doctors or nurses in terms of communication, patient safety and efficiency, while ensuring that assessment is independent of whether CDSS is used or not and organisational differences in the acute healthcare system.

Material and methods

Development of assessment tool

In November 2015, a literature search identified one relevant assessment tool, the RICE rating scale, measuring communication in OOH-TT only [18]. Two unpublished tools, KERNset and HAAKplus, were identified through Dutch colleagues. The KERNset comprises 24 items encompassing two domains: communicative and medical quality [19]. As KERNset is a comprehensively developed assessment tool measuring both communication and medical quality in OOH-TT, the Dutch version of the KERNset was forward-backward translated into Danish in accordance with modified WHO guidelines [20] and as proposed by Sousa et al. [21] (Figure 1).

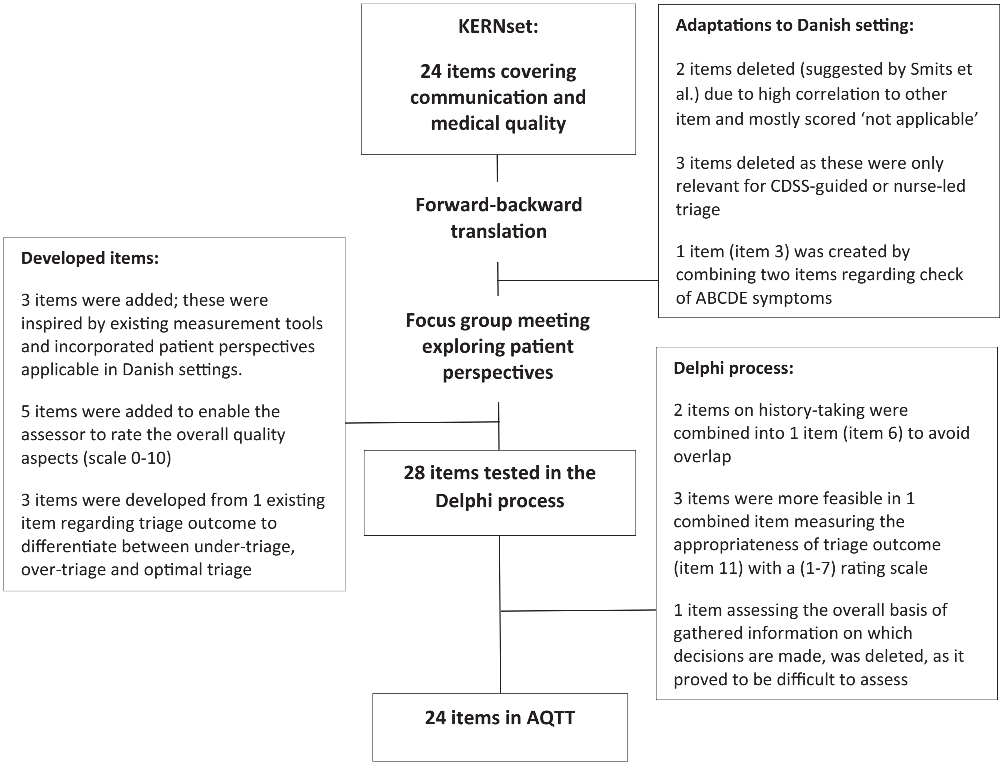
Adaptation of the assessment tool

As the KERNset was created to assess only the quality of nurse-led telephone triage [19], major adaptations needed to be made (see Figure 1). Testing of the original KERNset revealed a ceiling effect [19]. As displayed in Figure 1, we decided to rearrange, rephrase and extend the scale. This was done to include aspects of efficiency and to minimise the ceiling effect [19]. Consequently, AQTT would be more capable of differentiating between those contacts with poor quality (i.e. rated ‘1’ or ‘2’) needing improvements from those with sufficiently quality (i.e. ‘3’, ‘4’ or ‘5’) and identify aspects associated with good and poor quality. To strengthen the consistency of ratings, for each item we developed explicit descriptions for each rating in the rating manual. In the adaptation process, we incorporated relevant aspects from the RICE rating scale and HAAKplus and explored patient perspectives in a focus group interview with four patients with different age, sex, health and parent status (Figure 1). The patients expressed that the triage professional should conduct the conversation in an accommodating tone, which was incorporated in item 20. Additional patient input (e.g. ‘triage professional listens attentively’, ‘triage professional thinks aloud’ and ‘triage professional structures the call’) was incorporated in the rating manual for items 12, 13, 15, 16 and 19.

