Reliability and validity of emergency department triage systems

Ineke van der Wulp

Reliability and validity of emergency department triage systems

Utrecht, Universiteit Utrecht, Faculteit Geneeskunde
Thesis with a summary in Dutch
Proefschrift, met een samenvatting in het Nederlands

ISBN: 978-90-393-5276-2
Author: Ineke van der Wulp
Layout: Ineke van der Wulp
Cover design: Eva van der Wulp
Print: Gildeprint Drukkerijen

Financial support for this thesis was kindly given by:
Julius Center for Health Sciences and Primary Care

Reliability and validity of emergency department triage systems

Betrouwbaarheid en validiteit van triagesystemen op de spoedeisende hulp
(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht
op gezag van de rector magnificus, prof.dr. J.C. Stoof,
ingevolge het besluit van het college voor promoties in het openbaar te verdedigen
op donderdag 25 februari 2010 des middags te 12.45 uur

door

Ineke van der Wulp

geboren op 11 mei 1982
te Nuenen, Gerwen en Nederwetten

Promotor: Prof. dr. A.J.P. Schrijvers

Co-promotor: Dr. H.F. van Stel

'Nothing is more difficult, and therefore more precious, than to be able to decide'
Napoleon Bonaparte

Contents

1

General Introduction

Triage, the sorting of patients for treatment priority was first used on battlefields in the eighteenth century during the Napoleontic wars. The purpose of triage was to treat those soldiers who were at increased risk of dying without any treatment[1]. Nowadays, triage is applied in various healthcare settings such as in mass casualty incidents, the intensive care unit, and emergency departments (EDs)[2]. Triage systems tend to rely on three different healthcare values[3]. First, they intend to protect endangered human lives and human health. These systems therefore prioritise patients with urgent care needs to treatment while less severely ill or injured patients can safely wait. However, in case several patients have to wait for live-saving interventions because one patient needs too many resources, the latter patient will not be treated first. This situation is related to the second healthcare value, efficient use of resources. Because healthcare resources are scarce, these resources will be allocated to the patients in greatest need and with the largest probability of survival. The third and final value on which triage systems rely is fairness and refers to the use of established guidelines for allocating resources to patients. With these guidelines, decisions are made on the basis of standards instead of personal preferences.

ED triage

In descriptions of the concept 'ED triage' several aspects can be distinguished of which three of these, prioritisation of need, the use of guidelines and efficient use of resources are directly related to the above mentioned healthcare values. These aspects are found in the description of ED triage by Travers et al.[4]: '*a method of categorisation based upon a number of concerns, including the severity of illness or injury, prioritisation of patients for treatment, and making the most of ED operations*'. A fourth aspect of ED triage is time. The urgency of the patient's complaint should be assessed as soon as possible after arrival in the ED. This is described by Tanabe et al.[5]: '*triage is the ability to rapidly determine patient acuity*', and Elshove-Bolk et al.[6]: '*the rapid and preliminary assessment of patients identifying those who need to be seen quickly and those who can wait*'. Finally, ED triage is a dynamic process which means that triage must be repeated while patients are waiting in the ED. This aspect of ED triage is described in several triage guidelines[7-10] and in particular described by Drijver et al.[11]: *'the dynamic process of determining urgency and selecting an appropriate location for follow-up'*. No descriptions of ED triage in the literature are found in which all aspects of ED triage are described. Therefore, a new description of ED triage, consisting of a combination of the before mentioned descriptions, will be used in this thesis: '*ED triage is a dynamic process of rapidly and systematically categorising the patient's severity of illness or injury and the patient's priority for treatment, to efficiently use ED resources*'.

Implementation of ED triage systems

Healthcare reform in the United States in the 1960's caused an increasing number of patients presenting to the ED[12]. As a result, EDs implemented triage systems to ensure that patients who need to be seen quickly were actually treated first and that others could safely wait in the waiting room. These systems consisted of three categories. In Australia in the 1970's EDs experienced an increasing number of patients who presented to the ED by ambulance. This resulted in the development of several local five-level triage systems[12;13]. These triage systems formed the basis of the Australasian Triage Scale which was implemented in 1993. A few years later, in the United States in 1998, a study was published that evaluated the reliability and validity of a three-level triage system[14]. The results of this study were poor which resulted in the development of a new five-level triage system, the

Emergency Severity Index (ESI)[9]. In the same period, Canada and the United Kingdom developed and implemented the Canadian Triage and Acuity Scale and the Manchester Triage System (MTS) respectively, because of increasing patient volumes in the ED and the need to work more systematically[10;15].
In the Netherlands, patients increasingly bypassed their general practitioner and went straight to EDs[16;17]. As a result, patient volumes increased which jeopardized ED patient flow. Moreover, because of the aging population it is expected that in the upcoming years the demand for ED care will increase[17;18]. These developments have led to the implementation of ED triage systems at the beginning of the 21st century in the Netherlands. The MTS and the ESI are the most frequently implemented five-level triage systems, 55.3% and 7.1%, respectively[17]. Furthermore, a new five-level triage system has been developed recently, the Netherlands Triage System. This system provides uniform triage between different professionals in emergency medicine: ED nurses, general practitioners, and ambulance nurses. The system is currently being tested[11]. The focus in this thesis will be on the MTS and the ESI because these systems have predominantly been implemented on EDs in the Netherlands.

MTS and ESI
The MTS consists of 52 flowcharts each representing a patient complaint. An example of such a flowchart is presented in figure 1. Each flowchart contains discriminators which allow triage nurses to allocate a patient into an urgency category. An urgency category represents a maximum waiting time in the ED for the patient to see a doctor. Patients who need to be seen by a doctor immediately are triaged in the most urgent triage category of the system, category red. When patients are triaged in the second urgency category, orange, they can wait up to ten minutes. The third category, yellow, contains patients who can wait up to sixty minutes. Patients triaged in category green can wait up to 120 minutes and patients triaged in category blue can wait up to 240 minutes. An important discriminator of the MTS is pain. Pain can be assessed with the systems' pain ruler. This instrument consists of a visual analogue scale, a pain behaviour tool and a verbal descriptor scale. Pain assessments can direct a nurse immediately to the patient's level of urgency[10].
In contrast to the MTS, the ESI consists of one flowchart (figure 2) and does not define maximum waiting times to see a doctor. The discriminators of the two most urgent triage categories are similar to the MTS. Patients triaged in these categories need immediate life-saving interventions or have severely disturbed vital signs. These patients need to see a doctor as soon as possible. In the other triage categories, patients are triaged either because their vital signs are in a predefined danger zone or by counting the number of ED resources they need. Resources are in the system defined as: laboratory tests, specialty consultations, radiology, several simple and complex procedures e.g. laceration repair, intravenous fluids and medications or, intramuscular and nebulized medications. In case the triage nurse estimates that the patient needs two or more of these resources or the patient's vital signs are in a predefined danger zone, the patient is triaged in the third category. A patient is triaged in the fourth category when the triage nurse estimates that the patient needs one resource while patients who do not need any resources are triaged in the fifth triage category[9].

Figure 1. Flowchart Chest pain from the MTS

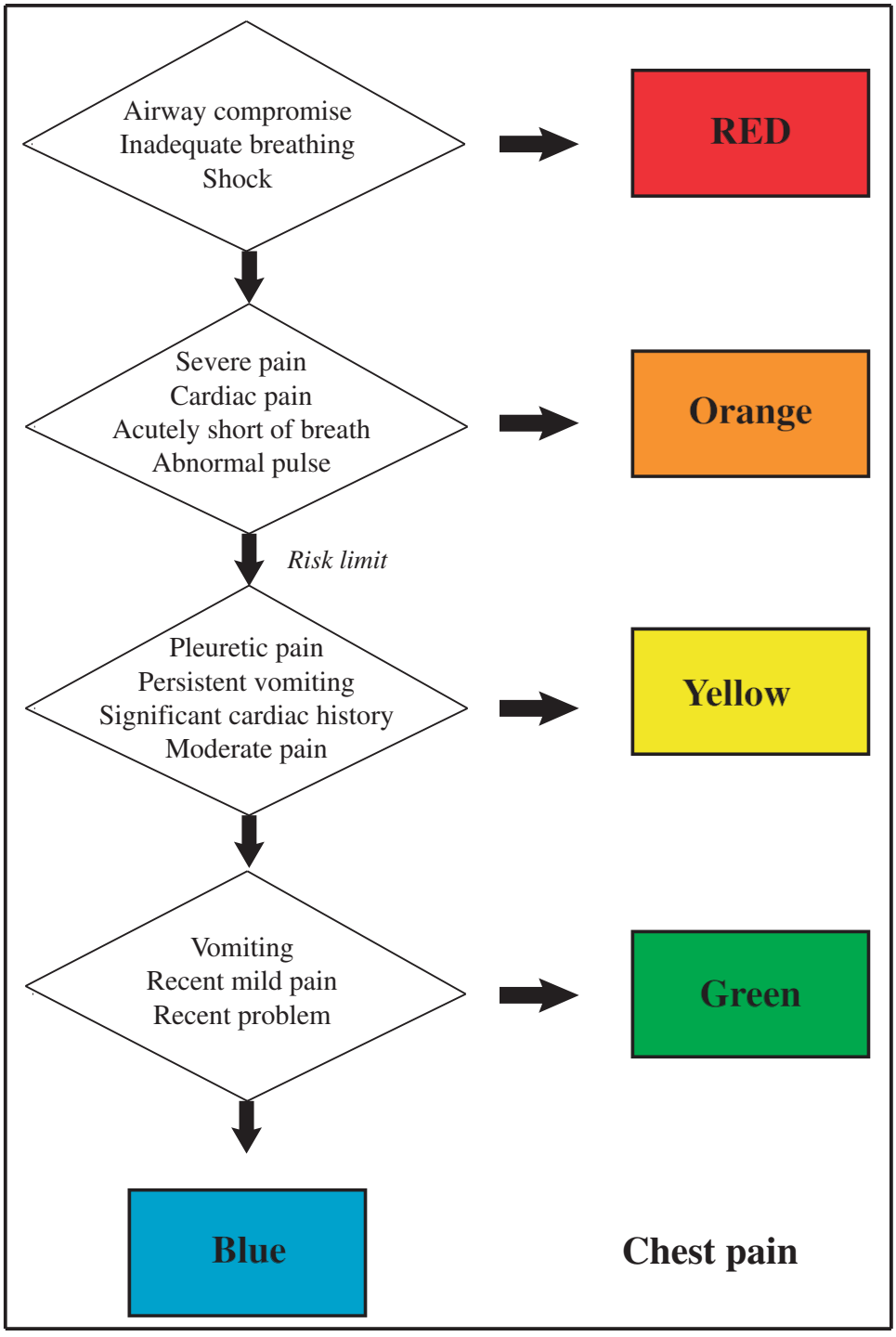
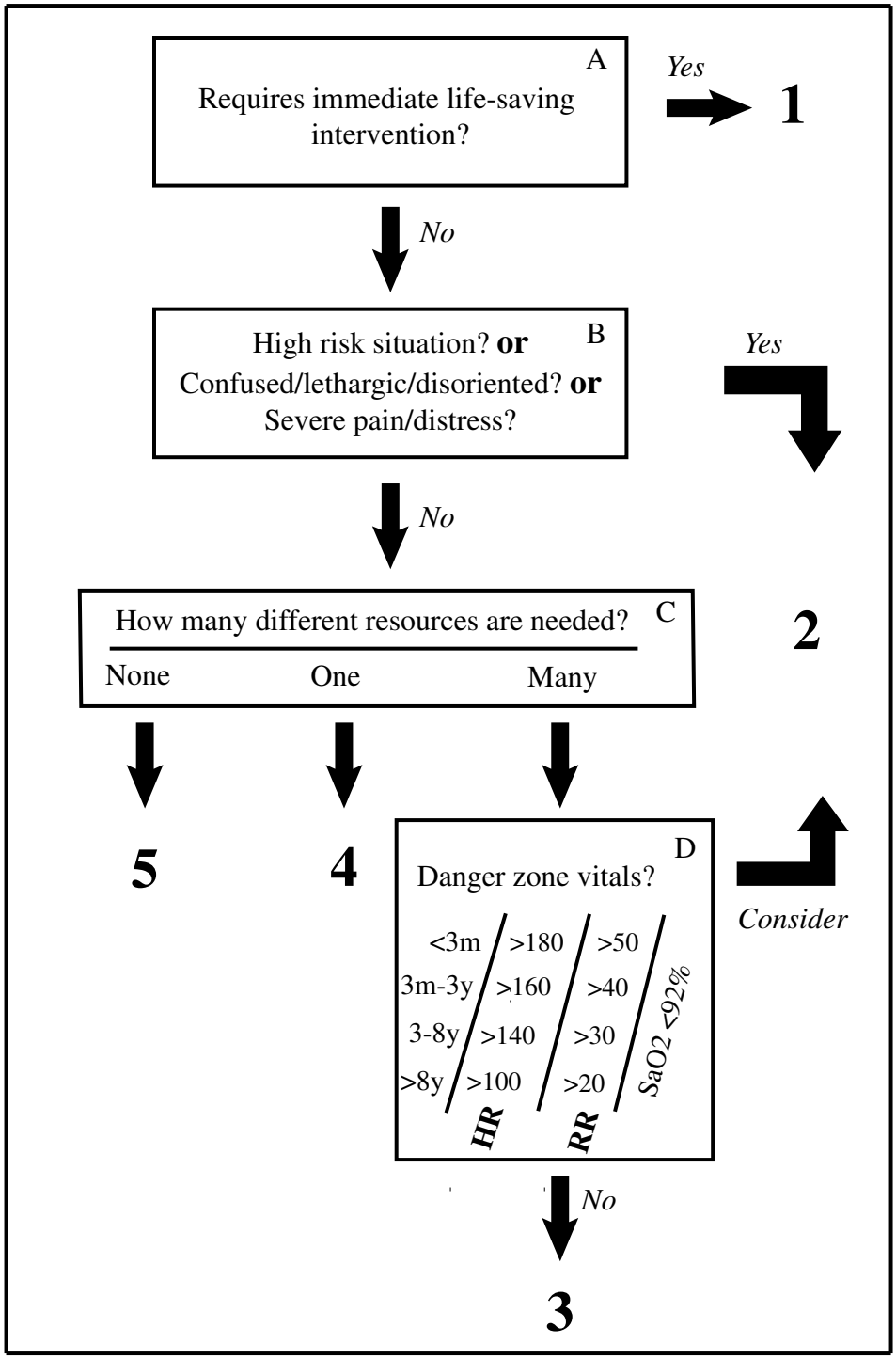


