Journal of Clinical Epidemiology 127 (2020) 117−124

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

# Limited reliability of experts' assessment of telephone triage in primary care patients with chest discomfort

Daphne C. Erkelens*, Frans H. Rutten, Loes T. Wouters, Esther de Groot, Roger A. Damoiseaux, Arno W. Hoes, Dorien L. Zwart

*Department of General Practice, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands*

Accepted 23 July 2020; Published online 28 July 2020

## Abstract

**Objective:** Root cause analyses of serious adverse events (SAE) in out-of-hours primary care (OHS-PC) often point to errors in telephone triage. Such analyses are, however, hampered by hindsight bias. We assessed whether experts, blinded to the outcome, recognize (un) safety of triage of patients with chest discomfort, and we quantified inter-rater reliability.

**Study Design and Setting:** This is a case-control study with triage recordings from 2013−2017 at OHS-PC. Cases were missed acute coronary syndromes (ACSs, considered as SAE). These cases were age- and gender-matched 1:8 with the controls, sampled from the remainder of people calling for chest discomfort. Fifteen experts listened to the recordings and rated the safety of triage. We calculated sensitivity and specificity of recognizing an ACS and the intraclass correlation.

**Results:** In total, 135 calls (15 SAE, 120 matched controls) were relistened. The experts identified ACSs with a sensitivity of 0.86 (95% CI: 0.71−0.95) and a specificity of 0.51 (95% CI: 0.43−0.58). Cases were rated significantly more often as unsafe than the controls (73.3% vs. 22.5%, $P < 0.001$). The inter-rater reliability for safety was poor: ICC 0.16 (95% CI: 0.00−0.32).

**Conclusions:** Blinded experts rated calls of missed ACSs more often as unsafe than matched control calls, but with a low level of agreement among the experts. 

*Keywords:* Acute coronary syndrome; Telephone triage; Patient safety; Serious adverse events; Out-of-hours primary care; Inter-rater reliability

## 1. Introduction

For people with chest discomfort calling the out-of-hours services in primary care (OHS-PC) adequate telephone triage is vital [1,2]. Telephone triage should help to differentiate acute coronary syndrome (ACS) from non-life−threatening conditions, to allow for timely intervention and to improve prognosis [3]. Similar to other European countries, Australia and New Zealand, telephone triage at Dutch OHS-PC is performed by triage nurses, who use a decision support tool called the "Netherlands Triage Standard." [4−11] During telephone triage, information on the patient's symptoms is collected and interpreted into an appropriate urgency level. Each level is linked with a corresponding type of care (e.g., ambulance, home visit, and consultation at OHS-PC or telephone advice, Appendix 1). It is, however, challenging to differentiate ACS from other conditions based on patient's complaints only [1,12,13]. Therefore, assigning a too low urgency level to callers with chest discomfort may occur and can result in a missed ACS, i.e., a serious adverse event (SAE).

In the Netherlands, an SAE is defined by the Healthcare Quality, Complaints, and Disputes Act as "an unintended or unexpected event related to the quality of care and resulting in death or a severe harmful event for the patient." [14].

Conflict of interest: None.

Ethics approval: The Medical Ethics Review Committee, Utrecht, The Netherlands.

Funding: This work was supported by the **Department of General Practice of the University Medical Center Utrecht**, Associate Professorship-promotion grant of D.L. Zwart, MD, PhD, the foundation "Netherlands Triage Standard" and the foundation "Stoffels-Hornstra." The views expressed are those of the authors and not necessarily those of the foundations. The funding foundations had no role in study design, data collection and analysis, preparation of the article, or decision to publish.

* Corresponding author. Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, STR 6.131, PO Box 85500, 3508 GA Utrecht, The Netherlands. Tel.: +31 (0)88 75 69289.

*E-mail address:* D.C.A.Erkelens@umcutrecht.nl (D.C. Erkelens).

https://doi.org/10.1016/j.jclinepi.2020.07.016

**What is new?**

**Key findings**

- Blinded experts rate SAE calls more often as unsafe than matched control calls.
- There is a low level of agreement among experts on the safety of OHS-PC telephone triage.

**What this adds to what was known?**

- This is the first study in which multiple expert assessments were combined to compare serious adverse events (SAE) evaluations.

**What are the implications and what should change now?**

- Learning from SAEs should not be based on single-case analysis only but also on a comparison with controls.
- Expert assessment of triage calls should be cautiously interpreted due to low inter-rater agreement.

In Dutch OHS-PC settings, almost half (46.2%) of all 240 SAEs in 2012 concerned missed acute cardiovascular disease, of which the majority were missed ACS (i.e., acute myocardial infarction and acute cardiac death) [15]. With a legally required root cause analysis investigation each SAE is scrutinized for factors contributing to the emergence of the SAE, intending to prevent similar events from occurring in the future [16,17]. Root cause analyses at the OHS-PC often pointed to triage-related errors: too low urgency allocations or assigning the incorrect type of care frequently played a role [15]. Importantly, however, investigators doing a root cause analysis are often hampered by hindsight bias, believing that what they learned during the assessment of the events, influenced by outcome knowledge, could have been known in foresight by the professionals involved in the process which led to SAEs [18,19]. As a result, investigators tend to judge harshly about the decision process of triage nurses and may draw false conclusions leading to inadequate improvement measures for future triage.