Qualitative evaluation of AQTT validity and adaptations

Delphi process

We explored face validity and content validity of the AQTT in an anonymous, two-round survey-based Delphi process. Invited experts were stakeholders (appointed health decision makers with knowledge of telephone triage from two key health-administrative regions, representatives of professional nurse and GP organisations), triage professionals (GPs, nurses and doctors representing both organisational telephone triage models) and communication experts. In the first round, experts were asked to rate the comprehensibility and relevance of each item and the accompanying rating manual, to state if important aspects were missing and to provide suggestions for improvements. In the second round, experts were asked to state


KERNset:
24 items covering communication and medical quality
Adaptations to Danish setting:
2 items deleted (suggested by Smits et al.) due to high correlation to other item and mostly scored 'not applicable'
3 items deleted as these were only relevant for CDSS-guided or nurse-led triage
1 item (item 3) was created by combining two items regarding check of ABCDE symptoms
Forward-backward translation
Developed items:
3 items were added; these were inspired by existing measurement tools and incorporated patient perspectives applicable in Danish settings.
5 items were added to enable the assessor to rate the overall quality aspects (scale 0-10)
3 items were developed from 1 existing item regarding triage outcome to differentiate between under-triage, over-triage and optimal triage
Focus group meeting exploring patient perspectives
Delphi process:
2 items on history-taking were combined into 1 item (item 6) to avoid overlap
3 items were more feasible in 1 combined item measuring the appropriateness of triage outcome (item 11) with a (1-7) rating scale
1 item assessing the overall basis of gathered information on which decisions are made, was deleted, as it proved to be difficult to assess
28 items tested in the Delphi process
24 items in AQTT

Rating scale of AQTT:

Not applicable	Only used if this aspect was correctly left out
Missing (1)	Should have been considered, but was incorrectly omitted and this could potentially have implications for patient safety or serious negative consequences for the development of the patient's situation
Insufficient (2)	Was insufficiently performed, and this could potentially have negative consequences for the development of the patient's situation
Sufficient (3)	Was just sufficiently performed, and this did probably not have negative consequences for the development of the patient's situation
Good (4)	Was well performed, although there was still room for minor improvements.
Optimal (5)	Was optimally performed, with no possibility for improvement.

Figure 1. Flowchart of AQTT development and the adapted rating scale.

whether they found the revised phrasings of items to be applicable in OOH-TT (yes/no).

Delphi results

Twenty-seven experts were included in the Delphi process in the first round, of whom 23 responded. They rated the relevance of items seven or above on a scale from one (i.e. not relevant at all) to nine (i.e. very relevant) in 87.2% of the communicative items and 89.0% of the health-related items. The authors evaluated and discussed the comments at several adaptation meetings, resulting in extensive modifications of items and rating manual (see Figure 1). During the Delphi process, it became evident that some items needed adaptations to be feasible. Items 1 and 2 could be either performed or not, resulting in a maximum rating of three. Item 11 was based on three original items and assessed the appropriateness of triage outcome on a seven-point scale. Optimal triage was rated '4', whereas increasing and decreasing ratings were used according to the degree of potential under-triage (towards 1 if patient safety is impaired) or over-triage (towards 7 in case of overuse of health resources).

Two experts resigned from the second round for personal reasons, leaving 25 experts of whom 20 responded. After the adjustments, in 25 out of 26 items at least 95% of experts stated that items were applicable in OOH-TT. Only minor comments remained, which were discussed among the authors.

The final measurement tool consisted of 24 items: elven health-related items, nine communicative items, and four items measuring overall quality of communication, health-related quality, patient safety and efficiency. The assessment tool was named "Assessment of Quality in Telephone Triage" (AQTT).

Quantitative assessment of item quality and reliability

Setting

The organisation of OOH-TT in Denmark currently differs between the five administrative regions. Four regions are organised as GP cooperatives (GPCs) where GPs conducts the telephone triage (as in The Central Danish Region), whereas the Capital Region of Denmark has established the medical helpline 1813 (MH-1813). At MH-1813 telephone triage is performed by nurses using a CDSS and doctors with diverse specialities. The two selected organisations are described in Box 1.

Recruitment of assessment panel

An assessment panel consisting of 24 doctors with triage experience from GPCs and MH-1813 was selected. Doctors were invited by email distributed to all doctors active in telephone triage in both organisations. The inclusion criterion was more than 1 year of triage experience. Fifty-six doctors from the GPC in the Central Denmark Region signed up, of which 16 GPs were randomly selected taking into account their age distribution (<45; 45–60; ≥60) (mean age: 50.8 (range: 36–75)) and gender (male/female: 9/7). Ten doctors from the MH-1813 signed up, but only eight met the inclusion criterion and were selected (mean age: 61.6 (range: 45–75), all male, specialisation: two GPs and six other specialists (internal medicine, paediatrics, anaesthesia, surgery)).