Figure 2. The Emergency Severity Index

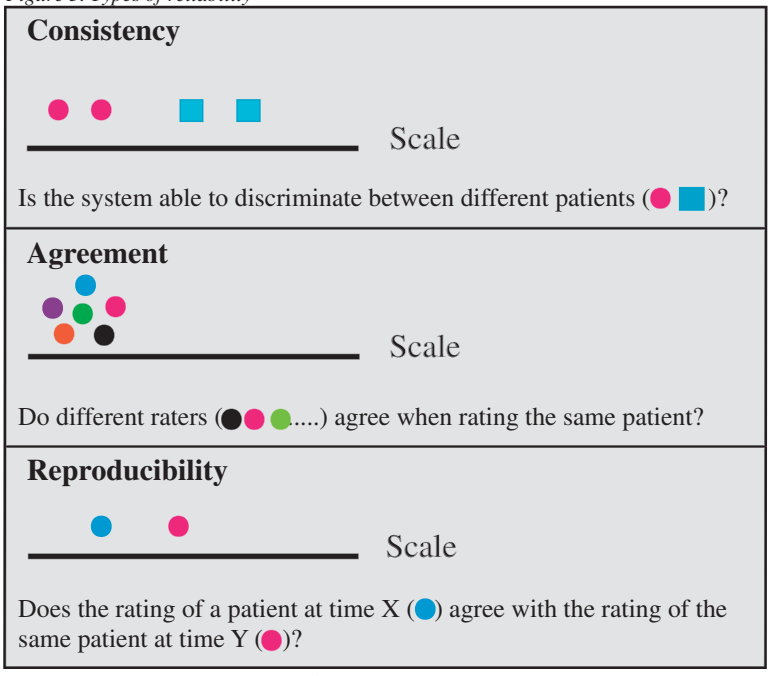


Reliability

Streiner and Norman define reliability as, '*a fundamental way to define the amount of error, both random and systematic, inherent in any measurement*'[19]. They distinguish three types of reliability: consistency, agreement and reproducibility (figure 3). Consistency is measured in case one is interested in the reliability of ratings, while agreement is measured when one is interested in the reliability of raters who use the same measurement scale. In case of consistency, one questions whether a triage system is able to discriminate between different patients. In these studies a reference standard can be applied which determines the extent to which patients differ by urgency. In case of agreement, one measures whether different raters rate a patient into the same urgency category. Reproducibility is assessed when measuring the stability of a scale over short periods of time (usually 2-14 days) within raters. This is also known as the test-retest reliability[19;20]. The reproducibility of a triage system can be assessed, for example, by rating the same patient cases twice by the same nurses.

The reliability of a measurement scale can be expressed by the intraclass correlation coefficient in case of continues measurement scales or the kappa statistic in case of categorical measurement scales such as triage systems. The kappa statistic expresses the amount of agreement between raters or methods corrected for agreement that exists due to chance. Kappa ranges between -1 (perfect disagreement) and 1 (perfect agreement)[21].

Figure 3. Types of reliability



Example derived from: de Vet et al.[40]

Validity

Is the scale measuring what it intends to measure, is known as validity[19]. For triage systems this comprises: does the system actually measure the urgency of the patient's complaint?

Three types of validity are frequently mentioned in the literature although they all aim to draw inferences from scores on scales. First, content validity which refers to the extent to which the content of a scale measures the characteristic or domain it is intended to measure[20]. For example, in the MTS this implies that a patient triaged in category orange is expected to deteriorate earlier after arrival on the ED than a patient triaged in the green category. The second type of validity is criterion validity which measures the correlation of a scale with a 'gold standard' or criterion measure. A gold standard can be a different scale which has been used and accepted in the field. Criterion validity can be studied by measuring the correlation between the scale and the 'gold standard' at the same time (concurrent validity) or at some time in the future (predictive validity)[19]. In triage systems this type of validity is difficult to study because a 'gold standard' is absent. Several studies therefore created expert panels who judged the 'true' urgency score[22-30]. Disagreement between the experts and the study participants (usually triage nurses) is calculated and reported as percentages over- and undertriage[27;31-33]. Overtriage is defined as: the assignment of a more urgent triage category than necessary, while undertriage is the opposite[34]. In addition to criterion validity, another option is to study the association between the measurement scale and a hypothetical construct. This is also known as construct validity, which is the third and final type of validity. There are several methods to measure the construct validity. First, by creating two groups of which one has the behaviour or trait while the other has not. A significant difference between the scores of the two groups is expected (discriminative validity). Second, by measuring correlations between the new scale and other variables which are related to the same construct (convergent validity).

The third method is opposite to the second and measures the correlation between the scale and variables that are unrelated to the construct (discriminant validity). The fourth and final method to measure construct validity is to simultaneously measure the convergent and discriminant validity (multitrait-multimethod matrix)[19]. In triage systems the construct validity has been studied by measuring the association between urgency categories and, for example, admission rates (convergent validity)[22;29;30;35-39]. In these studies it is hypothesized that patients who are triaged in the more urgent triage categories are more likely to be admitted than patients triaged in the less urgent triage categories. Other constructs that have been studied are associations between urgency categories and hospital length of stay, ED length of stay, mortality, survival time after ED visit and resource use[5;6;22;30;35;39].

Objectives and study questions

This thesis focuses on the reliability and validity of ED triage systems. Triage systems need to be reliable and valid because they can affect patient safety in the ED. In case of undertriage patients are at risk for deterioration. On the other hand, in case of overtriage too many ED resources are used for patients who do not necessarily need them. As a consequence, other (more urgent) patients have to wait which increases their risk for deterioration.

The objective of this thesis is to extend current knowledge about the reliability and validity of ED triage systems, and to improve triage research by evaluating methodological aspects. To reach this objective, three study questions are formulated:

- How reliable and valid are the MTS and ESI?
- What changes to the MTS and ESI are necessary to improve the reliability and validity?

- What changes in triage research methods are necessary to improve interpretations of the reliability and validity of triage systems?

Outline of this thesis

This thesis consists of two parts. The first part focuses on the reliability of triage systems and methodological aspects of reliability studies. Chapter 2 describes a study conducted to the reliability of the MTS by means of consistency and reproducibility. In this study, 48 triage nurses of two hospitals each rated fifty patient vignettes twice. Their ratings are compared with the ratings of two MTS experts. The behaviour of the kappa statistic in this study required further study. Chapter 3 therefore presents an alternative weighting scheme for the calculation of weighted kappa in triage reliability studies. The goal of this weighting scheme is to reflect clinical practice more precise than existing weighting schemes by taking into account the severity of over- and undertriage. Furthermore, kappa is also influenced by the distribution of ratings which have led to unequal comparisons of triage systems with different numbers of categories. Chapter 4 introduces an approach for interpreting the reliability especially in case comparisons between different systems are made. This approach also affects sample size calculations. Chapter 5 describes a study in which methods for calculating sample sizes in triage reliability studies are reviewed.

The second part of this thesis focuses on the validity of triage systems. In chapter 6 the construct validity of the MTS and the ESI is compared. In this study the association between the urgency categories of both triage systems and two outcome measures, admission and mortality, is assessed. Chapter 7 describes characteristics of patients who are triaged with the ESI as not needing any ED resources (ESI 5) but who were subsequently admitted to the hospital or sent to the outpatient department for follow-up (content validity). Chapter 8 focuses on the discriminator pain in the MTS (content validity). It evaluates the frequency that nurses use this tool at triage as well as characteristics of patients, nurses and triage conversations in which pain assessments are conducted. In this chapter also the reasons for not conducting pain assessments according to the system's guidelines were studied.

Finally, table 1 provides an overview of the study questions and the chapters in which these are addressed.

Table 1. Study questions related to the chapters of this thesis

Study question	Chapters
How reliable and valid are the MTS and ESI?	2-4, 6-8
What changes to the MTS and ESI are necessary to improve the reliability and validity?	7,8
What changes in triage research methods are necessary to improve interpretations of the reliability and validity of triage systems?	3-5

REFERENCES

1 Baker R, Strosberg M. Triage and equality: an historical reassessment of utilitarian analyses of triage. Kennedy Institute of Ethics Journal. 1992;2:103-123.

2 Iserson KV, Moskop JC. Triage in medicine, part I: Concept, history, and types. Annals of Emergency Medicine. 2007;49:275-281.

3 Moskop JC, Iserson KV. Triage in medicine, part II: Underlying values and principles. Annals of Emergency Medicine. 2007;49:282-287.

4 Travers DA, Waller AE, Bowling JM, Flowers D, Tintinalli J. Five-level triage system more effective than three-level in tertiary emergency department. Journal of Emergency Nursing. 2002;28:395-400.

5 Tanabe P, Gimbel R, Yarnold PR, Adams JG. The Emergency Severity Index (version 3) 5-level triage system scores predict ED resource consumption. Journal of Emergency Nursing. 2004;30:22-29.

6 Elshove-Bolk J, Mencl F, van Rijswijck BT, Simons MP, van Vugt AB. Validation of the Emergency Severity Index (ESI) in self-referred patients in a European emergency department. Emergency Medicine Journal. 2007;24:170-174.

7 Australasian College for Emergency Medicine. Policy Document: The Australasian Triage Scale. 2000. [URL: http://www.medeserv.com.au/acem/open/documents/triage.htm].

8 Beveridge R, Clarke B, Janes L, Savage N, Thompson J, Dodd G, et al. Implementation guidelines for the Canadian Emergency Department Triage & Acuity Scale (CTAS). Canadian Association of Emergency Physicians. 1998.

9 Gilboy N, Tanabe P, Travers D, Rosenau A, Eitel DR. Emergency Severity Index, Version 4: Implementation Handbook. Rockville: Agency for Healthcare Research and Quality; 2005.

10 Mackway-Jones K. Emergency Triage. 2 ed. London: Manchester Triage Group; 2006.

11 Drijver R, Jochems P, Herrmann G, in 't Veld K, ten Wolde W. Netherlands Triage System: towards uniform triage; 2006.

12 Fernandes CB, Groth SJ, Johnson LA, Rosenau AM, Sumner J.A., Begley D, et al. A uniform triage scale in emergency medicine. American College of Emergency Physicians; 1999.

13 LeVasseur S, Charles A, Considine J, Berry D, Orchard D, Woiwod M, et al. Consistency of triage in Victoria's emergency departments. Monash Institute of Health Services Research; 2001.

14 Wuerz RC, Fernandes CM, Alarcon J. Inconsistency of emergency department triage. Emergency Department Operations Research Working Group. Annals of Emergency Medicine. 1998;32:431-435.

15 Beveridge R. The Canadian Triage and Acuity Scale: a new and critical element in health care reform. The Journal of Emergency Medicine. 1998;16:507-511.

16 Coenen IGA, Hagemeijer A, de Caluwé R, de Voeght FJ, Jochems PJJ. Guideline Triage on the emergency department. Dutch Institute for Healthcare Improvement; 2004.

17 HAN University of applied sciences, the Dutch Emergency Nurses Association (NVSHV), Netherlands Centre of Excellence in Nursing (LEVV). Guideline triage on the emergency department revision; 2008.

18 Schrijvers AJP, Drijver R, de Kam I, Ravestein J, Smit-Kam M, van Velzen E, et al. Atlas en scenario's voor spoedzorg. Utrecht: Igitur, Utrecht Publishing & Archiving Services; 2008.

19 Streiner DL, Norman GR. Health Measurement Scales. 3 ed. Oxfort University Press; 2007.

20 Bowling A. Techniques of questionnaire design. In: Bowling A, Ebrahim S, editors. Handbook of Health Research Methods. 1 ed. Berkshire: Open University Press; 2005. p. 394-427.

21 Cohen J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement. 1960;20:37-46.

22 Baumann MR, Strout TD. Evaluation of the Emergency Severity Index (version 3) triage algorithm in pediatric patients. Academic Emergency Medicine. 2005;12:219-224.

23 Bergeron S, Gouin S, Bailey B, Amre DK, Patel H. Agreement among pediatric health care professionals with the pediatric Canadian triage and acuity scale guidelines. Pediatric Emergency Care. 2004;20:514-518.

24 Considine J, Ung L, Thomas S. Triage nurses' decisions using the National Triage Scale for Australian emergency departments. Accident and Emergency Nursing. 2000;8:201-209.

25 Considine J, LeVasseur SA, Villanueva E. The Australasian Triage Scale: examining emergency department nurses' performance using computer and paper scenarios. Annals of Emergency Medicine. 2004;44:516-523.

26 Goransson K, Ehrenberg A, Marklund B, Ehnfors M. Accuracy and concordance of nurses in emergency department triage. Scandinavian Journal of Caring Sciences. 2005;19:432-438.

27 Gravel J, Gouin S, Bailey B, Roy M, Bergeron S, Amre D. Reliability of a computerized version of the Pediatric Canadian Triage and Acuity Scale. Acadamic Emergency Medicine. 2007;14:864-869.

28 Maningas PA, Hime DA, Parker DE, McMurry TA. The Soterion Rapid Triage System: evaluation of inter-rater reliability and validity. Journal of Emergency Medicine. 2006;30:461-469.

29 Tanabe P, Gimbel R, Yarnold PR, Kyriacou DN, Adams JG. Reliability and validity of scores on the Emergency Severity Index version 3. Academic Emergency Medicine. 2004;11:59-65.

30 Wuerz RC, Travers D, Gilboy N, Eitel DR, Rosenau A, Yazhari R. Implementation and refinement of the emergency severity index. Academic Emergency Medicine. 2001;8:170-176.

31 Bergeron S, Gouin S, Bailey B, Patel H. Comparison of triage assessments among pediatric registered nurses and pediatric emergency physicians. Academic Emergency Medicine. 2002;9:1397-1401.