The substantial risk of hindsight bias in root cause analysis of SAEs is acknowledged by the Dutch Health and Youth Care Inspectorate, part of the Ministry of Health, Welfare, and Sport [20,21]. In scientific literature, numerous measures are proposed to decrease the influence of hindsight bias and to improve the quality of root cause analysis [17,20], for example, the use of an independent expert from another institution [20,22], using a group of multiple experts to obtain adequate coverage of the range of opinions with a weighted average as end point [23] and involving one or more experts with extensive knowledge of the subject in question [24−26]. Yet, the use of experts as a measure to decrease hindsight bias seems to be at least partly based on the assumption that experts are less sensitive to hindsight bias. From earlier research it is, however, known that experts also fall prey to hindsight bias [27−29]. To gain more insight into the effect of hindsight bias when using experts to evaluate safety and quality in root cause analysis of SAEs, a study is needed in which both SAEs and non-SAEs are included and experts are blinded to the final outcome while assessing the triage calls.

We assessed whether experts, unaware of the outcome, assess triage calls of patients calling the OHS-PC with chest discomfort in whom an ACS was missed (SAE) differently than triage calls of others' calls with chest discomfort, but in whom, the call did not end in an SAE. In addition, their inter-rater reliability was assessed.

## 2. Methods

### 2.1. Study design and setting

We conducted a retrospective, matched nested case-control study. The calls were part of a larger research project on telephone triage among callers with symptoms suggestive of acute cardiovascular disease in OHS-PC settings, of which the design is published elsewhere [30]. This case-control study included telephone triage recordings from all registered SAEs concerning missed ACS in the period 2013−2017 from a collaboration of six OHS-PC in the Netherlands. We matched SAEs with controls (1:8) based on age and gender from an existing database, which is part of the aforementioned larger research project and included telephone triage recordings of callers presenting with chest discomfort and other symptoms more or less suggestive of ACS [30]. Follow-up data on the final diagnoses of the controls were retrieved from the patients' own general practitioner's (GPs) electronic medical files.

### 2.2. Expert panel and data collection

A convenience sample of 15 GPs with ample triage consultation experience at the OHS-PC was approached by email and telephone to participate in the study. We defined triage experience as at least 5 years of experience with (telephone) triage in the OHS-PC setting, preferably in combination with additional training in cardiovascular disease or emergency medicine or with experience in the field of patient safety. The expert panel evaluated the quality and safety of the performed triage while being blinded to the final outcome (case/control status). Panel members received a description of the study domain (patients with symptoms suggestive of ACS who called the OHS-PC). Members were informed that there were SAEs within the sample but did not receive further information (e.g.,