Instruction of assessors

The assessors received a 2-day training course. Firstly, lectures were held to provide them with knowledge on important factors for quality in OOH-TT. Secondly, assessors received meticulous introduction to the items and the rating manual. Thirdly, they assessed a selection of contacts with various reasons for encounter with relevant topics and ratings were discussed in plenary sessions. After the course, participants assessed a pilot telephone contact. Each assessor received individual feedback on own ratings compared with the distribution of ratings among the other assessors.

Selection of contacts

The telephone contacts were selected from a larger-scale study consisting of approximately 1950 audio-recorded patient contacts aiming to compare OOH-TT conducted by nurses using CDSS (MH-1813), doctors with varying medical specialities (MH-1813) or GPs (GPC). The inclusion period was 2 weeks from November to December 2016. Of eligible contacts 1951 contacts were selected with equal distribution of contacts triaged by nurses, doctors and GPs. 1294 of the 1951 contacts were selected randomly among all reasons for encounter and 657 contacts were selected among a group of high-risk contacts defined as patients aged above 35 years with abdominal pain. All contacts were blinded with a beep tone to conceal the educational background of the triage professional and setting.

Box 1. Description of the OOH organisation in two Regions in Denmark.

	Capital Region of Denmark: Medical helpline 1813 (MH-1813)	Central Denmark Region: GP cooperatives
Population	1.8 m citizens [27]	1.2 m citizens [28]
Annual telephone contacts in 2014 [29]	Approx. 911,000 annual contacts	Approx. 697,000 annual contacts
Organisation	Organised by the regional administration Covers telephone triage and home visits Face-to-face consultations are located in hospital facilities and managed by the EDs	Organised by GPs in the region Covers telephone triage, home visits and face-to-face consultations GPs are obliged to take part in the service
Remuneration	Payment by the hour	Fee for service
Triaging professional	Nurses obliged to use a computer decision support system (CDSS) may redirect calls to a doctor on call Doctors with different specialisations and varying experience (a minority being GPs)	GPs or GP trainees in their final year of specialisation

To examine the test-retest (intra-rater) reliability, we randomly selected one contact among the 1950 contacts for each of the 24 assessors. This contact was re-assessed by the same assessor with a median 46-day interval (IQI: 37–58). To test the inter-rater reliability, we randomly selected another contact for each of the 24 assessors allocated to two other assessors; one from the MH-1813 and one from the GPC. These random selections were made without considering the educational background of the triage professional, the reason for encounter or the triage outcome. Assessments were performed individually at home and all assessors were payed an hourly payment.

Quantitative analyses

A floor or ceiling effect was considered to be present if an item was assigned the worst or best score by more than 15% of assessors. The construct validity was analysed using spearman correlation coefficient to explore the correlation between assessors overall assessed quality of communication and the mean score of all specific communicative items, and the overall health-related quality and the mean score of all the specific health-related items. Test-retest and inter-rater reliability was estimated using a two-way mixed-effect model with the absolute agreement intraclass correlation coefficient ($ICC_{agreement}$ (3,2)) described by and interpreted as suggested by Koo et al. (i.e. ICC values of <0.5 = poor reliability, 0.5–0.75 = moderate reliability, 0.75–0.9 = good reliability and >0.9 = excellent reliability) [22]. Post-assessment process interviews revealed that the assessors occasionally found it difficult to differentiate between scores 'not applicable' or '3'. Hence, as 'not applicable' potentially covers a correctly and sufficiently performed triage, handling it as missing would result in loss of relevant information. Consequently, the rating 'not applicable' was considered equal to '3' and coded accordingly in the $ICC_{agreement}$ analyses. With 'not applicable' centred, scales are assumed ordinal.

Fleiss' kappa was calculated for the inter-rater reliability, to explore the ability of AQTT to differentiate 'poor quality' (i.e. 1 or 2) from 'sufficient quality' (i.e. 'not applicable', 3, 4 or 5), as this is clinically relevant. The appropriateness of triage outcome (item 11) was analysed to differentiate optimal or near-optimal triage (i.e. 3, 4 or 5) from clinically relevant under-triage (1, 2) or over-triage (6,7). We interpreted Fleiss' kappa as suggested by Landis and Koch (i.e. <0.0 = poor, 0.0–0.20 = slight, 0.21–0.40 = fair, 0.41–0.60 = moderate, 0.61–0.80 = substantial, 0.81–1.0 = almost perfect) [23]. To descriptively explore the inter-rater agreement, percent agreement was analysed, both for the entire range of ratings and for agreement to differentiate between poor and sufficient quality [24].