32 Considine J, Ung L, Thomas S. Clinical decisions using the National Triage Scale: how important is postgraduate education? Accident and Emergency Nursing. 2001;9:101-108.

33 Goransson KE, Ehrenberg A, Marklund B, Ehnfors M. Emergency department triage: is there a link between nurses' personal characteristics and accuracy in triage decisions? Accident and Emergency Nursing. 2006;14:83-88.

34 Fernandes CM, Tanabe P, Gilboy N, Johnson LA, McNair RS, Rosenau AM, et al. Five-level triage: a report from the ACEP/ENA Five-level Triage Task Force. Journal of Emergency Nursing. 2005;31:39-50.

35 Baumann MR, Strout TD. Triage of geriatric patients in the emergency department: validity and survival with the Emergency Severity Index. Annals of Emergency Medicine. 2007;49:234-240.

36 Eitel DR, Travers DA, Rosenau AM, Gilboy N, Wuerz RC. The emergency severity index triage algorithm version 2 is reliable and valid. Academic Emergency Medicine. 2003;10:1070-1080.

37 van Gerven R, Delooz H, Sermeus W. Systematic triage in the emergency department using the Australian National Triage Scale: a pilot project. European Journal of Emergency Medicine. 2001;8:3-7.

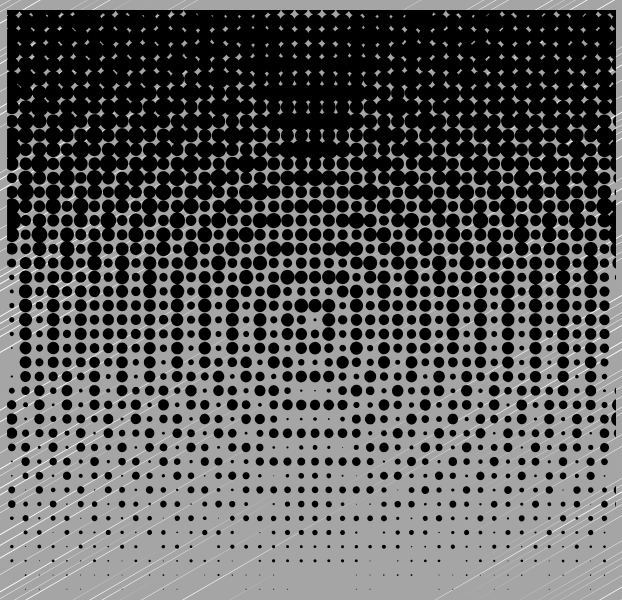
38 Roukema J, Steyerberg EW, van Meurs A, Ruige M, van der Lei J, Moll HA. Validity of the Manchester Triage System in paediatric emergency care. Emergency Medicine Journal. 2006;23:906-910.

39 Wuerz RC, Milne LW, Eitel DR, Travers D, Gilboy N. Reliability and validity of a new five-level triage instrument. Academic Emergency Medicine. 2000;7:236-242.

40 de Vet HC, Terwee CB, Knol DL, Bouter LM. When to use agreement versus reliability measures. Journal of Clinical Epidemiology. 2006;59:1033-1039

Part 1

Reliability of emergency department triage systems



2

Reliability and validity of the Manchester Triage System

Based on:
van der Wulp I, van Baar M.E, Schrijvers A.J.P. Reliability and validity of the Manchester Triage System in a general emergency department patient population in the Netherlands: results of a simulation study. Emergency Medicine Journal. 2008;25:431-434.

ABSTRACT

Objective: To assess the reliability and validity of the Manchester Triage System (MTS) in a general emergency department (ED) patient population.

Methods: A prospective evaluation study was conducted in two general hospitals in the Netherlands. ED nurses from both hospitals triaged fifty patient vignettes into one of five triage categories of the MTS. Triage ratings were compared with the ratings of two Dutch MTS experts to measure the consistency. Nineteen days after triaging the patient vignettes, triage nurses were asked to rate the same patient vignettes again, to measure the reproducibility. Reliability in relation to work experience of ED nurses was also studied. Validity was assessed by calculating percentages overtriage, undertriage, sensitivity and specificity.

Results: The consistency of the MTS was 'substantial' (quadratically weighted kappa: 0.62 (95% CI: 0.60 to 0.65)) and the reproducibility was high, (ICC: 0.75 (95% CI: 0.72 to 0.77)). No significant association was found between the number of years of experience of ED nurses with the MTS and the height of kappa. Undertriage occurred more frequently than overtriage, especially in elderly patients (25.3% versus 7.6%). Sensitivity for urgent patients in the MTS was 53.2% and specificity 95.1%. The patient vignettes representing children aged below sixteen years revealed a higher sensitivity (83.3%).

Conclusions: The consistency of the MTS is 'moderate' to 'substantial' and the reproducibility is high. The reliability of the MTS is not influenced by nurses' work experience. Undertriage mainly occurs in the MTS categories orange and yellow. The MTS is more sensitive for children who need immediate or urgent care than for other patients in the ED.

INTRODUCTION

Triage systems are often used in emergency departments (EDs). The need for triage systems originates in an increased demand for systemic working and increased workload[1]. The Manchester Triage System (MTS) has been implemented in several European countries[2]. In the Netherlands, most EDs have implemented the MTS in recent years[3]. The MTS, developed in Manchester (UK) in 1994, consists of fifty-two flowcharts. Each flowchart represents a specific complaint which can be subdivided into five categories: illness, lesion, children, abnormal behaviour and major incidents. A flowchart describes discriminators by which it is possible to assess the acuity of the patient's problem. Five categories can be distinguished: red (immediately), orange (can wait ten minutes), yellow (can wait sixty minutes), green (can wait two hours) and blue (can wait four hours)[1].
Because only a limited number of studies are available on the reliability and validity of the MTS, it is not known whether patients are correctly triaged when this system is used. From a patient safety point of view, this situation is not acceptable. Goodacre et al.[4] assessed agreement by means of interrater reliability in four emergency physicians who rated fifty randomly selected patient cases in three audit rounds. Agreement was found to be 'moderate'. Unfortunately this study measured agreement between physicians while triage is usually performed by nurses. A literature search found no studies which studied the relationship between work experience of ED nurses and reliability of the MTS.
The validity of triage systems is usually measured in terms of sensitivity, specificity and percentages of overtriage and undertriage (criterion validity). Four papers were found that described the results of studies to the validity of the MTS. Cooke et al[5]. studied a population of patients who needed admission to critical care areas. They concluded that the MTS was a sensitive tool. Speake et al[6]. measured the sensitivity and specificity of the MTS for identifying patients with high risk cardiac chest pain. Seven research workers reassessed patients with chest pain who visited the ED in a four-week period. The sensitivity was 86.8% and the specificity was 72.4%. Dann et al[7]. studied whether the MTS can detect critically ill patients. The sensitivity for allocating patients with mild pain into a nonurgent triage category was 14.9% while the specificity was 97.7%. Roukema et al[2]. studied the correlation between triage category and resource utilization and hospitalization in children. In addition, they compared triage categories with a predefined gold standard. All of these studies have focussed on specific patient subgroups. No studies were found that assessed the validity of the MTS in the general ED patient population.
The objective of our study was to prospectively assess the reliability and validity of the MTS in a general ED patient population in the Netherlands. The following study questions were therefore created:

- What is the consistency and reproducibility of the MTS?
- What is the association (if any) between nurses' work experience and reliability?
- How much over- and undertriage occurs when the MTS is used?
- What is the sensitivity and specificity of the MTS?

METHODS

Study design

The present study is a prospective evaluation of the MTS. The protocol was reviewed by the medical ethical committee of the University Medical Center Utrecht and all the participating hospitals.

Study setting and population

The study was conducted in two general hospitals located in the province of Utrecht. These hospitals implemented the MTS in 2005. The first hospital has an annual number of 36000 ED patient presentations with thirty-five nurses performing triage. The second hospital has an annual number of 23000 ED patient presentations with twenty nurses performing triage. All triage nurses had been instructed on how to use the MTS before triaging patients in their ED.

Data collection

Reliability and validity were prospectively measured by using fifty ED patient vignettes. These vignettes were generated from a random sample of ED patients from a general hospital, not participating in the study (Sint Elisabeth Hospital, Tilburg). The patient vignettes contained the following information: gender, age, time of arrival at the ED, complaint, transport to the ED, referrer and vital signs (table 1).

The triage nurses of both hospitals were asked to assign a triage category to every patient vignette using the MTS. They were allowed to use the same sources as they do when performing triage in their own ED (i.e. a computer program). Nineteen days later they were asked to complete the same patient vignettes again to assess reproducibility. These scores were compared with the triage category assigned by two Dutch MTS experts (reference standard). Each expert independently rated the vignettes. The ratings were then compared and discussed in case of discrepancy. The vignettes were rated by using the Dutch version of the book by Mackway-Jones et al.[1]. The experts are members of the Dutch Trauma Nursing Foundation (STNN) and work as ED nurses in two hospitals in the Netherlands. They translated and introduced the MTS into the Netherlands and teach nurses all over the country how to triage using the system. Both experts are certified MTS instructors.

Information was also collected on nurses' age, gender, and work experience as an ED nurse.

Table 1. Example of a patient vignette

Patient	Complaint	Arrival/referrer	Vital signs	Triage acuity
Gender: male Date of birth: 01.01.1983 Time: 03:39 a.m.	Experiences abdominal pain (right). Was examined in the previous week by an internist for the same complaint. Painkillers do not help.	Arrived with own transport. Self referral.	Temp: 37.5 °C HR: 85/min. RR: 122/79 mmHg Pain score: 5	□ Red □ Orange □ Yellow □ Green □ Blue

Data analysis

In this study, the reliability of the MTS is studied by means of consistency and reproducibility. The consistency was expressed in terms of interrater agreement between nurses and experts. Both, weighted and unweighted kappas (range: -1 to 1) were calculated. Former studies on the reliability of triage systems calculated unweighted kappa as a measure of interrater agreement[8;9]. Interrater agreement for ordinal scales (such as the MTS) is preferably measured by a weighted kappa statistic. There are different methods for calculating weighted kappa; we chose the most commonly used, quadratically weighted kappa[10]. The classification model designed by Landis and Koch[11] was used for the interpretation of kappa (table 2). These analyses were performed using Agree for Windows, a computer program for calculating a kappa statistic in numerous situations.

Reproducibility was calculated by the Intraclass Correlation Coefficient (ICC) (range: 0-1). The association between nurses' work experience and the interrater agreement score was calculated by Spearman's rho correlation coefficient. Validity was assessed by calculating sensitivity, specificity and percentages over- and undertriage. Overtriage was defined as the percentage of patient vignettes that received a more urgent triage category rating from the ED nurse than from the expert. Undertriage is the opposite of overtriage. These analyses were performed using SPSS for Windows Version 14.

Table 2. Interpretation of kappa

Kappa	Interpretation
< 0.00	Poor
0.00 - 0.20	Slight
0.21 - 0.40	Fair
0.41 - 0.60	Moderate
0.61 - 0.80	Substantial
0.81 – 1.00	Almost perfect

RESULTS

Response

Of the thirty-five nurses from hospital one, twenty-eight returned the vignettes in part one of the study and seventeen in part two. The overall response rate was 64.3%. No significant differences were found in the characteristics of the nurses between the responding and non responding groups. All twenty nurses from hospital two returned the vignettes in part one of the study and seventeen in part two. The overall response rate was 92.5%. Table 3 shows the characteristics of the respondents from both hospitals.

Table 3. Characteristics of participating nurses

Characteristic	Hospital 1	Hospital 2
Mean age (yrs)	46.9	41.3
% female	80.0	57.9
Mean number of years experience as nurse	26.4	19.8
Mean number of years experience as ED nurse	15.2	7.4

Reliability

Unweighted kappa was 0.48 (95% CI: 0.45 to 0.50), which represents 'moderate' agreement between nurses and the experts. Quadratically weighted kappa was 0.62 (95% CI: 0.60 to 0.65), which represents 'substantial' agreement. The ICC between nurses who participated in the two parts of the study was 0.75 (95% CI: 0.72 to 0.77), which indicates that nurses are consistent in triaging patients.

Nursing experience

Analyses of the ED nurses' work experience by outcome (kappa statistic) showed no significant association. There was a small negative, non-significant correlation for ED nurses' work experience (Spearman's rho = -0.22, p=0.141). Similar results were found for general nursing experience (Spearman's rho = -0.12, p=0.419).

Validity

Almost one-third (32.9%) of the patient vignettes were rated differently from the ratings from the experts, of which 7.6% showed overtriage and 25.3% undertriage (table 4). Most patient vignettes triaged one level above or under the triage rating of the experts. Undertriage mainly occurred when the experts triaged a patient vignette into the orange category while nurses triaged the vignette into the yellow category (32% of all undertriage). Similarly, experts triaged patient vignettes into the yellow category where nurses triaged it mainly into the green category (57% of all undertriage). Only 2.1% of the patient vignettes were triaged two or more levels above or below the triage rating of the expert. The sensitivity for MTS categories red and orange was moderate (53.2%) and the specificity for MTS categories yellow, green and blue was high (95.1%, table 4).

Eleven of the fifty patient vignettes concerned children below the age of sixteen years. A closer look at these vignettes reveals 8.2% overtriage and 24.7% undertriage. Undertriage in children mainly occurred in the yellow category where nurses triaged the patient into the green category. The sensitivity for MTS categories red and orange in children was 83.3% while specificity was 93.7%.

Six patient vignettes concerned patients above the age of sixty-five. Undertriage occurred more frequently than overtriage (43.3% and 5.0%, respectively). Sensitivity and specificity were comparable with the general ED patient population (52.5% and 97.5%, respectively).