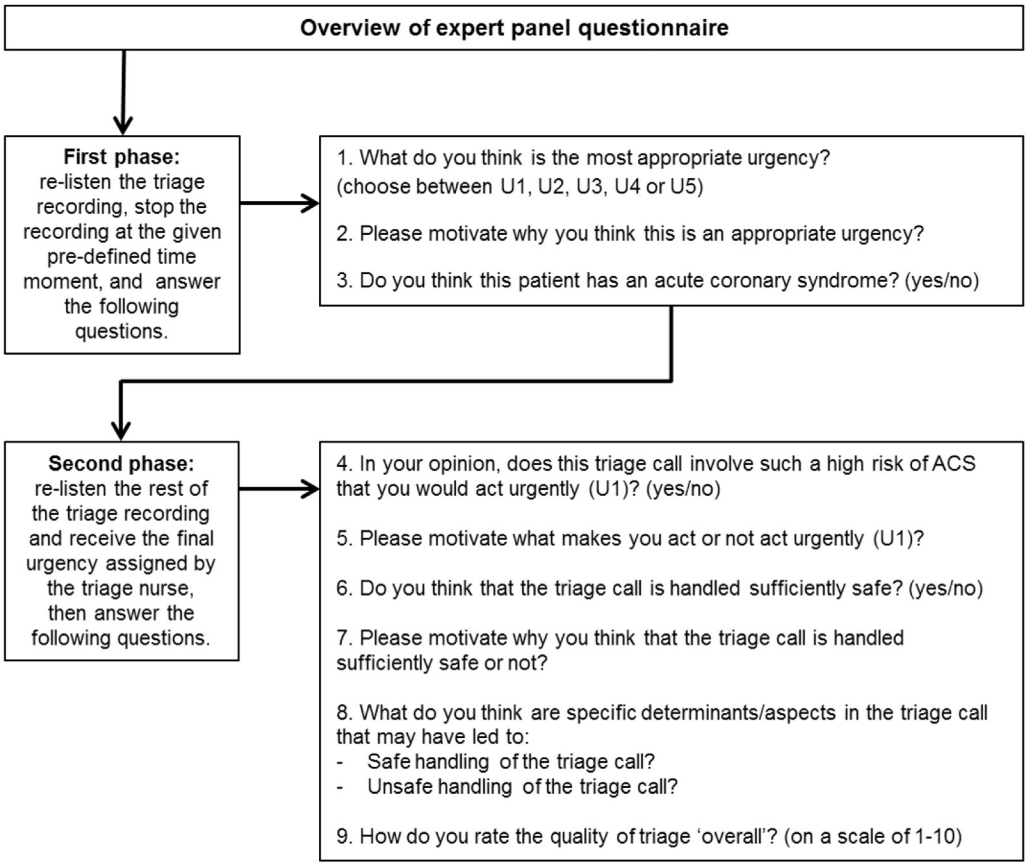


**Fig. 1.** Expert panel questionnaire.

number of SAEs and final diagnoses). We randomly allocated the triage calls to the GPs in such a way that every triage call was listened to by 2 different experts, thus with two expert assessments per triage call. For every call, we determined the exact time moment just before the triage nurse said anything that revealed the urgency or the assigned type of medical help. Panel members were obliged to stop the recordings at these predefined time moments. They received an overview of the predefined time moments for all the triage calls they had to assess. At this predefined time moment, the experts had to determine what they considered the most proper urgency allocation for that specific call (ranging from U1 to U5, see Appendix 1) [5,31] and whether or not they considered ACS present, without being influenced by the urgency assignment of the triage nurse. The experts continued listening to the remaining part of the tape (including new information about the triage nurse's final urgency allocation). At the end of the call, experts were asked to assess whether the triage nurse handled the telephone contact safely and to give an overall appraisal of the triage quality (visual analog scale ranging from zero to ten). Experts could not reconsider and change their response after completion of a question. An overview of the complete questionnaire is available in Figure 1.

### 2.3. Data analysis

We dichotomized the experts' appraisal of triage quality and defined poor triage quality as five or less on a scale from zero to ten. For differences in experts' assessments of urgency, quality, and safety of triage, we used conditional logistic regression analysis [32,33]. In addition, we calculated the sensitivity and specificity (with corresponding 95% confidence intervals) of ACS identification by the triage nurses and experts within the control group. For the triage nurses' urgencies we used an U1 urgency allocation as an expression of ACS identification, whereas for the experts, we used their answers to question number 4 from the questionnaire (see Figure 1): "Does this triage call involve such a high risk of ACS that you would act urgently (U1)?"

To quantify inter-rater reliability of the expert assessments (that is two assessments per triage recording with different experts, and not every recording was listened to by the same two experts), we calculated intraclass correlation coefficients (ICC) as an approximation of weighted Kappa's [34,35]. As suggested in previous literature on the interpretation of ICC, we considered values less than 0.40 indicative of poor inter-rater reliability, values between 0.40 and 0.75 indicative of fair to good reliability, and values greater than 0.75 indicative of excellent

reliability [34,36−38]. A $P$-value $<0.05$ was considered statistically significant for all analyses. Statistical analyses were performed using both SPSS version 25.0 and SAS version 9.4.

## 3. Results

### 3.1. Characteristics of the expert panel

Of the 15 GPs, nine were men (60%) and work experience as a GP ranged from 5 to 36 years with a median work experience of 23 years (interquartile range 6−27 years). The areas of expertise varied: four GPs completed additional training in cardiovascular disease, five in emergency medicine, and five GPs had experience in the field of patient safety (e.g., member of incident/SAE board and patient safety researcher or advisor).

### 3.2. Experts' assessments

Each triage call was assessed twice by different experts, resulting in a total amount of 270 assessments (2 × 15 SAE cases and 2 × 120 controls). Five cases concerned a missed myocardial infarction, and ten were acute cardiac deaths. Of the controls, 17.5% had an ACS, 75% had a diagnosis other than ACS of non-life−threatening origin (e.g., musculoskeletal, respiratory, or gastrointestinal disease), and in 7.5%, the exact final diagnosis was missing.

In Table 1, both the actual assigned urgencies by triage nurses and the experts' blinded assessments on the most appropriate urgency allocation are displayed. Considering the highest urgency level (U1) appropriate in suspected ACS, experts correctly allocated a high urgency in 55.2% of all calls vs. 51.1% by the triage nurses. In SAE calls, triage nurses undertriaged 28 cases (93.3% U2−U5), whereas experts would have undertriaged 13 cases (43.3% U2−U5). If we would also consider U2 as adequate, experts would have correctly allocated U1 or U2 urgency in 85.2% of all calls vs. 84.4% by the triage nurse. In SAE calls, triage nurses undertriaged 14 cases (46.7% U3−U5), whereas experts would have undertriaged 2 cases (6.6% U3−U5).