All analyses were performed in Stata 14.2 (StataCorp. 2015. *Stata Statistical Software: Release 14.2*. College Station, TX: StataCorp LP).

Results

Distribution of ratings

Table 1 summarizes the rates for each item among the 48 contacts selected in either the test-retest or the inter-rater design. The majority of ratings, in which rating was relevant ('not applicable' was excluded), were centred around 3 (i.e. 'just sufficiently performed'), and ratings were distributed across the entire scale. We found no floor effect. The Spearman correlation coefficient between the mean rating of all communicative items and the overall assessed communication was 0.86. The corresponding correlation for health-related items and overall assessed health-related quality was 0.85.

Test-retest reliability

Table 2 displays the intra-rater $ICC_{agreement}$ (3,2) in the test-retest design for each of the 24 items. Items 21–24 describe the assessors' overall rating of the quality-related aspects, and all had an estimated 'good' $ICC_{agreement}$ reliability (0.86, 0.83, 0.81 and 0.77). The majority of specific items showed 'moderate' to 'good' reliability. Two items showed 'poor' reliability (item 3, 10).

Inter-rater reliability

We analysed the inter-rater $ICC_{agreement}$ for the 24 contacts evaluated by groups of three assessors as displayed in Table 3. $ICC_{agreement}$ reliability was for most items poor, except item 2. Fleiss' kappa estimates for most items revealed a 'slight' to 'fair' inter-rater reliability when grouping the rates 1 and 2 into 'poor quality' and rates 'not applicable', 3, 4 and 5 into 'sufficient quality'. Table 3 displays the percent agreement. The average complete percent agreement within the entire scale for all items was 0.40 (range: 0.25–0.93). The average complete agreement was higher for health-related items (0.46) than for communicative items (0.32). The average percent agreement to differentiate between 'poor' (1 or 2) and 'sufficient quality' ('not applicable', 3, 4 or 5) was 0.75 for all

Table 1. Distribution of ratings in 48 selected contacts in the reliability design. A total of 24 were selected in the test-retest design and 24 in the inter-rater design.

Items	Percent rated as 'not applicable'	Floor (rating = 1)	Percent rated as 'only just sufficiently performed'	Ceiling (rating = 5)	Mean (95% CI)
INTRODUCTION					
1: Collects information about location (scale 1-3)	56.3%	9.5%	71.4%	–	2.6 (2.3–2.9)
2: Asks to speak to the patient when the caller has briefly described the situation (scale 1–3)	79.2%	50.0%	40.0%	–	1.9 (1.2–2.6)
IDENTIFICATION AND UNCOVERING					
3: Identifies and acts appropriately on signs that could be critical or life-threatening for the patient (signs of problems according to the ABCDE criteria)	75.0%	8.3%	41.7%	0.0%	2.7 (2.1–3.2)
4: Identifies and uncovers problems, including symptoms and their development	2.1%	2.1%	29.8%	10.6%	3.3 (3.0–3.6)
5: Identifies and states the purpose of the patient's contact	16.7%	2.5%	50.0%	12.5%	3.3 (2.9–3.6)
6: Prioritises the presented problems and symptoms in an appropriate way	2.1%	6.4%	25.5%	12.8%	3.3 (3.0–3.7)
7: Asks, as a minimum, all the essential questions concerning the problem(s) and symptom(s) required for optimal triage	0.0%	4.2%	27.1%	8.33%	3.2 (2.9–3.5)
8: Asks the relevant questions concerning previous medical history and medications	22.9%	10.8%	21.6%	8.11%	2.9 (2.5–3.3)
TRIAGE					
9: Gives relevant advice on self-care	39.6%	13.8%	17.2%	13.8%	3.1 (2.6–3.6)
10: Gives relevant advice on safety netting	43.8%	7.4%	14.8%	22.2%	3.6 (3.1–4.0)
11: Choses the optimal triage decision (scale 1–7)	8.3%	0.0%	75.0%*	0.0%	3.9 (3.7–4.1)*
COMMUNICATION					
12: Gives the caller sufficient time and space to describe the situation	0.0%	4.2%	31.3%	27.1%	3.7 (3.4–4.0)
13: The conversation is conducted in understandable language adapted to the caller's situation	0.0%	0.0%	16.7%	37.5%	3.9 (3.6–4.2)
14: Ensures that the triage decision and the advice given are understandable and feasible	2.1%	0.0%	38.3%	31.9%	3.9 (3.6–4.2)
15: Ensures that the caller agrees on the triage decision and advice given and is	8.3%	4.6%	38.6%	22.7%	3.5 (3.1–3.8)

(*continued*)

Table 1. Continued.