Table 4: Percentages of agreement between the ratings of nurses and experts

Nurses ↓ Experts →	Red	Orange	Yellow	Green	Blue	Total
Red	0	0.8	0.3	0	0	1.1
Orange	0	9.8	3.4	0.2	0	13.4
Yellow	0	8.1	35.3	2.8	0	46.2
Green	0	1.2	14.4	22.0	0	37.6
Blue	0	0	0.5	1.1	0	1.6
Total	0	19.9	53.9	26.1	0	100

DISCUSSION

In this study, the reliability and validity of the MTS in a general ED population were studied by means of consistency, reproducibility and criterion validity. The consistency was shown to be 'moderate' to 'substantial' and the reproducibility was high. No significant associations were found between the number of years of (ED) work experience and consistency. Undertriage occurred frequently, especially in the MTS categories orange and yellow. Patient vignettes concerning children aged below sixteen years showed higher sensitivity. A focus on elderly patients over the age of sixty-five revealed considerable differences in undertriage compared with the general ED patient population.

Our study has some limitations and our results should be interpreted with care. Brenner et al[12]. concluded that the number of scale categories influences the kappa statistic. Unweighted kappa decreases when the number of categories increases, while (quadratically) weighted kappa increases. Furthermore, the marginal distribution in the contingency table has an influence on the kappa statistic[13].

We assessed the reproducibility by using a time period of nineteen days. It is possible that, after this time, nurses could still remember some of their former ratings. Nevertheless, the period between the first and second part of the study was based on recommendations by Streiner and Norman[14] who stated that a retest interval of 2-14 days is usual. No

associations between work experience and consistency were found. This might have been due to the small number of nurses who participated in the study. A larger population with a broader range of experience should be included in future studies.

One limitation of validity studies for triage systems is the lack of a gold standard. However, two solutions for this problem are available by either using an ED database or MTS experts. We chose to cooperate with two MTS experts. Their ratings were based on expert consensus. It is possible that some vignettes would be rated slightly different if different experts performed the rating. However, owing to their considerable theoretical and practical experience in using the MTS, we do not expect substantial deviations from current ratings. This would therefore only have a marginal influence on the results.

We refrained from using the patient's condition as a reference standard because we did not have access to the medical files of the patients and, in addition, no information on the real clinical decisions and treatments given to the patients from the vignettes could be obtained. It was therefore decided to use expert opinions as a reference standard. We hope to be able to include the patient's clinical condition as a reference standard in future research.

Percentages over- and undertriage were used as an indicator of validity. This has been used in previous triage research[2;15]. The MTS evaluates the maximum waiting time for a patient to see a doctor based on their complaint. When the waiting time is structurally too high or low, the system is not working correctly. Although the clinical condition of an individual patient is preferred for determining over- and undertriage, we think the use of experts is the second best solution. The percentages over- and undertriage may have been influenced by the limited number of patient vignettes (n=50) and because paper-based vignettes lead to more undertriage than real triage[16].

A comparison of our results with former studies shows some interesting differences. Goodacre et al.[4] found similar results compared to the present study ($0.31<\kappa<0.63$). However, it is not clear whether they calculated a weighted or unweighted kappa.

Our results concerning the high prevalence of undertriage are not in line with the findings of Roukema et al.[2] who reported 40% overtriage and 15% undertriage. The sensitivity and specificity also differed, with a sensitivity for detecting MTS categories red and orange of 63% and specificity of 78% while the present study showed sensitivity of 53% and specificity of 95%. These differences could possibly be explained by the fact that they studied the MTS in a children's hospital and used a different reference standard (ED database). The number of studies that can be compared with the present study is limited because other studies focused on specific diagnostic groups of patients while we studied a general ED population.

In conclusion, the consistency of the MTS is moderate to substantial and the reproducibility is high. There is no association between the amount of nurses' work experience and the use of the MTS. Undertriage (especially in elderly patients) in the MTS categories orange and yellow is a problem. The MTS shows moderate sensitivity for recognising patients who need immediate or urgent care. In children, the MTS is more sensitive.

REFERENCES

1 Manchester Triage Group. Emergency Triage. 1 ed. London: BMJ Publishing Group; 1997.

2 Roukema J, Steyerberg EW, van Meurs A, Ruige M, van der Lei J, Moll HA. Validity of the Manchester Triage System in paediatric emergency care. Emergency Medicine Journal. 2006;23:906-910.

3 van Baar ME, Giesen P, Grol R, Schrijvers AJP. Reporting results of a preliminary study to Emergency Medicine. Julius Center for Health Sciences and Primary Care & Center for Quality of Care Research WOK; 2007.

4 Goodacre SW, Gillett M, Harris RD, Houlihan KP. Consistency of retrospective triage decisions as a standardised instrument for audit. Journal of Accident and Emergency Medicine. 1999;16:322-324.

5 Cooke MW, Jinks S. Does the Manchester triage system detect the critically ill? Journal of Accident and Emergency Medicine. 1999;16:179-181.

6 Speake D, Teece S, Mackway-Jones K. Detecting high-risk patients with chest pain. Emergency Nurse. 2003;11:19-21.

7 Dann E, Jackson R, Mackway-Jones K. Appropriate categorisation of mild pain at triage: a diagnostic study. Emergency Nurse. 2005;13:28-32.

8 Brillman JC, Doezema D, Tandberg D, Sklar DP, Davis KD, Simms S, et al. Triage: limitations in predicting need for emergent care and hospital admission. Annals of Emergency Medicine. 1996;27:493-500.

9 Wuerz RC, Fernandes CM, Alarcon J. Inconsistency of emergency department triage. Emergency Department Operations Research Working Group. Annals of Emergency Medicine. 1998;32:431-435.

10 Graham P, Jackson R. The analysis of ordinal agreement data: beyond weighted kappa. Journal of Clinical Epidemiology. 1993;46:1055-1062.

11 Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33:159-174.

12 Brenner H, Kliebsch U. Dependence of weighted kappa coefficients on the number of categories. Epidemiology. 1996;7:199-202.

13 Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. Journal of Clinical Epidemiology. 1990;43:543-549.

14 Streiner DL, Norman GR. Health Measurement Scales: A practical guide to their development and use. 3 ed. Oxfort University Press; 2007.

15 Travers DA, Waller AE, Bowling JM, Flowers D, Tintinalli J. Five-level triage system more effective than three-level in tertiary emergency department. Journal of Emergency Nursing. 2002;28:395-400.

16 Worster A, Sardo A, Eva K, Fernandes CM, Upadhye S. Triage Tool Inter-rater Reliability: A Comparison of Live Versus Paper Case Scenarios. Journal of Emergency Nursing. 2007;33:319-323.

An alternative kappa weighting scheme which accounts for severity of mistriage

Based on:
van der Wulp I, van Stel H.F. Adjusting weighted kappa for the severity of mistriage decreases reported reliability of emergency department triage systems: a comparative study. Journal of Clinical Epidemiology. 2009;62:1196-1201.

ABSTRACT

Objective: Mistriage can have serious consequences for patients with urgent complaints. We therefore reviewed the assessment of triage reliability, and propose an alternative weighted kappa that accounts for severity of mistriage.

Methods: A systematic literature search was performed in three online databases and a search engine. An alternative kappa weighting scheme was developed (triage weighted kappa), and kappas of previous conducted reliability studies were recalculated.

Results: Kappa is the most frequently used statistic in triage reliability studies (n=33). However, more than half of the studies did not report the type of kappa that was used. Linear and quadratically weighted kappa values do not reflect the seriousness of mistriage. Several studies reported almost perfect agreement, whereas percentages of mistriage ranged between 11.1% and 43.4%. In all studies, triage weighted kappa was smaller than the reported kappas with a mean difference of 0.17 (range: 0.04 to 0.32).

Conclusions: No existing studies on reliability of triage systems account for mistriage. Using triage weighted kappa, which reflects the severity of mistriage, shows that the reliability of triage systems is smaller than previously reported.

INTRODUCTION

Triage is defined as: "*the initial assessment and sorting of patients in an emergency setting to determine clinical priority of need and, in some emergency departments, appropriate area for treatment*"[1]. Several emergency department (ED) triage systems have been developed in recent years for example, the Manchester Triage System, the Emergency Severity Index, the Canadian Triage and Acuity Scale and the Australasian Triage Scale[2]. The accuracy of a triage decision is very important. In case of overtriage, which is an overestimation of urgency, a truly urgent patient may have to wait because resources, such as a bed, are scarce. In case of undertriage, which is an underestimation of urgency, a patient has to wait though at risk for deterioration[3]. From a patient safety point of view, it therefore is necessary to assess the performance of triage systems in the ED for example, by measuring reliability.
Kappa is a measure that is frequently used in reliability studies. It is a measurement of observer agreement, corrected for agreement expected by chance. Kappa ranges between -1, representing perfect disagreement and 1, representing perfect agreement[4]. Two types of kappa are frequently used: unweighted- and weighted kappa. Unweighted kappa is used for measuring exact agreement in nominal scales. Weighted kappa also measures partial agreement and is used for ordinal scales. Weights can be calculated with a linear or quadratic algorithm, or an own set of weights[4;5].
As triage systems are ordinal scales, weighted kappa seems the correct measure for evaluating reliability. Moreover, weights used for the calculation of weighted kappa in triage reliability studies should take into account the severity of mistriage. To sort out which kappas were used in triage reliability studies, we conducted a literature review. Furthermore, we studied which type of kappa is suitable for use in triage reliability studies and developed an alternative kappa weighting scheme. From previously conducted triage reliability studies, kappas were recalculated using this alternative weighting scheme.

METHODS

Literature search

A literature search was performed in three online databases: Pubmed, Embase and Cochrane. Furthermore, we used Google Scholar[6]. Table 1 presents the search terms that were used and the number of hits. Studies were included when they reported about the reliability of a triage system in the ED, and provided an abstract in English or Dutch. Articles and abstracts not providing statistic measures of agreement were excluded from the study. Finally, reference lists of included studies were reviewed for additional relevant articles.

Table 1. Search terms used in online databases and Google Scholar

Database	Search terms (hits)	Results
Pubmed	"Triage (Mesh)" AND "Emergency Service Hospital (Mesh)" (1060)	23
	"Triage (Mesh)" AND "Validation studies (publication type)" (69)	0
	"Triage (Mesh)" AND "Reliability" (64)	5
	"Triage (Mesh)" AND "Emergency Service Hospital (Mesh)" (AND) "Validation studies (publication type)" (29)	0
	"Emergency Severity Index" (21)	1
	"Manchester Triage System" (6)	1
	"Canadian Triage and Acuity Scale" (31)	2
	"Australasian Triage and Acuity Scale" (10)	0
Cochrane	"Triage" AND "Emergency Department" (163)	0
	"Triage" AND "Emergency Department" (AND) "Reliability" (12)	0
	"Emergency Severity Index" (1)	0
	"Manchester Triage System" (1)	0
	"Canadian Triage and Acuity Scale" (1)	0
	"Australasian Triage Scale" (1)	0
Embase	"Triage" AND "Emergency ward (Emtree)" (1164)	4
	"Triage" AND "Reliability (Emtree)" (100)	0
	"Triage" AND "Validation study (Emtree)" (24)	0
	"Triage" AND "Emergency ward (Emtree)" AND "Reliability (Emtree)" (28)	0
	"Emergency Severity Index" (22)	1
	"Manchester Triage System" (8)	0
	"Canadian Triage and Acuity Scale" (19)	0
	"Australasian Triage Scale" (25)	0
Google Scholar	"triage" AND "emergency department" AND "interrater reliability" (571)	2
Hand search		2
Total		41

Kappa coefficient of agreement

To calculate a weighted kappa, two algorithms or an own set of weights can be applied. The algorithms for linear and quadratically weighted kappas are: $Wij = 1-\{|i - j| / |c - 1|\}$ and $Wij = 1- \{(i - j)^2/ (c – 1)^2\}$, respectively. In these algorithms, 'c' represents the number of categories a scale consists of, 'i' represents the category rated by rater 'x' and 'j' represents the category rated by rater 'y'(7). In tables 2a and 2b, the weights for both algorithms were calculated. These weights are based on five-level triage systems, such as the Emergency Severity Index, the Manchester Triage System, the Canadian Triage and Acuity Scale and the Australasian Triage Scale. The weights are symmetrically distributed, with the larger the distance from perfect agreement, the smaller the weight. However, in these weighting schemes, substantial disagreement will be rewarded with a high kappa value. For example, a patient who is triaged as nonurgent, but in fact, needs care immediately, can get seriously ill or even die. This is not reflected in the weighting schemes: a disagreement of two levels still counts as 75% of exact agreement. For clinical practice, different weights should be considered (quadratic weights). Furthermore, a second limitation is that in both weighting schemes, because of the distribution of weights, agreement will be higher when raters constantly rate in the middle triage category.

In our opinion, a weighting scheme, suitable for triage reliability studies, must meet the criterion that the weights fit to what is clinically more acceptable in triage practice than current weighting schemes. No study has previously adjusted the weights in triage

reliability studies. We therefore designed an alternative weighting scheme (table 3). This weighting scheme is based on two principles. First, undertriage can cause unnecessary health loss in patients. Second, in case of overtriage, the waste of ED resources must be prevented, because this can also cause prolonged waiting times for truly urgent patients. Therefore several cells were set to zero. In the least urgent triage category, two levels of overtriage were allowed because overtriage of the least urgent patients is not likely to have a major impact on the speed of treatment onset for patients in the urgent triage categories.
An alternative algorithm for calculating weights in the remaining cells was designed. The algorithm is based on the algorithm used to calculate quadratic weights. We added an extra weight into the existing algorithm. As, in our opinion, undertriage is more serious than overtriage, lower weights were given to undertriage. To keep the weights between zero and one, two algorithms were created. In case of overtriage, weights are calculated with the algorithm: $Wij= (1-((i-j)^2/ (c-1)^2)) / (i/j)$. In case of undertriage, a square to the algorithm was added to create the smaller weights: $Wij= (1-((i-j)^2/ (c-1)^2))* (i/j)^2$. In both algorithms, 'i' is defined as the category assigned by a reference standard and 'j' as the category assigned by a rater. The calculated weights can be used in studies examining the reliability by means of consistency in five-level triage systems. In these systems, the first category represents the highest urgency. We denote this alternative weighting scheme 'triage weighted kappa'. The software program AGREE 7 for Windows was used to calculate (if not reported) quadratically weighted kappa and triage weighted kappa in previous studies[8]. For this analysis, only full text articles were included.