Table 2 shows the results of the experts' assessments. We compared assigned urgencies by triage nurses and experts (dichotomized into high or low). In cases, the difference between a high assigned urgency category by experts and a low assigned urgency by triage nurses was significantly more often present: 50.0% vs. 9.6%, $P < 0.001$ (U1 vs. U2−U5) and 43.4% vs. 5.4%, $P < 0.001$ (U1−U2 vs. U3−U5).

Experts considered, after listening to the first part of the triage call, ACS the most likely diagnosis in 18 of 30 case assessments, which was quite comparable with controls (60.0% vs. 56.7%, $P = 0.73$). We compared the experts' assessments that urgent acting was required due to a high risk of ACS (question 4, see Figure 1) with the actual final diagnosis of ACS and found that this relation was significantly more often present for the cases (63.3% vs. 16.2%, $P < 0.001$). A poor triage quality (≤5 on a scale of 0−10) was significantly more given by experts to calls of cases than calls of controls (33.3% vs. 10.9%, $P = 0.004$). In addition, experts considered the handling of calls of cases more often unsafe than calls of controls (73.3% vs. 22.5%, $P < 0.001$).

Within the controls, we calculated sensitivity and specificity of ACS identification by the triage nurses and experts. For the triage nurses, we compared a U1 urgency allocation (as an expression of ACS within the domain of patients calling with chest discomfort) with the other urgencies for patients with and without a final diagnosis of ACS. In 36 of 42 ACS calls, the triage nurses allocated an U1 urgency; sensitivity of 0.86 (95% CI 0.71−0.95). In addition, in 88 of 180 non-ACS calls, the triage nurse correctly allocated an urgency level lower than U1; specificity of 0.49 (95% CI 0.41−0.56).

For the experts, we compared whether urgent acting was required due to a high risk of ACS or not (question 4, see Figure 1) for patients with and without a final diagnosis of ACS. Comparable with the triage nurses, 36 of 42 ACS calls were identified by the experts as at high risk of ACS for which an U1 urgency was required; sensitivity of 0.86 (95% CI 0.71−0.95). In addition, 91 of 180 non-ACS calls were correctly assessed as non-ACS by the experts; specificity of 0.51 (95% CI 0.43−0.58).

**Table 1.** Urgencies allocated by triage nurses and experts

| | Cases (n = 30 assessments in n = 15 cases) | Controls (n = 240 assessments in n = 120 controls) | Total (n = 270 assessments) |
|---|---|---|---|
| Actual urgency allocation by triage nurses (*n* (%)) | U1: 2 (6.7)<br>U2: 14 (46.6)<br>U3: 10 (33.3)<br>U4: 2 (6.7)<br>U5: 2 (6.7) | U1: 136 (56.7)<br>U2: 76 (31.7)<br>U3: 22 (9.1)<br>U4: 0 (0)<br>U5: 6 (2.5) | U1: 138 (51.1)<br>U2: 90 (33.3)<br>U3: 32 (11.9)<br>U4: 2 (0.7)<br>U5: 8 (3.0) |
| Urgency allocation by experts (after hearing the first part of the triage call) and blinded to the actual triage nurses' allocation (*n* (%)) | U1: 17 (56.7)<br>U2: 11 (36.7)<br>U3: 1 (3.3)<br>U4: 1 (3.3)<br>U5: 0 (0.0) | U1: 132 (55.0)<br>U2: 70 (29.2)<br>U3: 30 (12.5)<br>U4: 6 (2.5)<br>U5: 2 (0.8) | U1: 149 (55.2)<br>U2: 81 (30.0)<br>U3: 31 (11.5)<br>U4: 7 (2.6)<br>U5: 2 (0.7) |

**Table 2.** Results of experts' assessments of urgency, quality, and safety of triage

| | Cases (n = 30 assessments in n = 15 cases) | Controls (n = 240 assessments in n = 120 controls) | *P*-value |
|---|---|---|---|
| Urgency allocation by experts, blinded to the actual triage nurses' allocation | U1: 17 (56.7)<br>U2: 11 (36.7)<br>U3: 1 (3.3)<br>U4: 1 (3.3)<br>U5: 0 (0.0) | U1: 132 (55.0)<br>U2: 70 (29.2)<br>U3: 30 (12.5)<br>U4: 6 (2.5)<br>U5: 2 (0.8) | 0.47 |
| Highest urgency allocation (U1) by experts, blinded to the actual triage nurses' allocation | 17 (56.7) | 132 (55.0) | 0.87 |
| High urgency allocation (U1–U2) by experts, blinded to the actual triage nurses' allocation | 28 (93.3) | 202 (84.2) | 0.21 |
| Assigned urgency by experts vs. triage nurse differs: | | | |
| Expert high (U1) vs. triage nurse low (U2-U5) | 15 (50.0) | 23 (9.6) | <0.001 |
| Expert low (U2–U5) vs. triage nurse high (U1) | 0 (0.0) | 27 (11.3) | 0.23 |
| Expert high (U1–U2) vs. triage nurse low (U3–U5) | 13 (43.4) | 13 (5.4) | <0.001 |
| Expert low (U3–U5) vs. triage nurse high (U1–U2) | 1 (3.3) | 23 (9.6) | 0.30 |
| ACS considered likely | 18 (60.0) | 136 (56.7) | 0.73 |
| Poor triage quality on a scale of 0-10 (≤5 is considered "poor") | 10 (33.3) | 26 (10.9) | 0.004 |
| Urgent acting required because of high risk of ACS | 19 (63.3) | 135 (56.3) | 0.47 |
| Urgent acting required because considered high risk of ACS and final outcome indeed ACS | 19 (63.3) | 36 (16.2) | <0.001 |
| Contact safely handled | 8 (26.7) | 186 (77.5) | <0.001 |
| Contact both safely handled and sufficient triage quality on a scale of 0–10 (6≥ is considered sufficient) | 8 (26.7) | 179 (74.6) | <0.001 |