Items	Percent rated as 'not applicable'	Floor (rating = 1)	Percent rated as 'only just sufficiently performed'	Ceiling (rating = 5)	Mean (95% CI)
accommodating in case of disagreement					
16: Structures the conversation	4.2%	2.2%	23.9%	6.5%	3.4 (3.1–3.7)
17: Masters suitable questioning techniques (including suitable use of open-ended, closed-ended and non-leading questions)	0.0%	4.2%	37.5%	8.3%	3.3 (3.0–3.5)
18: Summarises (if relevant), verifies and adjusts if needed	6.3%	13.3%	37.8%	4.4%	2.9 (2.5–3.2)
19: Pays attention to the caller's experience and situation	8.3%	4.6%	36.4%	4.6%	3.0 (2.8–3.3)
20: Conducts the conversation in an accommodating and friendly tone	0.0%	2.1%	20.8%	43.8%	4.0 (3.7–4.3)
OVERALL QUALITY		Minimum (ratings: 0–1)		Maximum (ratings: 9–10)	Median rate(IQR)
21: How would you rate the overall communication in the telephone triage? (scale 0–10)		4.2%		18.8%	8 (5–8)
22: How would you rate the overall health-professional quality in the telephone triage? (scale 0–10)		8.3%		29.2%	7 (3.5–9)
23: How would you rate the overall patient safety in the telephone triage? (scale 0–10)		6.3%		47.9%	8 (5.5–9)
24: How would you rate the overall efficiency in the telephone triage? (scale 0–10)		4.2%		33.3%	8 (5–9)

Floor (rating = 1), ceiling (rating = 5) and just sufficiently (rating = 3) are the percent of ratings in contacts in which assessment is relevant ('not applicable' is excluded). Mean values are the means of ratings in which assessment is relevant ('not applicable' was excluded).
Item 11: *Centre of scale is 4 on a 1–7 scale.
Scales: Items 1 and 2 range from 1 to 3, item 11 ranges from 1 to 7, and items 21–24 range from 0 to 10
IQR: Interquartile range.

items (range: 0.61–0.94), 0.74 for health-related items and 0.76 for communicative items.

Discussion

Principal findings

We have developed the first assessment tool (AQTT) assessing the quality of communication, patient safety and efficiency of OOH-TT conducted by nurses using CDSS or doctors. The AQTT comprises 24 items with an accompanying rating manual. The AQTT demonstrated a high degree of face validity, content and construct validity. The test-retest reliability of the AQTT was satisfactory. The inter-rater reliability appeared reduced and revealed considerable disagreement among experienced and working triage professionals. However, in descriptive analyses of percent agreement when differentiating 'poor' from 'sufficient' quality, the agreement was satisfactory.

Strengths and limitations

The high degree of face validity and content validity are major strengths of the AQTT secured by the extensive development process incorporating input from patients and relevant stakeholders. The detailed AQTT rating manual is a strength as it aims to ensure the best possible consistency of assessments. Our inclusion of mostly random contacts ensured high

Table 2. Test-retest reliability in the 24 OOH telephone contacts assessed twice by the 24 assessors.

Items	$ICC_{agreement}$ (items rated 'not applicable' coded as 3)
INTRODUCTION	
1: Collects information about location (scale 1–3)	0.90
2: Asks to speak to the patient when the caller has briefly described the situation (scale 1–3)	0.86
IDENTIFICATION AND UNCOVERING	
3: Identifies and acts appropriately on signs that could be critical or life-threatening for the patient (signs of problems according to the ABCDE criteria)	0.43
4: Identifies and uncovers problems, including symptoms and their development	0.75
5: Identifies and states the purpose of the patient's contact	0.60
6: Prioritises the presented problems and symptoms in an appropriate way	0.66
7: Asks, as a minimum, all the essential questions concerning the problem(s) and symptom(s) required for optimal triage	0.77
8: Asks the relevant questions concerning previous medical history and medications	0.70
TRIAGE	
9: Gives relevant advice on self-care	0.63
10: Gives relevant advice on safety netting	0.44
11: Choses the optimal triage decision (scale 1–7)	0.53[#]
COMMUNICATION	
12: Gives the caller sufficient time and space to describe the situation	0.82
13: The conversation is conducted in understandable language adapted to the caller's situation	0.57
14: Ensures that the triage decision and the advice given are understandable and feasible	0.88
15: Ensures that the caller agrees on the triage decision and advice given and is accommodating in case of disagreement	0.68
16: Structures the conversation	0.69
17: Masters suitable questioning techniques (including suitable use of open-ended, closed-ended and non-leading questions)	0.52
18: Summarises (if relevant), verifies and adjusts if needed	0.65
19: Pays attention to the caller's experience and situation	0.77
20: Conducts the conversation in an accommodating and friendly tone	0.71
OVERALL QUALITY	
21: How would you rate the overall communication in the telephone triage? (scale 0–10)	0.86
22: How would you rate the overall health-professional quality in the telephone triage? (scale 0–10)	0.83
23: How would you rate the overall patient safety in the telephone triage? (scale 0–10)	0.81
24: How would you rate the overall efficiency in the telephone triage? (scale 0–10)	0.77