Table 2a. Linear weights for an ordinal scale with 5 categories

Y ↓ X →	1	2	3	4	5
1	1.00	0.75	0.50	0.25	0.00
2	0.75	1.00	0.75	0.50	0.25
3	0.50	0.75	1.00	0.75	0.50
4	0.25	0.50	0.75	1.00	0.75
5	0.00	0.25	0.50	0.75	1.00

Algorithm linear weighted kappa: $Wij=1-\{|i-j| / |c-1|\}$

Table 2b. Quadratic weights for an ordinal scale with 5 categories

Y ↓ X →	1	2	3	4	5
1	1.00	0.94	0.75	0.44	0.00
2	0.94	1.00	0.94	0.75	0.44
3	0.75	0.94	1.00	0.94	0.75
4	0.44	0.75	0.94	1.00	0.94
5	0.00	0.44	0.75	0.94	1.00

Algorithm quadratically weighted kappa: $Wij=1-\{(i-j)^2/ (c-1)^2\}$

Table 3. Triage weighted kappa

Ref stand. → Rater ↓	1	2	3	4	5
1	1.00	0.47	0.00	0.00	0.00
2	0.23	1.00	0.63	0.00	0.00
3	0.00	0.42	1.00	0.70	0.45
4	0.00	0.00	0.53	1.00	0.75
5	0.00	0.00	0.00	0.60	1.00

Overtriage: Wij= (1-((i-j)²/(c-1)²))/(i/j)
Undertriage: Wij= (1-((i-j)²/(c-1)²))(i/j)²*
Ref stand.= reference standard

RESULTS

The literature search resulted in forty-one eligible articles of triage reliability studies published between 1975 and 2008. One abstract was included that provided statistical measures of agreement (table 1). Most studies were conducted in the United States (n=18) and Canada (n=12). In thirty studies, the interrater reliability of a triage system between different raters was measured (agreement), and in twenty-three studies, a reference standard was used (consistency). Eleven studies did both. Three methods were mainly used to measure triage reliability: patient scenarios (n=23), triage of real patient presentations in the ED (n=14) and the review of triage notes or patients charts (n=11). In six studies, more than one method was applied.

Reliability reported in literature

Kappa appeared to be the most frequently reported measure in triage reliability studies (n=33). Seven studies reported unweighted kappa of which six also reported a weighted kappa[9-15]. In total, weighted kappa was reported in twenty-two studies[1;9-13;15-30]. One study was found that reported, besides a weighted kappa also the prevalence and bias adjusted kappa (PABAK)[21]. This kappa is adjusted for prevalence and bias by substituting actual values in the cells of a contingency table with average values[31]. In ten studies, information on the type of kappa used (weighted or unweighted) was missing[32-41]. From the presented contingency tables it was concluded that three of these studies reported an unweighted kappa[32;33;40].

From the studies reporting a weighted kappa (n=22), eight reported a quadratically weighted kappa[9-11;15;19;21;25;28], of which two also reported a linear weighted kappa[10;11]. Fourteen studies did not report the type of weighted kappa that was used[1;12;13;16-18;20;22-24;26;27;29;30]. Recalculation of kappa by using the reported contingency tables revealed that five studies reported a linear weighted kappa[13;16;24;29;30] and two studies reported a quadratically weighted kappa[1;22]. In seven studies, the type of weighted kappa could not be determined[12;17;18;20;23;26;27]. No studies were found that used an own set of weights to calculate weighted kappa.

Besides kappa, other reported measures for interrater reliability and test-retest reliability included: the percentage absolute agreement between raters[42-47], the intraclass correlation coefficient[15;48;49], Kendall's correlation[41] and Relative to an Identified Distribution (RIDIT) analysis[50], which is a probability measure. Two studies which reported one of the before mentioned reliability measures, also reported a kappa[15;41].

Kappa recalculated

Eleven articles reported a contingency table with frequencies of ratings from a five-level triage system compared with a reference standard. If it was not presented, we calculated quadratically weighted kappa. Moreover, triage weighted kappa was calculated and the results are presented in table 4. Compared with quadratically weighted kappa, triage weighted kappa appeared to be smaller in all studies, ranging from 0.04 to 0.32 (mean difference=0.17). According to the classification model of Landis and Koch[51](table 5), the strength of agreement in thirteen out of fifteen reported quadratically weighted kappas decreased one level when triage weighted kappa was calculated. Of these, nine kappas decreased from almost perfect agreement to substantial agreement[1;13;14;16;18;24;33;44], three kappas decreased from substantial agreement to moderate agreement[14;15;33], and one kappa decreased from moderate agreement to fair agreement[40]. In two studies, triage weighted kappa decreased marginally (0.04 and 0.06)[1;24]. A closer look at these contingency tables revealed a substantial percentage of exact agreement, 79.7% and 88.9%, respectively. In these studies, mistriage mainly occurred within one category from perfect agreement, resulting in only a minimal decrease in kappa.

Another way of interpreting the difference between triage weighted kappa and quadratically weighted kappa is from a clinical point of view (table 5)[52]. Twelve out of fifteen reported quadratically weighted kappas decreased by one or more levels in the classification scheme. Eight of these fell from the category 'excellent' to the category 'good'[13;14;18;30;33;44], two from 'excellent' to 'fair'[14;33], one from 'good' to 'fair'[15], and one from 'fair' to 'poor'[40]. Three of the reported quadratically weighted kappas, remained in the same category after calculating triage weighted kappa[1;16;24].

Table 4 also presents percentages mistriage. In several studies, quadratically weighted kappa was high, between 0.81 and 0.91, whereas mistriage occurred frequently: between 11.1% and 43.4%. Triage weighted kappa seems to reflect this mistriage somewhat better as this kappa ranged between 0.61 and 0.87[1;14;16;18;33;40;44].

Table 4. Agreement with squared weighted kappa and triage weighted kappa

Study	Agreement	N ratings	Quadratic weighted (95% CI)	Triage weighted (95% CI)	% over triage	% under triage
Considine	Nurses/ experts	310	0.86 (0.84-0.89)	0.67 (0.61-0.72)	20.97	20.97
Wuerz	Nurses/ researcher	219	0.80 (0.75-0.87)	0.71 (0.62-0.80)	10.96	12.79
Considine	Nurses/ experts	1176	0.82 (0.80-0.84)	0.64 (0.60-0.67)	19.98	15.47
	Nurses/ experts	1133	0.89 (0.87-0.90)	0.74 (0.71-0.77)	13.95	18.18
	Nurses/ experts	1173	0.81 (0.79-0.82)	0.62 (0.59-0.65)	25.83	17.56
	Nurses/ experts	1132	0.80 (0.78-0.82)	0.58 (0.55-0.61)	22.88	22.88
Wollaston.	ATS*/ ATTT**	58	0.52 (0.37-0.67)	0.27 (0.11-0.42)	37.93	15.52
Tanabe	Nurses/ researcher	359	0.83 (0.78-0.88)	0.77 (0.71-0.82)	9.19	11.14
Bergeron	Nurses/ experts	1485	0.84 (0.82-0.85)	0.65 (0.62-0.68)	14.75	21.55
Baumann	Nurses/ researcher	20	0.91 (0.82-1.00)	0.76 (0.57-0.95)	5.00	15.00
Manningas	Nurses/ researcher	423	0.91 (0.88-0.94)	0.87 (0.83-0.91)	5.20	5.91
v/d Wulp	Nurses/ experts	2382	0.62 (0.60-0.65)	0.52 (0.49-0.55)	7.56	25.27
Göransson	Nurses/ experts	7550	0.86 (0.85-0.86)	0.65 (0.64-0.66)	28.42	13.93
Gravel.	Nurses/ experts	970	0.81 (0.79-0.83)	0.61 (0.58-0.64)	12.80	30.10
	Nurses/ experts	974	0.79 (0.77-0.82)	0.47 (0.42-0.51)	14.20	31.10

** ATS= Australasian Triage Scale*

*** TATTT= Toowoomba Adult Trauma Triage Tool*

Table 5. Interpretation of kappa

Kappa	Interpretation strength of agreement	Kappa	Interpretation clinical significance
< .00	Poor	< .40	Poor
.00 - .20	Slight	.40 - .59	Fair
.21 - .40	Fair	.60 - .74	Good
.41 - .60	Moderate	.75 – 1.00	Excellent
.61 - .80	Substantial		
.81 – 1.00	Almost perfect		

DISCUSSION

It was studied how triage reliability is reported in the literature. Most studies examined the reliability with the kappa statistic. However, a large number of studies did not report what type of kappa or weighted kappa was used. It was also found that the weights of linear and quadratically weighted kappa do not fit with triage practice as they reward mistriage too much. Moreover, triaging constantly in the middle triage category causes unfounded high agreement. Therefore, an alternative weighted kappa was designed and kappas from published studies were recalculated. With this triage weighted kappa, agreement in previous conducted studies decreased between 0.04 and 0.32. It therefore seemed more in accordance with the percentages reported mistriage, although percentages agreement must always be reported. With the development of this alternative weighting scheme, an attempt was made to create a measurement which is more in accordance with clinical practice. It is difficult to determine a weighting scheme that exactly reflects clinical practice. Further research should focus on improving our weighting scheme or on developing a method that generates clinically sensible weights. Furthermore, triage weighted kappa is designed for studies in which the reliability is studied by means of consistency. Studies measuring agreement still face the same problems with weighted kappa as described earlier. In these studies, unweighted kappa can be used, as mistriage cannot be assessed.

Interpretation of the kappa coefficient has to be done with care. First, the kappa value depends on the distribution of ratings. There was an instance where the percentage observed agreement between raters was high (80%) with a kappa of only 0.32[51;53]. Second, the number of categories on the scale being measured influenced the kappa value, especially when quadratic weights were used. Quadratically weighted kappa increased with the number of categories on the scale. An opposite effect was seen when linear weighted kappa was used[54]. Manos et al.[25] therefore stated that it was very difficult to compare the reliability of triage systems with various numbers of categories. Other measures, such as correlation coefficients, Yule's Y, and the odds ratio, were therefore suggested as options to measure reliability. However, Cicchetti et al.[55] concluded that those measures did not correct for agreement expected by chance, or if they did, got similar results as kappa. Triage weighted kappa does not solve these shortcomings, but currently is the only measure that is fitted to triage practice. Furthermore, these statistics do not allow for giving different weights according to the seriousness of mistriage. Triage weighted kappa was designed with this objective in mind.

No studies were found, in and outside the field of triage, which used weighted kappa with an own set of weights. A limitation of using an own set of weights is that it is more difficult to compare the results with other studies that used the standard weighting schemes.

However, when the standard weighting schemes do not reflect clinical practice, one is forced to create its own weights.
In conclusion, it is difficult to compare the results of previously conducted triage reliability studies because of the lack of information about the type of kappa that is used. Linear and quadratically weighted kappas were not suitable when measuring reliability of triage systems, because from a clinical point of view, disagreement is rewarded to a great extent. With the use of triage weighted kappa, an attempt was made to evaluate the reliability of triage systems in a way that represents triage practice. As a result, reliability of triage systems will be interpreted in accordance with clinical practice.

REFERENCES

1 Tanabe P, Gimbel R, Yarnold PR, Kyriacou DN, Adams JG. Reliability and validity of scores on The Emergency Severity Index version 3. Academic Emergency Medicine. 2004;11:59-65.

2 Fernandes CM, Tanabe P, Gilboy N, Johnson LA, McNair RS, Rosenau AM, et al. Five-level triage: a report from the ACEP/ENA Five-level Triage Task Force. Journal of Emergency Nursing. 2005;31:39-50.

3 Gilboy N, Tanabe P, Travers D, Rosenau A, Eitel DR. Emergency Severity Index, Version 4: Implementation Handbook. Rockville: Agency for Healthcare Research and Quality; 2005.

4 Cohen J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement. 1960;20:37-46.

5 Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychology Bulletin. 1968;70:213-220.

6 Google Scholar available at: http://scholar.google.com. 2008.

7 Fleiss JL, Levin B, Paik MC. The measurement of inter rater agreement. Statistical methods for rates and proportions. New York: Wiley; 2003.

8 AGREE for Windows [computer program]. Version 7.002 1999.

9 Dong SL, Bullard MJ, Meurer DP, Colman I, Blitz S, Holroyd BR, et al. Emergency triage: comparing a novel computer triage program with standard triage. Academic Emergency Medicine. 2005;12:502-507.

10 Dong SL, Bullard MJ, Meurer DP, Blitz S, Ohinmaa A, Holroyd BR, et al. Reliability of computerized emergency triage. Academic Emergency Medicine. 2006;13:269-275.

11 Dong SL, Bullard MJ, Meurer DP, Blitz S, Holroyd BR, Rowe BH. The effect of training on nurse agreement using an electronic triage system. Canadian Journal of Emergency Medicine. 2007;9:260-266.

12 Durani Y, Brecher D, Walmsley D, Attia MW, Loiselle JM. The Emergency Severity Index (Version 4): Reliability in pediatric patients. Academic Emergency Medicne. 2007;14(Suppl 1):S95-S96.

13 Goransson K, Ehrenberg A, Marklund B, Ehnfors M. Accuracy and concordance of nurses in emergency department triage. Scandinavian Journal of Caring Sciences. 2005;19:432-438.

14 Gravel J, Gouin S, Bailey B, Roy M, Bergeron S, Amre D. Reliability of a computerized version of the Pediatric Canadian Triage and Acuity Scale. Academic Emergency Medicine. 2007;14:864-869.

15 van der Wulp I, van Baar ME, Schrijvers AJP. Reliability and validity of the Manchester Triage System in a general emergency department patient population in the Netherlands: Results of a simulation study. Emergency Medicine Journal. 2008;25:431-434.

16 Baumann MR, Strout TD. Evaluation of the Emergency Severity Index (version 3) triage algorithm in pediatric patients. Academic Emergency Medicine. 2005;12:219-224.

17 Bergeron S, Gouin S, Bailey B, Patel H. Comparison of triage assessments among pediatric registered nurses and pediatric emergency physicians. Academic Emergency Medicine. 2002;9:1397-1401.