### 3.3. Inter-rater reliability

Table 3 shows the overall inter-rater reliability for the experts' assessments of the presence of ACS, urgency levels, urgently acting, and safety and quality of triage and for the cases and controls separately. The overall inter-rater reliability of the presence of ACS after pausing the triage call at the given time was poor (ICC = 0.34), as well as for the controls (ICC = 0.32), whereas inter-rater reliability was fair for the cases (ICC = 0.44). The inter-rater reliability on urgency level was fair for cases (ICC = 0.48), controls (ICC = 0.53), and overall (0.53). Yet, overall agreement on urgently acting due to high risk of ACS (ICC = 0.37) and agreement for cases (ICC = 0.28) and controls (ICC = 0.37) apart were poor.

**Table 3.** Inter-rater reliability of experts' assessments

| | Cases ICC (95% CI) | Controls ICC (95% CI) | Overall ICC (95% CI) |
|---|---|---|---|
| After listening to triage call recording until given stop time: | | | |
| Presence of ACS | 0.44 (0.00–0.89) | 0.32 (0.16–0.48) | 0.34 (0.18–0.49) |
| After listening to whole triage call recording | | | |
| Urgency level (U1–U5) | 0.48 (0.06–0.91) | 0.53 (0.40–0.66) | 0.53 (0.40–0.65) |
| Urgently acting required because of high risk of ACS | 0.28 (0.00–0.79) | 0.37 (0.22–0.53) | 0.37 (0.22–0.51) |
| Safely handled triage call | 0.00 (0.00–0.66) | 0.09 (0.00–0.27) | 0.16 (0.00–0.32) |
| Poor triage quality | 0.40 (0.00–0.87) | 0.05 (0.00–0.23) | 0.17 (0.00–0.33) |

Agreement on the safety of the triage was lower than previous assessments, with poor inter-rater reliability for controls (ICC = 0.09) and overall (ICC = 0.16). Among the cases, agreement was the lowest (ICC = 0.00). Similarly, the inter-rater reliability for poor triage quality (defined as ≤6 on a scale of 1-10) overall (ICC = 0.17) and for controls only (ICC = 0.05) was slight. However, agreement on poor triage quality among cases was fair (ICC = 0.40).

## 4. Discussion

Fifteen GP experts assessed the urgency, safety, and quality of telephone triage calls in recordings of patients with symptoms suggestive of ACS, while blinded to the final outcome. We found that in SAE calls with missed ACS cases, triage nurses undertriaged 14 cases (46.7% U3−U5), whereas experts would have undertriaged 2 cases (6.6% U3−U5) when considering both U1 and U2 adequate. Our analysis of control calls suggests that experts are reasonably capable of identifying ACS with a sensitivity of 0.86 but less able to rule out ACS with a specificity of 0.51. We calculated sensitivity and specificity only for the control calls because we believe that the population was a more realistic reflection of our study domain of suspected ACS, and the prevalence of ACS within this control group (17.5%) approximates the prevalence of ACS within our study domain (10−12%) [39−42]. Sensitivity and specificity of ACS identification should be interpreted carefully for triage nurses because they worked under the stress of real life decision, while the experts interpreted the calls, however, without having direct responsibility. Cases were rated significantly more often as unsafe and were of lower triage quality than controls, which might suggest that experts recognize unsafety in triage of patients with chest discomfort without knowing the final outcome. However, the poor inter-rater reliability remains striking.

Many international studies have been performed on the reliability of triage systems [35,43−45], also for triage of chest pain [46−49], yet, studies on the reliability of blinded experts' assessments on recordings of triage of missed ACS have never been carried out [50]. In a previous retrospective study from New Zealand, an expert panel of GPs did examine emergency department discharges with the aim to identify cases that could have been managed in primary care. Between 37% and 50% of all cases were considered ''primary care appropriate'' by the experts, who were not blinded to the final diagnosis. Nevertheless, similar to our study, there was poor to fair agreement (Kappa 0.35−0.45) between panel members about which cases were appropriate. In addition, in 15% of all cases, GPs gave a different response to the same individual case on different occasions, illustrating the variability of clinicians' assessments, even while knowing the final diagnosis [51].