Scales: Items 1 and 2 range from 1 to 3, item 11 ranges from 1 to 7, and items 21-24 range from 0 to 10.
Item 11: [#]'not applicable' coded as 4.

representativeness of presented health problems and patient characteristics and thus simulating real-life.

The design and generally low variance of ratings holds some limitations. Preferably, a larger setup with all 24 contacts assessed by all 24 assessors could optimally have been conducted, but this was not feasible. One could challenge the generalisability of our reliability estimates due to the potentially described overrepresentation of contacts with abdominal pain. However, as these contacts are recognised as difficult to triage [25,26], we hypothesise assessment consequently is difficult possibly leading to an underestimation of the reliability.

Interpretation of findings

The satisfactory test-retest reliability indicates that assessors were consistent when assessing the same contact. Only item 3 and 10 had 'fair' reliability. This reduced reliability could be related to the frequent use of 'not applicable'. As 'not applicable' was recoded into 3, the majority of ratings were located in the centre of the scale, resulting in a low general variance in the distribution of ratings. ICC calculations consider the general variance in the denominator. Consequently, a minor variance between assessments will thus result in a largely reduced ICC reliability estimate.

Table 3. Inter-rater reliability in the 24 OOH telephone contacts assessed by three different assessors.

Item	ICC$_{\text{agreement}}$ (items rated 'not applicable' coded as 3)	Fleiss' Kappa (1,2 vs. 'not applicable', 3,4,5)	Percent agreement (No recoding)	Percent agreement (1,2 vs. 'not applicable', 3,4,5)
INTRODUCTION				
1: Collects information about location (scale 1–3)	<0.0001	−0.02	0.61	0.78
2: Asks to speak to the patient when the caller has briefly described the situation (scale 1–3)	0.67	0.68	0.93	0.94
IDENTIFICATION AND UNCOVERING				
3: Identifies and acts appropriately on signs that could be critical or life-threatening for the patient (signs of problems according to the ABCDE criteria	0.26	0.20	0.54	0.75
4: Identifies and uncovers problems, including symptoms and their development	0.22	0.33	0.40	0.72
5: Identifies and states the purpose of the patient's contact	0.15	0.16	0.31	0.75
6: Prioritises the presented problems and symptoms in an appropriate way	0.31	0.29	0.42	0.72
7: Asks, as a minimum, all the essential questions concerning the problem(s) and symptom(s) required for optimal triage	0.34	0.19	0.29	0.64
8: Asks the relevant questions concerning previous medical history and medications	0.40	0.26	0.31	0.64
TRIAGE				
9: Gives relevant advice on self-care	0.33	0.14	0.36	0.67
10: Gives relevant advice on safety netting	0.48	0.12	0.36	0.69
11: Choses the optimal triage decision (scale 1–7)	0.30[§]	0.36[^]	0.50[*]	0.89[£]
COMMUNICATION				
12: Gives the caller sufficient time and space to describe the situation	0.10	0.10	0.31	0.75
13: The conversation is conducted in understandable language adapted to the caller's situation	0.31	0.30	0.35	0.83
14: Ensures that the triage decision and the advice given are understandable and feasible	0.08	−0.03	0.40	0.94
15: Ensures that the caller agrees on the triage decision and advice given and is accommodating in case of disagreement	<0.00001	−0.11	0.25	0.81
16: Structures the conversation	0.29	0.21	0.42	0.69
17: Masters suitable questioning techniques (including suitable use of open-ended, closed-ended and non-leading questions)	0.19	0.17	0.35	0.64
18: Summarises (if relevant), verifies and adjusts if needed	0.14	0.18	0.25	0.61

(*continued*)

Table 3. Continued.