18 Bergeron S, Gouin S, Bailey B, Amre DK, Patel H. Agreement among pediatric health care professionals with the pediatric Canadian triage and acuity scale guidelines. Pediatric Emergency Care. 2004;20:514-518.

19 Beveridge R, Ducharme J, Janes L, Beaulieu S, Walter S. Reliability of the Canadian emergency department triage and acuity scale: interrater agreement. Annals of Emergency Medicine. 1999;34:155-159.

20 Eitel DR, Travers DA, Rosenau AM, Gilboy N, Wuerz RC. The emergency severity index triage algorithm version 2 is reliable and valid. Academic Emergency Medicine. 2003;10:1070-1080.

21 Grafstein E, Innes G, Westman J, Christenson J, Thorne A. Inter-rater reliability of a computerized presenting-complaint-linked triage system in an urban emergency department. Canadian Journal of Emergency Medicine. 2003;5:323-329.

22 Kim TG, Cho JK, Kim SH, Lee HS, Gu HD, Chung SW. Reliability and validity of the modified Emergency Severity Index-2 ad a triage tool. Journal of the Korean Society of Emergency Medicine. 2006;17:154-164.

23 Maningas PA, Hime DA, Parker DE. The use of the Soterion Rapid Triage System in children presenting to the Emergency Department. Journal of Emergency Medicine. 2006;31:353-359.

24 Maningas PA, Hime DA, Parker DE, McMurry TA. The Soterion Rapid Triage System: evaluation of inter-rater reliability and validity. Journal of Emergency Medicine. 2006;30:461-469.

25 Manos D, Petrie DA, Beveridge RC, Walter S, Ducharme J. Inter-observer agreement using the Canadian Emergency Department Triage and Acuity Scale. Canadian Journal of Emergency Medicine. 2002;4:16-22.

26 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. Journal of Clinical Epidemiology. 2006;59:615-621.

27 Travers DA, Waller AE, Bowling JM, Flowers D, Tintinalli J. Five-level triage system more effective than three-level in tertiary emergency department. Journal of Emergency Nursing. 2002;28:395-400.

28 Worster A, Gilboy N, Fernandes CM, Eitel D, Eva K, Geisler R, et al. Assessment of inter-observer reliability of two five-level triage and acuity scales: a randomized controlled trial. Canadian Journal of Emergency Medicine. 2004;6:240-245.

29 Wuerz RC, Milne LW, Eitel DR, Travers D, Gilboy N. Reliability and validity of a new five-level triage instrument. Academic Emergency Medicine. 2000;7:236-242.

30 Wuerz RC, Travers D, Gilboy N, Eitel DR, Rosenau A, Yazhari R. Implementation and refinement of the emergency severity index. Academic Emergency Medicine. 2001;8:170-176.

31 Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Physical Therapy. 2005;85:257-268.

32 Brillman JC, Doezema D, Tandberg D, Sklar DP, Davis KD, Simms S, et al. Triage: limitations in predicting need for emergent care and hospital admission. Annals of Emergency Medicine. 1996;27:493-500.

33 Considine J, LeVasseur SA, Villanueva E. The Australasian Triage Scale: examining emergency department nurses' performance using computer and paper scenarios. Annals of Emergency Medicine. 2004;44:516-523.

34 Gerdtz MF, Bucknall TK. Influence of task properties and subjectivity on consistency of triage: a simulation study. Journal of Advanced Nursing. 2007;58:180-190.

35 Gill JM, Reese CL, Diamond JJ. Disagreement among health care professionals about the urgent care needs of emergency department patients. Annals of Emergency Medicine. 1996;28:474-479.

36 Goodacre SW, Gillett M, Harris RD, Houlihan KP. Consistency of retrospective triage decisions as a standardised instrument for audit. British Medical Journal. 1999;16:322-324.

37 Hay E, Bekerman L, Rosenberg G, Peled R. Quality assurance of nurse triage: consistency of results over three years. American Journal of Emergency Medicine. 2001;19:113-117.

38 Maldonado T, Avner JR. Triage of the pediatric patient in the emergency department: are we all in agreement? Pediatrics. 2004;114:356-360.

39 Nakagawa J, Ouk S, Schwartz B, Schriger DL. Interobserver agreement in emergency department triage. Annals of Emergency Medicine. 2003;41:191-195.

40 Wollaston A, Fahey P, McKay M, Hegney D, Miller P, Wollaston J. Reliability and validity of the Toowoomba adult trauma triage tool: a Queensland, Australia study. Accident and Emergency Nursing. 2004;12:230-237.

41 Wuerz RC, Fernandes CM, Alarcon J. Inconsistency of emergency department triage. Emergency Department Operations Research Working Group. Annals of Emergency Medicine. 1998;32:431-435.

42 Albin SL, Wassertheil-Smoller S, Jacobson S, Bell B. Evaluation of emergency room triage performed by nurses. American Journal of Public Health. 1975;65:1063-1068.

43 Bazarian JJ, Stern RA, Wax P. Accuracy of ED triage of psychiatric patients. American Journal of Emergency Medicine. 2004;22:249-253.

44 Considine J, Ung L, Thomas S. Triage nurses' decisions using the National Triage Scale for Australian emergency departments. Accident and Emergency Nursing. 2000;8:201-209.

45 Cooke MW, Jinks S. Does the Manchester triage system detect the critically ill? Journal of Accident and Emergency Medicine. 1999;16:179-181.

46 Durojaiye L, O'Meara M. A study of triage of paediatric patients in Australia. Emergency Medicine (Fremantle). 2002;14:67-76.

47 Toulson K, Laskowski-Jones L, McConnell LA. Implementation of the five-level emergency severity index in a level I trauma center emergency department with a three-tiered triage scheme. Journal of Emergency Nursing. 2005;31:259-264.

48 Worster A, Sardo A, Eva K, Fernandes CM, Upadhye S. Triage Tool Inter-rater Reliability: A Comparison of Live Versus Paper Case Scenarios. Journal of Emergency Nursing. 2007;33:319-323.

49 Fernandes CM, Wuerz R, Clark S, Djurdjev O. How reliable is emergency department triage? Annals of Emergency Medicine. 1999;34:141-147.

50 van Gerven R, Delooz H, Sermeus W. Systematic triage in the emergency department using the Australian National Triage Scale: a pilot project. European Journal of Emergency Medicine. 2001;8:3-7.

51 Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33:159-174.

52 Cicchetti D, Bronen R, Spencer S, Haut S, Berg A, Oliver P, et al. Rating scales, scales of measurement, issues of reliability. Journal of Nervous and Mental Disease. 2006;194, 557-564.

53 Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. Journal of Clinical Epidemiology. 1990;43:543-549.

54 Brenner H, Kliebsch U. Dependence of weighted kappa coefficients on the number of categories. Epidemiology. 1996;7:199-202.

55 Cicchetti DV, Feinstein AR. High agreement but low kappa: II. Resolving the paradoxes. Journal of Clinical Epidemiology. 1990;43:551-558.

Interpreting the reliability of emergency department triage systems

Based on:
van der Wulp I, van Stel H.F. Calculating kappas from adjusted data improved the comparability of the reliability of triage systems: a comparative study.
Submitted after revision

ABSTRACT

Objective: It is difficult to compare the reliability of triage systems with the kappa statistic. In this paper, a method for comparing triage systems was developed and applied to previously conducted triage reliability studies.

Methods: From simulations with theoretical distributions, the minimum, normal, and maximum obtainable weighted kappas for three- to five-level triage systems were computed. To compare the reliability of triage systems in previously conducted triage reliability studies, the normal kappa was calculated. Furthermore, the reported quadratically weighted kappas were compared with the minimum, normal, and maximum weighted kappa to characterize the degree and direction of skewness of the data.

Results: The normal kappa was higher in three-level triage systems (median: κ=0.84) compared to four-level systems (median: κ=0.37) and five-level systems (median: κ=0.57). In three-level triage systems, the percentages observed agreement were unequally distributed which resulted in small quadratically weighted kappas. In four- and five-level triage systems the percentages observed agreement were more equally distributed compared to three-level triage systems which resulted in higher quadratically weighted kappa values.

Conclusions: When comparing triage systems with different numbers of categories one should report both, the normal kappa and quadratically weighted kappa. Calculating normal kappas from previously conducted triage reliability studies revealed substantial theoretical differences in interrater reliability of triage systems than previously reported.

INTRODUCTION

In recent years, several five-level emergency department (ED) triage systems such as the Emergency Severity Index (ESI), the Canadian Triage and Acuity Scale (CTAS), the Australasian Triage Scale (ATS) and the Manchester Triage System (MTS) have been developed and implemented in EDs[1]. The purpose of creating these systems was mainly to create uniformity in the triage of patients in the ED[2-5]. Before the implementation of five-level triage systems, the existing triage systems in the ED consisted of three or four categories[1]. Studies conducted to three- and four-level triage systems reported that interrater reliability of these systems was small with kappa values between 0.19 and 0.53[6-9]. Studies conducted to five-level triage systems reported higher reliability, kappa ranged between 0.38 and 1.00. In these studies it was concluded that five-level triage systems were more reliable than three- or four-level triage systems while the percentages disagreement were not taken into account. In one of these studies these percentages differed only by 1%[8;10-17].

It is difficult to compare kappas of triage systems. The kappa statistic is a measure in which agreement of two or more raters or methods is corrected for agreement due to chance. The chance correction in kappa depends on the distribution of ratings and the number of categories of the measurement scale[18-21]. These influences can be important when comparing systems, especially when these systems have different numbers of categories. Because triage reliability studies frequently included patients that were representative of ED casemix this resulted, in some studies, in unequal distributions of categories[22-28]. This can lead to substantial percentages agreement but small to moderate kappa values. Also the opposite situation could occur, high kappa values while percentages agreement are small[29;30]. Possibly this behaviour of the kappa statistic is more sensitive for systems with five levels than for systems with fewer categories. This is because more cells in the contingency table can be filled so that the data in these tables is skewed earlier in these systems compared to three- or four-level systems. Furthermore, when comparing triage systems with different numbers of categories, the kappa statistic is influenced by that the proportion of expected agreement in scales with for example five levels is generally smaller compared to three-level scales. As a result, the kappa values of five-level systems will be higher by definition. However, this difference will always exist because it is a characteristic of kappa. Finally, we have previously reported that heterogeneity in the use of kappa exists and that information about the type of kappa reported was missing in several studies[31]. In case it is unknown what type of kappa is used in studies which reported the reliability of triage systems, comparisons cannot be made between triage systems because weighted kappas are generally higher than unweighted kappa. The before mentioned, could have led to improper comparisons of the reliability of triage systems. The reliability of triage systems is often compared with existing triage systems to support decisions for the implementation of triage systems in the ED, or to revise triage guidelines. It is therefore important to account for the distribution of ratings when comparing the reliability of triage systems.

The objective of this study was to compare the reliability of three-, four-, and five-level ED triage systems by developing and applying a method which accounts for influences of kappa due to the distribution of ratings. From the literature, reliability studies of ED triage systems were selected and the reported data were used to calculate a kappa that accounts for distribution differences. Finally, the reliability of triage systems was compared by means of the normal kappa and the quadratically weighted kappa.

METHODS

Literature search

A search in four online databases was conducted: Pubmed, Embase, Cinahl and Cochrane. Also the online search engine Google Scholar was used[32]. The search terms were used in a previous study (table 1)[31]. Studies were included when they reported about the reliability of three-, four- or five-level ED triage systems. Studies reporting a single kappa without contingency tables or percentages agreement were excluded from the present study because the distribution of the data could not be derived. Finally, the reference lists from all included studies were reviewed for additional relevant articles.

Table 1. Search terms used in online databases and an online search engine

Database	Search terms (hits)	Results
Pubmed	"Triage (Mesh)" AND "Emergency Service Hospital (Mesh)" (1613)	24
	"Triage (Mesh)" AND "Validation studies (publication type)" (77)	2
	"Triage (Mesh)" AND "Reliability" (71)	1
	"Triage (Mesh)" AND "Emergency Service Hospital (Mesh)" (AND) "Validation studies (publication type)" (32)	0
	"Emergency Severity Index" (25)	1
	"Manchester Triage System" (12)	1
	"Canadian Triage and Acuity Scale" (38)	3
	"Australasian Triage and Acuity Scale" (10)	0
Cochrane	"Triage" AND "Emergency Services Hospital" (67)	0
	"Triage" AND "Emergency Department" (AND) "Reproducibility of results" (2)	0
	"Emergency Severity Index" (1)	0
	"Manchester Triage System" (1)	0
	"Canadian Triage and Acuity Scale" (1)	0
	"Australasian Triage Scale" (1)	0
Embase	"Triage" AND "Emergency ward (Emtree)" (467)	1
	"Triage" AND "Reliability (Emtree)" (159)	0
	"Triage" AND "Validation study (Emtree)" (18)	0
	"Triage" AND "Emergency ward (Emtree)" AND "Reliability (Emtree)" (43)	0
	"Emergency Severity Index" (18)	0
	"Manchester Triage System" (19)	0
	"Canadian Triage and Acuity Scale" (22)	0
	"Australasian Triage Scale" (14)	1
Cinahl	(MH "triage") OR (MH "triage (Iowa NIC)") AND "emergency nursing" (513)	1
	(MH "reliability") OR (MH "reliability and validity") OR (MH "interpreter reliability") OR (MH "intrastate reliability") OR (MH "test retest reliability") AND (MH "triage") OR (MH"triage (Iowa NIC)") (73)	0
	"Emergency Severity Index" (21)	0
	"Manchester Triage System" (13)	0
	"Canadian Triage and Acuity Scale" (14)	0
	"Australasian Triage Scale" (17)	0
Google Scholar	"triage" AND "emergency department" AND "interrater reliability" (371) "triage" AND "emergency nursing" AND "reliability" (446)	2 0
Hand search		3
Total		40