In our study, experts may have recognized unsafe triage without outcome knowledge, illustrated by the observations that they allocated higher urgencies to cases and rated them more often as unsafe. In contrast, the experts' assessments showed poor inter-rater reliability, which implies ''one expert is no expert.'' The value of a single expert assessment such as in root cause analysis, therefore, seems questionable [52]. We believe that, contrary to the individual case−oriented approach of root cause analysis [17], a blinded assessment of multiple SAEs by experts, preferably in a case-control manner provides a more realistic view on safety and quality of telephone triage in the context of daily practice.

Furthermore, a substantial proportion of the triage calls from the control group were considered as unsafe (22.5%), of poor triage quality (10.9%) or both (7.9%). This suggests room for improvement of OHS-PC telephone triage of patients with symptoms suggestive of ACS, irrespective of an adverse outcome. Similarly, in 26.7% of the SAE triage calls, the safety and quality were considered sufficient. On the other hand, aforementioned finding and the poor inter-rater reliability among experts may reflect the inherent complexity of handling triage calls concerning chest discomfort. One could argue that telephone triage is such a complex task that a certain complication rate is inevitable, in line with a recent publication on the risks of striving for ''zero harm.'' [53−56]

A strength and key feature in our study design was the blinding of the expert panel to the final outcome (case/control status), which limited the influence of hindsight bias. However, it is conceivable that experts' knowledge that SAEs were present in the study might have raised their awareness for potential SAEs. Our study is limited by missing values on the final diagnosis of some of the controls. It is possible that this 7.5% of callers could have had a ''silent'' ACS, but the possibility of an SAE as the final outcome is virtually ruled out as these control calls were not registered as SAEs in the OHS-PC database nor reported in an incident reporting system. Therefore, there were no missing values on the final outcome. Another limitation is the relatively small sample size of our study, which is mainly problematic for the reported absolute numbers and percentages in Table 2 but not for the ICC. Yet, our study is unique in blinded assessments of multiple SAEs of missed ACS, and it provides an alternative to solely a qualitative root cause analysis of a single SAE.

## 5. Conclusion

Blinded experts rated telephone triage calls of SAEs in which ACS was missed more often as unsafe and of lower triage quality than matched control calls. However, there is such a low level of agreement among the experts that the value of a single expert assessment is questionable.

## CRediT authorship contribution statement

**Daphne C. Erkelens:** Conceptualization, Data curation, Formal analysis, Methodology, Project administration,

Writing - original draft. **Frans H. Rutten:** Conceptualization, Funding acquisition, Methodology, Supervision, Writing - review & editing. **Loes T. Wouters:** Writing - review & editing. **Esther de Groot:** Writing - review & editing. **Roger A. Damoiseaux:** Writing - review & editing. **Arno W. Hoes:** Writing - review & editing. **Dorien L. Zwart:** Conceptualization, Funding acquisition, Methodology, Supervision, Writing - review & editing.

## Acknowledgments

The authors thank the out-of-hours primary care foundation ‘‘Primair Huisartsenposten’’ and the participating general practitioners of the expert panel. The authors also thank Harmke Kirkels for her help with the data collection and Paul Westers and Peter Zuithoff for their help with the data analysis.

## Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclinepi.2020.07.016.

## References

[1] International Working Group on Chest Pain in Primary C, Aerts M, Minalu G, Bosner S, Buntinx F, Burnand B, et al. Pooled individual patient data from five countries were used to derive a clinical prediction rule for coronary artery disease in primary care. J Clin Epidemiol 2017;81:120−8.

[2] Mol KA, Smoczynska A, Rahel BM, Meeder JG, Janssen L, Doevendans PA, et al. Non-cardiac chest pain: prognosis and secondary healthcare utilisation. Open Heart 2018;5(2):e000859.

[3] Bosner S, Haasenritter J, Becker A, Karatolios K, Vaucher P, Gencer B, et al. Ruling out coronary artery disease in primary care: development and validation of a simple prediction rule. CMAJ 2010;182(12):1295−300.

[4] Smits M, Rutten M, Keizer E, Wensing M, Westert G, Giesen P. The development and performance of after-hours primary care in The Netherlands: a Narrative review. Ann Intern Med 2017;166:737−42.

[5] Netherlands triage standard [Nederlandse triage Standaard]. 2019. Available at www.de-nts. Accessed October 7, 2019.

[6] Berchet C, Nader C. The organisation of out-of-hours primary care in OECD countries. In: OECD Health Working Papers no. 89. Paris: OECD Pub; 2016.

[7] Schoen C, Osborn R, Squires D, Doty M, Rasmussen P, Pierson R, et al. A survey of primary care doctors in ten countries shows progress in use of health information technology, less in other areas. Health Aff (Millwood) 2012;31(12):2805−16.

[8] Hansen EH, Hunskaar S. Telephone triage by nurses in primary care out-of-hours services in Norway: an evaluation study based on written case scenarios. BMJ Qual Saf 2011;20(5):390−6.