Item	$ICC_{agreement}$ (items rated 'not applicable' coded as 3)	Fleiss' Kappa (1,2 vs. 'not applicable', 3,4,5)	Percent agreement (No recoding)	Percent agreement (1,2 vs. 'not applicable', 3,4,5)
19: Pays attention to the caller's experience and situation	0.19	0.28	0.29	0.75
20: Conducts the conversation in an accommodating and friendly tone	0.22	0.09	0.28	0.86
OVERALL QUALITY				
21: How would you rate the overall communication in the telephone triage? (scale 0–10)	0.40		0.22$^{\$}$	0.63$^{\#}$
22: How would you rate the overall health-professional quality in the telephone triage? (scale 0–10)	0.36		0.17$^{\$}$	0.56$^{\#}$
23: How would you rate the overall patient safety in the telephone triage? (scale 0–10)	0.30		0.25$^{\$}$	0.57$^{\#}$
24: How would you rate the overall efficiency in the telephone triage?(scale 0–10)	0.32		0.15$^{\$}$	0.51$^{\#}$

Scales: Items 1 and 2 range from 1 to 3, item 11 ranges from 1 to 7, and items 21–24 range from 0 to 10.
Item 11: §'not applicable' coded as 4, ^ kappa analysis differentiating near-optimal (i.e. not applicable,3,4,5) from clinically relevant under-triage or over-triage (i.e. 1,2,6,7), *percent agreement within entire scale, $^{£}$percent agreement within three groups (1,2; 3,4,5; 6,7).
Overall items (21–24): exact agreement in the entire scale ($^{\$}$) and agreement within three groups (#): (0–2), (3–7), (8–10).

The inter-rater reliability was unsatisfactory for all items, except item 2, but interpretation of estimates is difficult due to the general low variance. Interpretation of estimates in the inter-rater design was complicated by a small sample size and by the fact that each contact was only assessed by groups of three and varying assessors. Due to the difficulties interpreting the inter-rater reliability, we performed descriptive analyses and found an average total percent agreement of 40%, and an average total percent agreement of 75% when differentiating between 'poor' and 'sufficient' quality. Interpretation of percent agreement is difficult as the expected agreement by chance is not accounted for [24] and no relevant cut-off exist. In our design with three assessors, an average percent agreement of 33% would in general indicate that two of three assessors agree. Consequently, the AQTT is a relevant and feasible tool as it can differentiate between 'poor' and 'sufficient' quality of telephone triage. The reduced inter-rater reliability may suggest a suboptimal assessment tool or inadequate use of the rating manual. However, from post-assessment interviews, assessment panel described thorough assessment processes (20–30 min per contact) with a constant need for consulting the rating manual. Alternatively, and merely we think it could be the result of different opinions on the quality of OOH-TT, which could be supported by the satisfactory test-retest reliability of AQTT. Hence, the variations seen between assessors could reflect true variations between how experienced triage professionals perceive the quality of triage rather than inconsistent assessments.

Findings in relation to other studies

Although a comparison with KERNset is difficult due to different designs, our findings are in line with the validation of the KERNset. Smits et al. [19] also found better agreement for test-retest reliability than for inter-rater reliability. Our study with a more thorough training and a more comprehensive rating manual confirms the impaired agreement among assessors in perception of assessment.

Implications for clinician or policy makers

Owing to the good test-retest reliability and high face validity, content validity and construct validity, the AQTT seems to be a feasible and clinically relevant assessment tool of the quality of OOH-TT conducted by doctors or nurses. Although AQTT is supplemented by a rating manual, the inter-rater agreement points at the importance of assessors to be in line with the best practice of telephone triage as reflected in AQTT. The AQTT could be used for quality assurance in

OOH-TT services on an organisational level identifying areas for improvement. Moreover, it could be used in audits to identify individual triage professionals' areas for improvement or serve as a model to educate future or practicing triage professionals.

The AQTT could in future studies explore and compare the quality on an organisational level of OOH-TT conducted by nurses using CDSS, doctors with different medical specialisations or GPs. Additionally, future research could explore the ability of AQTT to distinguish between individual triage professionals.

Acknowledgements

The authors would like to thank the 24 assessors who participated in the assessment process, the participating experts in the Delphi process and patients who gave valuable feedback. The authors thank the MH-1813 and GP cooperative organisations.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This study was supported by the Danish foundation TrygFonden, the Foundation for Primary Health Care Research (*Praksisforskningsfonden*) in the Central Denmark Region, the Committee for Quality Improvement and Continuing Medical Education in general practice (*Kvalitets- og Efteruddannelsesudvalget*) in the Central Denmark Region and the Foundation for General Practice (Praktiserende Laegers Uddannelses- og Udviklingsfond).

References

[1] Deloitte. Under pressure the funding of patient care in general practice. Epub ahead of print 2014. https://www.queensroadpartnership.co.uk/mf.ashx?ID=406a083a-144f-457d-b14b-aad537f67fc9

[2] Huibers L, Moth G, Andersen M, et al. Consumption in out-of-hours health care: Danes double Dutch? Scand J Prim Health Care. 2014;32:44–50.

[3] Christensen MB, Olesen F. Out of hours service in Denmark: evaluation five years after reform. BMJ 1998;316:1502–1506.