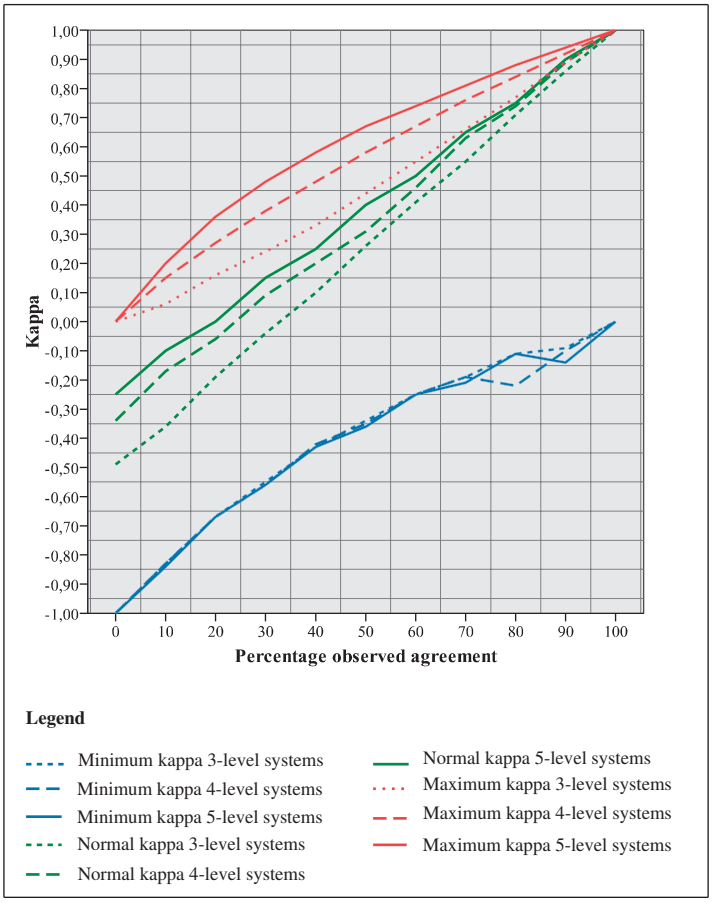
Equations for estimating the normal kappa for scales with 3, 4 or 5 categories

In order to solve the paradox of low kappa values and high agreement, Lantz and Nebenzahl showed how kappa depends on the distribution of the data[21]. They developed the minimum, maximum, and normal kappa. The minimum kappa was obtained when all agreement was located in one cell of the contingency table while disagreement was equally distributed. For the maximum kappa, exactly the opposite situation occurred. Normal kappa was obtained when both, agreement and disagreement were symmetrically distributed over the cells of the contingency table. In the literature, the latter is also known as the prevalence and bias adjusted kappa[33]. The approach of Lantz and Nebenzahl can have several advantages for comparing the reliability of triage systems. Plotting kappa against the minimum, normal, and maximum kappa provides knowledge about the extensiveness and the distribution of disagreement between raters. As a result, the reliability of triage systems can be interpreted in more detail. A second advantage is that by calculating the normal kappa one corrects for influences due to the distribution of ratings. Lantz and Nebenzahl developed their equations for unweighted kappa and binary measurement scales. However, ED triage systems mostly have three to five categories. Therefore, simulations were conducted with one hundred ratings each to develop equations for measurement scales with three to five categories. For every 10% increase of agreement in ratings, the minimum, normal, and maximum linear and quadratically weighted kappas were calculated in three- to five-level systems. To construct the contingency tables for the simulations, an algorithm was applied to determine the number of agreeing and/or disagreeing observations in each cell. The algorithm: (proportion observed agreement or disagreement/number of scale categories) * total number of ratings, was used to symmetrically distribute agreement and disagreement data over the contingency table. Asymmetries in the data which occurred when calculating the minimum or maximum kappa were handled differently. In case of asymmetry in agreement data (minimum kappa), the data can be allocated to every cell of the contingency table that represents agreement, because the weights in these cells are equal (1.00). To obtain a minimum kappa in binary scales, the disagreement data was equally divided over the cells of the contingency table[21]. In the simulations, it appeared that the minimum kappa in three- to five-level scales was obtained when the disagreement data was equally distributed over the two cells that were located within one category from the agreement data. Asymmetries in disagreement data for calculating a maximum kappa were obtained by placing all disagreement data within one cell from agreement. This was done because the largest weights of both weighting schemes are located within one category from agreement. The contingency tables for every simulation were imported into AGREE for Windows (a computer program for calculating weighted and unweighted kappas)[34], to calculate linear and quadratically weighted kappa. The results of the simulations were exported into SPSS for Windows version 14 for further analyses and these results are presented in figures 1 (linear weighted kappa) and 2a-b (quadratically weighted kappa). In these figures, the lines represent the minimum, normal, and maximum weighted kappa in three- to five-level scales as a function of the percentages observed agreement. The percentages expected agreement of the minimum, normal, and maximum weighted kappas were compared in table 2 to sort out which kappa (minimum, normal, or maximum) was suitable for the comparison of triage systems. It was decided to calculate the expected percentages agreement for unweighted kappa because it shows the effect of this percentage on the different numbers of categories more clearly than weighted kappa. In this study, a kappa was defined suitable for comparing triage systems when the expected percentages agreement were equal for every percentage observed agreement and did not exceed the

probability on agreement in general. This was the case for normal kappa in which the expected percentages agreement (33%) were independent of the percentages observed agreement.

Because the normal kappa in figures 1 and 2 approximate a straight line, linear functions (y=ax+b) were created to calculate a normal weighted kappa from the selected triage reliability studies. In these functions the 'a' is obtained by dividing the distance between two outcomes on the y axis by the distance between two outcomes on the x axis. The 'b' can be calculated after the calculation of 'a' by filling in the linear algorithm (Appendix A). As a result, for studies reporting about the reliability of triage systems with five categories, linear weighted and quadratically weighted normal kappa are equal and approximate the algorithm: kappa=1.250(Po)-0.225. The algorithms: quadratically weighted normal kappa= 1.340(Po)-0.380 and linear weighted normal kappa= 1.340(Po)-0.360 approximate normal kappa for systems with four categories. Finally, for systems with three categories the algorithms: quadratically weighted normal kappa= 1.480(Po)-0.502 and linear weighted normal kappa= 1.490(Po)-0.496 were created. In these algorithms (Po) is the proportion observed agreement.

Figure 1. Linear weighted kappa in three to five-level systems



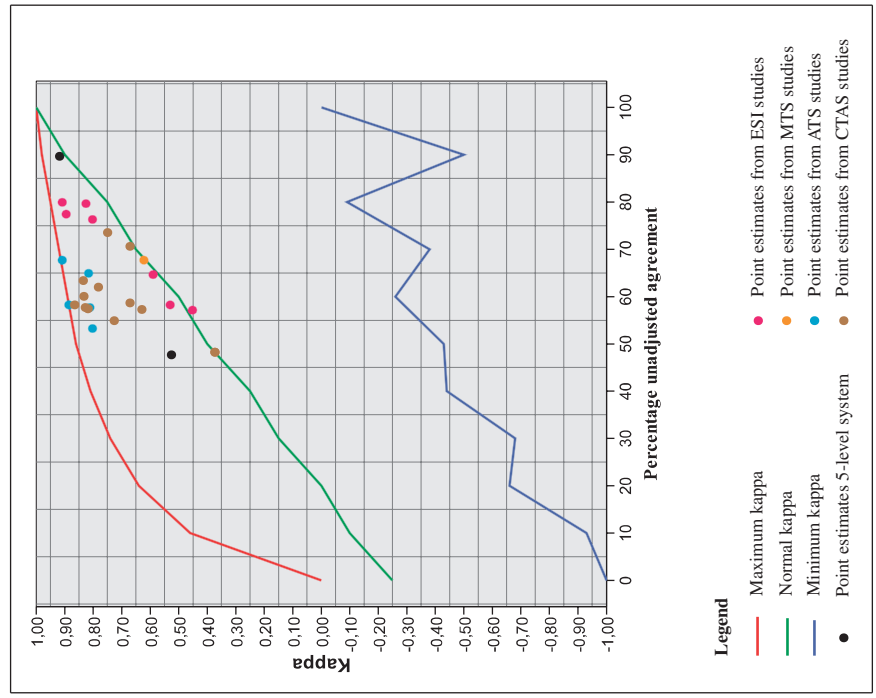


Figure 2b Quadratically weighted kappa of five-level triage systems

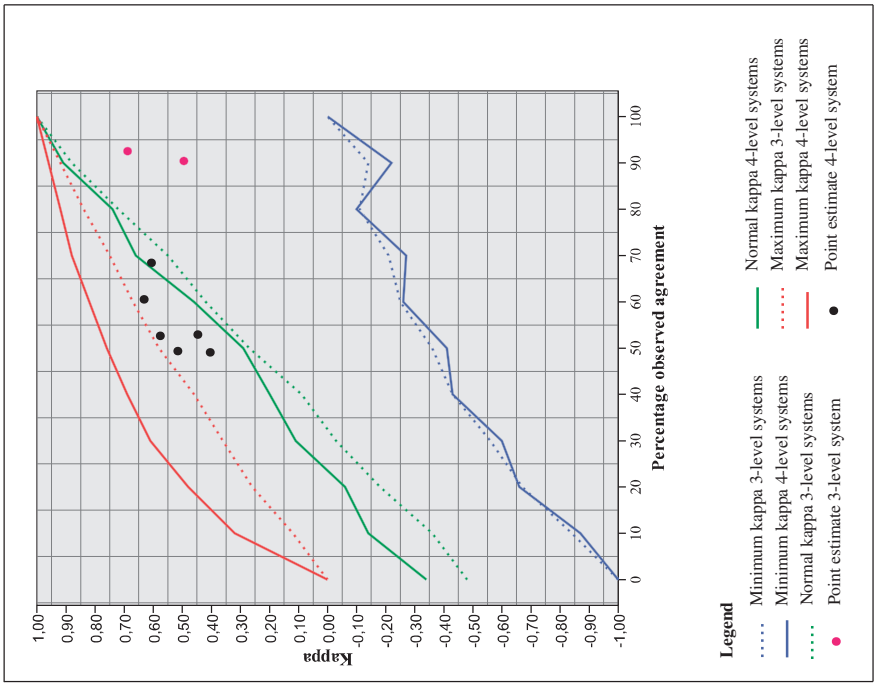


Figure 2a Quadratically weighted kappa of three to five-level triage systems

*Table 2. Expected percentages agreement for minimum, normal, and maximum unweighted kappa**

% observed agreement	Expected % agreement minimum kappa	Expected % agreement normal kappa	Expected % agreement maximum kappa
0	50.0	33.4	0.0
10	50.6	33.4	6.7
20	52.0	33.4	11.8
30	54.6	33.4	17.0
40	58.0	33.4	21.6
50	62.6	33.4	24.9
60	68.0	33.4	28.0
70	74.6	33.4	27.6
80	82.0	33.4	32.0
90	90.6	33.4	33.0
100	100.0	33.4	33.4

**Numbers are calculated for measurement scales with three levels*

RESULTS

Initially sixty-six articles were included in the study. Twenty-six studies had to be excluded because: they reported a single kappa (n=8), no abstract and/or full-text articles were available (n=6), the articles did not report the (overall) reliability of a triage system (n=6), did report about a triage system which allocated patients to different types of caregivers (n=3) or were focused on specific complaints e.g. trauma (n=3). In total forty articles were available for further analysis (table 1).
Three common methods of measuring the reliability of a triage system were used: paper based case scenarios (n=21), retrospective review of charts and/or triage notes (n=12) and the triage of patients in the ED (n=11). One study used all three methods and two studies combined two methods. Of the studies measuring the reliability with case scenarios, the distribution of the numbers of cases per triage category were in four studies based on the casemix of the EDs under study[12;28;35;36].
In eight studies[15;37-43] the investigator selected scenarios which were not derived from actual ED casemix. In nine studies this was not mentioned[10;17;44-50].

Comparing kappas from adjusted data

The algorithms for calculating linear and quadratically weighted normal kappa in three- to five-level systems differ slightly. For each algorithm, the normal kappa was calculated for every 10% increase in the proportion observed agreement. The outcomes of these calculations were transported into SPSS for Windows and the mean differences between kappa values were compared between three-, four- and five-level systems. The mean differences between kappas of three-level systems compared to four-level systems (mean difference: 0.09; 95% CI: 0.04 to 0.13) and three-level systems compared to five-level systems (mean difference: 0.12; 95% CI: 0.07 to 0.18) increased slightly with the number of categories. The same effect was seen when comparing linear weighted normal kappas. As the differences are similar for linear weighted and quadratically weighted normal kappa, the reliability of triage systems can be compared with both types of normal kappa. In this study triage systems will be compared with quadratically weighted normal kappa because it is the most frequently used measure in reliability studies[51]. Moreover, in this study triage systems are compared by interpreting the normal kappa along with quadratically weighted

kappa, plotted against the minimum, normal, and maximum kappa. Calculating normal kappa with quadratic weights will therefore result in more precise interpretations of the reliability of triage systems.

From the forty included studies fifty-eight normal kappas were calculated from the reported data (table 3). Of these, twenty-nine studies reported about a five-level triage system, six studies about a four-level triage system and four about a three-level triage system. One study reported about a three- and a five-level triage system. Interpretation of the normal kappas according to the classification scheme of Landis and Koch[52] revealed that a majority of the five-level triage systems had moderate (n=19 kappas) or substantial (n=11 kappas) reliability (median normal kappa: 0.57 range: 0.08 to 0.93)[8;10;12;16;17;23;25;27;28;35;36;38-41;44-48;50;53-56]. Reliability of four-level triage systems was fair (n=6 kappas) to moderate (n=4 kappas) (median normal kappa: 0.37 range: 0.25 to 0.53) while the majority of the three-level triage systems had almost perfect reliability (n=6 kappas), (median normal kappa: 0.84 range: 0.53 to 0.98)[6;15;24;26;38;42;57-59].

Because five-level triage systems were evaluated frequently, the reliability of the ESI (n=10), CTAS (n=13), ATS (n=8), and MTS (n=5) could be compared. The median normal kappa of the ESI (median: 0.74 range: 0.49 to 0.81) appeared to be the highest [8;10;16;17;45;50;55;56]. The median normal kappa of the ATS (median: 0.54 range: 0.45 to 0.62)[12;35;40;41;44] and the CTAS (median: 0.52 range: 0.37 to 0.70)[22;23;25;27;36;38;39;46-48;54] were moderate while the MTS had the smallest median normal kappa (median: 0.33 range: 0.08 to 0.61)[28;53;60].