[9] Leibowitz R, Day S, Dunt D. A systematic review of the effect of different models of after-hours primary medical care services on clinical outcome, medical workload, and patient and GP satisfaction. Fam Pract 2003;20:311−7.

[10] Dunt D, Day SE, Kelaher M, Montalto M. Impact of standalone and embedded telephone triage systems on after hours primary medical care service utilisation and mix in Australia. Aust New Zealand Health Policy 2005;2:30.

[11] St George I, Cullen M, Gardiner L, Karabatsos G. Universal telenursing triage in Australia and New Zealand - a new primary health service. Aust Fam Physician 2008;37(6):476−9.

[12] Burman RA, Zakariassen E, Hunskaar S. Management of chest pain: a prospective study from Norwegian out-of-hours primary care. BMC Fam Pract 2014;15:51.

[13] Rawshani N, Rawshani A, Gelang C, Herlitz J, Bang A, Andersson JO, et al. Could ten questions asked by the dispatch center predict the outcome for patients with chest discomfort? Int J Cardiol 2016;209:223−5.

[14] Healthcare Quality, Complaints and Disputes Act. (WKKGZ); 2016. https://www.rijksoverheid.nl/onderwerpen/kwaliteit-van-de-zorg/wet-kwaliteit-klachten-en-geschillen-zorg. Accessed July 2, 2020.

[15] Rutten MH, Kant J, Giesen P. What can we learn from calamities at out-of-hours services in primary care? [Wat kunnen we leren van calamiteiten op de huisartsenpost?]. Huisarts Wet 2018;6(61).

[16] Reason J. Human error. New York, NY: Cambridge University Press; 1990.

[17] Peerally MF, Carr S, Waring J, Dixon-Woods M. The problem with root cause analysis. BMJ Qual Saf 2017;26(5):417−22.

[18] Henriksen K, Kaplan H. Hindsight bias, outcome knowledge and adaptive learning. Qual Saf Health Care 2003;12:ii46−50.

[19] Fischhoff B. Hindsight not equal to foresight: the effect of outcome knowledge on judgment under uncertainty. Qual Saf Health Care 1975;12:304−11.

[20] Eeuwijk J, van den Bosch J, van der Wal G, Robben PBM. With the wisdom afterwards. Hindsight and outcome bias in supervision. [Met de wijsheid achteraf. Hindsight en outcome bias in het toezicht.]. Tijdschrift voor Toezicht 2015;6(3):6−20.

[21] Robben PBM, Vedder A, Braams N, Mannée Y. Learning from five years of research, Knowledge Report, Academic Workshop Supervision. [Leren van vijf jaar onderzoek, Kenniscahier, Academische Werkplaats Toezicht]. In: Utrecht: Health and Youth Care Inspectorate. Ministry of Health, Welfare and Sport; 2017.

[22] Nicolini D, Waring J, Mengis J. Policy and practice in the use of root cause analysis to investigate clinical adverse events: mind the gap. Soc Sci Med 2011;73:217−25.

[23] Morgan MG. Use (and abuse) of expert elicitation in support of decision making for public policy. Proc Natl Acad Sci U S A 2014;111:7176−84.

[24] Pham JC, Kim GR, Natterman JP, Cover RM, Goeschel CA, Wu AW, et al. ReCASTing the RCA: an improved model for performing root cause analyses. Am J Med Qual 2010;25(3):186−91.

[25] Christensenszalanski JJJ, Willham CF. The hindsight bias - a meta-analysis. Organ Behav Hum 1991;48(1):147−68.

[26] Woloshynowych M, Rogers S, Taylor-Adams S, Vincent C. The investigation and analysis of critical incidents and adverse events in healthcare. Health Technol Assess 2005;9:1−143.

[27] Arkes HR, Wortmann RL, Saville PD, Harkness AR. Hindsight bias among physicians weighing the likelihood of diagnoses. J Appl Psychol 1981;66(2):252−4.

[28] Dawson NV, Arkes HR, Siciliano C, Blinkhorn R, Lakshmanan M, Petrelli M. Hindsight bias: an impediment to accurate probability estimation in clinicopathologic conferences. Med Decis Making 1988;8:259−64.

[29] Marks Knoll MAZ. The Effects of Expertise on the Hindsight Bias [dissertation from the internet]. Ohio: The Ohio State University; 2009. Available at https://etd.ohiolink.edu/!etd.send_file?accession=osu1242920562&disposition=inline. Accessed July 2, 2020.

[30] Erkelens DC, Wouters LT, Zwart DL, Damoiseaux RA, De Groot E, Hoes AW, et al. Optimisation of telephone triage of callers with symptoms suggestive of acute cardiovascular disease in out-of-hours primary care: observational design of the Safety First study. BMJ Open 2019;9(7):e027477.

[31] van Ierland Y, van Veen M, Huibers L, Giesen P, Moll HA. Validity of telephone and physical triage in emergency care: The Netherlands Triage System. Fam Pract 2011;28:334−41.