[4] Smits M, Rutten M, Keizer E, et al. The development and performance of after-hours primary care in the Netherlands: a narrative review. Ann Intern Med. 2017;166:737–742.

[5] Huibers L, Giesen P, Wensing M, et al. Out-of-hours care in western countries: assessment of different organizational models. BMC Health Serv Res. 2009;9:105.

[6] Nørøxe KB, Huibers L, Moth G, Vedsted P. Medical appropriateness of adult calls to Danish out-of-hours primary care: a questionnaire-based survey. BMC Fam Pract. 2017;18:34.

[7] Shipman C, Dale J. Responding to out-of-hours demand: the extent and nature of urgent need. Fam Pract. 1999;16:23–27.

[8] Campbell JL, Fletcher E, Britten N, et al. The clinical effectiveness and cost-effectiveness of telephone triage for managing same-day consultation requests in general practice: a cluster randomised controlled trial comparing general practitioner-led and nurse-led management systems with usual car. Health Technol Assess (Rockv). 2015;19:1–212.

[9] Huibers L, Smits M, Renaud V, et al. Safety of telephone triage in out-of-hours care: a systematic review. Scand J Prim Health Care. 2011;29:198–209.

[10] Wheeler SQ, Greenberg ME, Mahlmeister L, et al. Safety of clinical and non-clinical decision makers in telephone triage: a narrative review. J Telemed Telecare. 2015;21:305–322.

[11] Murdoch J, Barnes R, Pooler J, et al. Question design in nurse-led and GP-led telephone triage for same-day appointment requests: a comparative investigation. BMJ Open. 2014;4:e004515.

[12] Banning M. A review of clinical decision making: models and current research. J Clin Nurs. 2008;17:187–195.

[13] Croskerry P. A universal model of diagnostic reasoning. Acad Med. 2009;84:1022–1028.

[14] Holmström I. Decision aid software programs in telenursing: not used as intended? Experiences of Swedish telenurses. Nurs Heal Sci. 2007;9:23–28.

[15] Greatbatch D, Hanlon G, Goode J, et al. Telephone triage, expert systems and clinical expertise. Sociol Health Illn. 2005;27:802–30.

[16] Murdoch J, Barnes R, Pooler J, et al. The impact of using computer decision-support software in primary care nurse-led telephone triage: interactional dilemmas and conversational consequences. Soc Sci Med. 2015;126:36–47.

[17] Boon H, Stewart M. Patient-physician communication assessment instruments: 1986 to 1996 in review. Patient Educ Couns. 1998;35:161–176.

[18] Derkx HP, Rethans JJE, Knottnerus JA, et al. Assessing communication skills of clinical call handlers working at an out-of-hours centre: development of the RICE rating scale. Br J Gen Pract. 2007;57:383–387.

[19] Smits M, Keizer E, Ram P, et al. Development and testing of the KERNset: an instrument to assess the quality of telephone triage in out-of-hours primary care services. BMC Health Serv Res. 2017;17:798.

[20] WHO. Process of translation and adaptation of instruments. Available from: http://www.who.int/substance_abuse/research_tools/translation/en/.

[21] Sousa VD, Rojjanasrirat W. Translation, adaptation and validation of instruments or scales for use in cross-cultural health care research: a clear and user-friendly guideline. J Eval Clin Pract. 2011;17:268–274.

[22] Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med. 2016;15:155–163.

[23] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977; 33: 159–174.

[24] Gwet KL. Handbook of inter-rater reliability: the definitive guide to measuring the extent of agreement among raters. 4th ed. Gaithersburg, MD: Advanced Analytics, LLC; 2014.

[25] Philips H, Van Bergen J, Huibers L, et al. Agreement on urgency assessment between secretaries and general practitioners: an observational study in out-of-hours general practice service in Belgium. Acta Clin Belg. 2015;3286:1–6.

[26] Kristoffersen JE. Out-of-hours primary care and the patients who die. A survey of deaths after contact with a suburban primary care out-of-hours service. Scand J Prim Health Care. 2000;18:139–142.

[27] The Capital Region of Denmark. [Demographic information]. https://www.regionh.dk/om-region-hovedstaden/fakta/befolkning/Sider/Befolkning.aspx (accessed 9 April 2018).

[28] The Central Denmark Region. [Demographic information]. https://www.rm.dk/regional-udvikling/regionen-i-tal/ (accessed 9 April 2018).

[29] Ebert JF, Huibers L, Lippert FK, et al. Development and evaluation of an "emergency access button" in Danish out-of-hours primary care: a study protocol of a randomized controlled trial. BMC Health Serv Res. 2017;17:1–8.