[32] Pearce N. Analysis of matched case-control studies. BMJ 2016;352: i969.
[33] Grobbee DE, Hoes AW. Clinical Epidemiology: Principles, Methods, and Applications for Clinical Research. Second ed. Burlington, MA: Jones & Bartlett Learning; 2015:255−301.
[34] Fleiss J, Levin B, Paik M. Statistical Methods for Rates and Proportions. Third ed. Hoboken, NJ: John Wiley & Sons Inc; 2003: 598−626.
[35] van der Wulp I, van Stel HF. Adjusting weighted kappa for severity of mistriage decreases reported reliability of emergency department triage systems: a comparative study. J Clin Epidemiol 2009;62: 1196−201.
[36] Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979;86(2):420−8.
[37] Fleiss J. Design and analysis of clinical experiments. New York: Wiley Classics Library; 1999.
[38] Landis JR, Koch GG. Measurement of observer agreement for categorical data. Biometrics 1977;33:159−74.
[39] Rawshani A, Larsson A, Gelang C, Lindqvist J, Gellerstedt M, Bang A, et al. Characteristics and outcome among patients who dial for the EMS due to chest pain. Int J Cardiol 2014;176: 859−65.
[40] Hoorweg BB, Willemsen RT, Cleef LE, Boogaerts T, Buntinx F, Glatz JF, et al. Frequency of chest pain in primary care, diagnostic tests performed and final diagnoses. Heart 2017;103: 1727−32.
[41] Frese T, Mahlmeister J, Heitzer M, Sandholzer H. Chest pain in general practice: frequency, management, and results of encounter. J Fam Med Prim Care 2016;5(1):61−6.
[42] Plat FM, Peters YAS, Loots FJ, de Groot CJA, Eckhardt T, Keizer E, et al. Ambulance dispatch versus general practitioner home visit for highly urgent out-of-hours primary care. Fam Pract 2017.
[43] Rutschmann OT, Kossovsky M, Geissbuhler A, Perneger TV, Vermeulen B, Simon J, et al. Interactive triage simulator revealed important variability in both process and outcome of emergency triage. J Clin Epidemiol 2006;59:615−21.
[44] van der Wulp I, van Stel HF. Calculating kappas from adjusted data improved the comparability of the reliability of triage systems: a comparative study. J Clin Epidemiol 2010;63:1256−63.
[45] Kuriyama A, Urushidani S, Nakayama T. Five-level emergency triage systems: variation in assessment of validity. Emerg Med J 2017; 34(11):703−10.
[46] Nishi F, de Oliveira Motta Maia F, de Souza Santos I, de Almeida Lopes Monteiro da Cruz D. Assessing sensitivity and specificity of the Manchester Triage System in the evaluation of acute coronary syndrome in adult patients in emergency care: a systematic review. JBI Database System Rev Implement Rep 2017;15(6):1747−61.
[47] Leite L, Baptista R, Leitao J, Cochicho J, Breda F, Elvas L, et al. Chest pain in the emergency department: risk stratification with Manchester triage system and HEART score. BMC Cardiovasc Disord 2015;15:48.
[48] Nishi FA, Polak C, Cruz D. Sensitivity and specificity of the Manchester Triage System in risk prioritization of patients with acute myocardial infarction who present with chest pain. Eur J Cardiovasc Nurs 2018;17(7):660−6.
[49] Matias C, Oliveira R, Duarte R, Bico P, Mendonca C, Nuno L, et al. The Manchester Triage System in acute coronary syndromes. Rev Port Cardiol 2008;27(2):205−16.
[50] Bertens LC, Broekhuizen BD, Naaktgeboren CA, Rutten FH, Hoes AW, van Mourik Y, et al. Use of expert panels to define the reference standard in diagnostic research: a systematic review of published methods and reporting. PLoS Med 2013;10(10):e1001531.
[51] Elley CR, Randall PJ, Bratt D, Freeman P. Can primary care patients be identified within an emergency department workload? N Z Med J 2007;120(1256):U2583.
[52] van Houten CB, Naaktgeboren CA, Ashkenazi-Hoffnung L, Ashkenazi S, Avis W, Chistyakov I, et al. Expert panel diagnosis demonstrated high reproducibility as reference standard in infectious diseases. J Clin Epidemiol 2019;112:20−7.
[53] Thomas EJ. The harms of promoting 'Zero Harm. BMJ Qual Saf 2020;29(1):4−6.
[54] Amalberti R, Vincent C. Managing risk in hazardous conditions: improvisation is not enough. BMJ Qual Saf 2020;29(1):60−3.
[55] Carthey J, de Leval MR, Reason JT. Institutional resilience in healthcare systems. Qual Health Care 2001;10(1):29−32.
[56] Welsh D, Bush J, Thiel C, Bonner J. Reconceptualizing goal setting's dark side: the ethical consequences of learning versus outcome goals. Organ Behav Hum 2019;150:14−27.