Comparing kappas from unadjusted data

From the studies not reporting a quadratically weighted kappa, the reported contingency tables were used to calculate this type of weighted kappa. In order to express the magnitude of deviations from the minimum, normal, and maximum kappa, the kappas of the included studies were integrated with dots into figure 2a/b. For three-level triage systems only one study was found that reported a quadratically weighted kappa combined with a percentage agreement or a contingency table[26]. In this study, in which two kappas were reported, percentages observed agreement ranged between 70% and 93% while quadratically weighted kappa ranged between 0.57 and 0.68. Figure 2a shows that these kappas deviate substantially from the normal kappa, which indicates asymmetries in the distribution of the data. This could be confirmed by the contingency tables in the report. In both phases of that study, 87% of all agreement was centered in the third category.

Four studies reporting about four-level triage systems are presented in figure 2a. Three kappas substantially deviate from the normal kappa in exactly the opposite direction as was seen in three-level triage systems, which indicates skewness in disagreement. The percentages observed agreement ranged between 49% and 68% and quadratically weighted kappa ranged between 0.40 and 0.64[6;15;57;58]. From studies reporting about the reliability of five-level triage systems, twenty-seven kappas from twenty studies were entered into figure 2b. In these studies, the percentages observed agreement ranged between 47% and 89% while quadratically weighted kappa ranged between 0.36 and 0.91 [10;14;16;17;22;23;25;27;28;35;36;38-40;43;46;48;50;54;55]. The distribution of kappa values in five-level triage systems was similar to the distribution of four-level triage systems. Fourteen kappas were located close to the maximum kappa while the others approached the normal kappa closely. The colored dots in the figure represent the four most frequently implemented triage systems: CTAS (brown), ATS (blue), ESI (pink) and MTS (orange). The figure shows that the reliability of the CTAS and ATS was mainly located close to the maximum kappa. Reliability of the ESI and

MTS was located around the normal kappa or in between the maximum and normal quadratically weighted kappa.

DISCUSSION

In this study, we have compared the reliability in previously reported studies on triage systems with three to five categories. Previous studies compared triage systems by interpreting kappas that did not account for the distribution of the data. In the present study this approach was extended. We propose to compare triage systems with the normal kappa because the chance corrected agreement is independent of the distribution of the data. Interpretation of this type of kappa revealed substantial theoretical differences in the reliability of three-, four- and five-level triage systems. Furthermore, it is important to plot kappa against the minimum, normal, and maximum kappa because part of the differences in observed reliability of three-, four- and five-level triage systems can be explained by differences in the distribution of ratings. In one study reporting about the reliability of a three-level triage system, skewness in the distribution of percentages agreement was observed. In studies reporting about the reliability of four- and five-level triage systems more skewness was found in the distribution of disagreement compared to three-level triage systems. In these studies, disagreement was mainly located within one level of agreement. Therefore, it seems that four- and five-level triage systems have less extensive mistriage compared to three-level triage systems. However, one must be cautious with the interpretation of these data because of the difference in the numbers of included studies.
To interprete the reliability, it is important to report both, quadratically weighted kappa from unadjusted data and normal kappa from adjusted data. The normal kappa should only be used for comparing the reliability of triage systems, it is not informative when a single triage system is evaluated. By calculating quadratically weighted kappa from unadjusted data and plotting this kappa to the minimum, normal, and maximum kappa, important information about mistriage or disagreement is revealed. For example, a kappa which approaches the maximum kappa indicates a skewness in disagreement, close to agreement[21]. Although in the maximum kappa the proportion expected agreement is skewed, this situation is preferred when evaluating triage systems because, from a validity point of view, it indicates that no extensive mistriage or disagreement occurred. This was seen in the ATS and the CTAS. Conversely, studies reporting kappas which approached the normal kappa value, had extensive mistriage or disagreement because disagreement in these studies was distributed more symmetrically over the cells in the contingency tables. This was found in two studies evaluating the ESI and one study evaluating the MTS. Therefore, when comparing the reliability of triage systems one should take into account both outcomes and should consider to add more weight to the quadratically weighted kappa than to the normal kappa.
In addition, it is important to report the way in which patients or scenarios were selected in a reliability study. A majority of the studies accounted for actual ED casemix when measuring reliability. Studies that were not based on the ED casemix could have overestimated the reliability of a triage system. For example, in these studies more urgent patients were triaged than would normally present to the ED. These patients were probably recognized more easily by raters because in most triage systems the guidelines for patients in the most urgent categories are very straightforward[2-5]. This could result in higher agreement rates.

Our simulations showed that the mean difference between linear and quadratically weighted normal kappa remained equal when the number of categories of the measurement scale increased. These results differ from the study of Brenner and Kliebsch[18] who preferred the use of linear weighted kappa when comparing measurement scales. In their study kappa was calculated from unadjusted data and increased when the number of categories of the measurement scale increased. In our study the normal kappa was calculated, which could explain these differences because in their study the ratings were normally and exponentially distributed. Another difference that occurred in this study was that, compared to the study of Lantz and Nebenzahl[21], the curves representing minimum, normal, and maximum kappa were slightly different. In the present study, the curves are less smooth and the shape of the curves is slightly different. These differences could be explained by the fact that Lantz and Nebenzahl calculated the unweighted minimum, normal, and maximum kappa for binary scales while the present study calculated linear and quadratically weighted kappa for three- to five-level scales.

In conclusion, when comparing the reliability of triage systems (with differences in numbers of categories) one should report both, the normal kappa, and the quadratically weighted kappa. Interpretation of the reliability should be done in accordance with both of these outcomes. This study showed that substantial theoretical differences in interrater reliability existed between three-, four- and five-level triage systems than previously reported.

Table 3. Quadratically weighted normal kappa calculated from previous studies

Study	Nr. categories	Po	Nr of ratings	κ normal
Beveridge[39]	5 (CTAS)	0.598	450	0.52
		0.566	400	0.48
Worster[27]		0.579	271	0.50
Grafstein[54]		0.737	3990	0.70
Bergeron.[38]		0.637	1485	0.57
		0.597	715	0.52
Manos[36]		0.634	820	0.57
Göransson[46;47]		0.576	7550	0.50
Gravel[48]		0.570	970	0.49
		0.550	974	0.46
Dong[23]		0.592	569	0.52
Dong[22]		0.479	693	0.37
Gravel[25]		0.709	499	0.66
Wuerz[50]	5 (ESI)	0.770	351	0.74
Kim[55]		0.647	955	0.58
		0.586	353	0.51
		0.573	328	0.49
Durani[45]		0.830	1320	0.81
Travers[8]		0.710	305	0.66
Baumann[10]		0.800	20	0.78
Toulson.[56]		0.800	71	0.78
Tanabe[16]		0.797	359	0.77
Wuerz[17]		0.763	219	0.73
Maningas[13]	5 (SRTS)	0.920	117	0.93
Maningas[14]		0.889	423	0.89
Wollaston.[43]	5 (TATTT)	0.466	58	0.36
Considine[40]	5 (ATS)	0.581	310	0.50
Gerdtz[12]		0.659	744	0.60
Considine[35]		0.645	1176	0.58
		0.679	1133	0.62
		0.566	1173	0.48
		0.542	1132	0.45
Doherty[44]		0.625	96	0.57
Goodacre[60]	5 (MTS)	0.300	200	0.15
		0.240	200	0.08
		0.440	200	0.33
van der Wulp[28]		0.671	2382	0.61
Cooke[53]		0.670	91	0.61
George[24;57]	4	0.491	1213	0.28
Brillman[7]		0.679	2737	0.53
		0.598	1212	0.42
Brillman[6]		0.678	3949	0.53
		0.471	3618	0.25
Bergeron[37]		0.642	2145	0.48
		0.535	1320	0.34
Nakagawa[58]		0.532	201	0.33
		0.493	201	0.28
Rutschmann[15]		0.580	1113	0.40
Hay[26]	3	0.905	1310	0.84
		0.930	1576	0.87
Maldonado[42]		0.750	396	0.61
		0.750	612	0.61
		0.910	492	0.84
		0.910	360	0.84
		1.00	900	0.98
Travers[8]		0.700	303	0.53
Wong[59]		0.903	534	0.83

REFERENCES

1 Fernandes CM, Tanabe P, Gilboy N, Johnson LA, McNair RS, Rosenau AM, et al. Five-level triage: a report from the ACEP/ENA Five-level Triage Task Force. Journal of Emergency Nursing. 2005;31:39-50.

2 Australasian College for Emergency Medicine. Policy Document: The Australasian Triage Scale. 2000.

3 Beveridge R, Clarke B, Janes L, Savage N, Thompson J, Dodd G, et al. Guidelines for the Canadian Emergency Department Triage & Acuity Scale (CTAS); 1998.

4 Gilboy N, Tanabe P, Travers D, Rosenau A, Eitel DR. Emergency Severity Index, Version 4: Implementation Handbook. Rockville: Agency for Healthcare Research and Quality; 2005.

5 Mackway-Jones K. Emergency Triage. 2 ed. London: Manchester Triage Group; 2005.

6 Brillman JC, Doezema D, Tandberg D, Sklar DP, Davis KD, Simms S, et al. Triage: limitations in predicting need for emergent care and hospital admission. Annals of Emergency Medicine. 1996;27:493-500.

7 Brillman JC, Doezema D, Tandberg D, Sklar DP, Skipper BJ. Does a physician visual assessment change triage? American Journal of Emergency Medicine. 1997;15:29-33.

8 Travers DA, Waller AE, Bowling JM, Flowers D, Tintinalli J. Five-level triage system more effective than three-level in tertiary emergency department. Journal of Emergency Nursing. 2002;28:395-400.

9 Wuerz RC, Fernandes CM, Alarcon J. Inconsistency of emergency department triage. Emergency Department Operations Research Working Group. Annals of Emergency Medicine. 1998;32:431-435.

10 Baumann MR, Strout TD. Evaluation of the Emergency Severity Index (version 3) triage algorithm in pediatric patients. Academic Emergency Medicine. 2005;12:219-224.

11 Eitel DR, Travers DA, Rosenau AM, Gilboy N, Wuerz RC. The emergency severity index triage algorithm version 2 is reliable and valid. Academic Emergency Medicine. 2003;10:1070-1080.

12 Gerdtz MF, Bucknall TK. Influence of task properties and subjectivity on consistency of triage: a simulation study. Journal of Advanced Nursing. 2007;58:180-190.

13 Maningas PA, Hime DA, Parker DE. The use of the Soterion Rapid Triage System in children presenting to the Emergency Department. Journal of Emergency Medicine. 2006;31:353-359.

14 Maningas PA, Hime DA, Parker DE, McMurry TA. The Soterion Rapid Triage System: evaluation of inter-rater reliability and validity. Journal of Emergency Medicine. 2006;30:461-469.

15 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. Journal of Clinical Epidemiology. 2006;59:615-621.

16 Tanabe P, Gimbel R, Yarnold PR, Kyriacou DN, Adams JG. Reliability and validity of scores on The Emergency Severity Index version 3. Academic Emergency Medicine. 2004;11:59-65.

17 Wuerz RC, Travers D, Gilboy N, Eitel DR, Rosenau A, Yazhari R. Implementation and refinement of the emergency severity index. Academic Emergency Medicine. 2001;8:170-176.

18 Brenner H, Kliebsch U. Dependence of weighted kappa coefficients on the number of categories. Epidemiology. 1996;7:199-202.

19 Cohen J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement. 1960;20:37-46.

20 Maclure M, Willett WC. Misinterpretation and misuse of the kappa statistic. American Journal of Epidemiology. 1987;126:161-169.

21 Lantz CA, Nebenzahl E. Behavior and interpretation of the kappa statistic: resolution of the two paradoxes. Journal of Clinical Epidemiology. 1996;49:431-434.

22 Dong SL, Bullard MJ, Meurer DP, Colman I, Blitz S, Holroyd BR, et al. Emergency triage: comparing a novel computer triage program with standard triage. Academic Emergency Medicine. 2005;12:502-507.

23 Dong SL, Bullard MJ, Meurer DP, Blitz S, Ohinmaa A, Holroyd BR, et al. Reliability of computerized emergency triage. Academic Emergency Medicine. 2006;13:269-275.

24 George S, Read S, Westlake L, Fraser-Moodie A, Pritty P, Williams B. Differences in priorities assigned to patients by triage nurses and by consultant physicians in accident and emergency departments. Journal of Epidemiology and Community Health. 1993;47:312-315.

25 Gravel J, Gouin S, Manzano S, Arsenault M, Amre D. Interrater Agreement between Nurses for the Pediatric Canadian Triage and Acuity Scale in a Tertiary Care Center. Academic Emergency Medicine. 2008;15:1262-1267.

26 Hay E, Bekerman L, Rosenberg G, Peled R. Quality assurance of nurse triage: consistency of results over three years. American Journal of Emergency Medicine. 2001;19:113-117.

27 Worster A, Sardo A, Eva K, Fernandes CM, Upadhye S. Triage Tool Inter-rater Reliability: A Comparison of Live Versus Paper Case Scenarios. Journal of Emergency Nursing. 2007;33:319-323.

28 van der Wulp I, van Baar ME, Schrijvers AJP. Reliability and validity of the Manchester Triage System in a general emergency department patient population in the Netherlands: Results of a simulation study. Emergency Medicine Journal. 2008;25:431-434.

29 Cicchetti DV, Feinstein AR. High agreement but low kappa: II. Resolving the paradoxes. Journal of Clinical Epidemiology. 1990;43:551-8.

30 Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. Journal of Clinical Epidemiology. 1990;43:543-549.

31 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. Journal of Clinical Epidemiology. 2009;62:1196-1201.

32 Google Scholar available at: http://scholar.google.com. 2008.

33 Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. Journal of Clinical Epidemiology. 1993;46:423-429.

34 AGREE for Windows [computer program]. Available at: http://www.scienceplus.nl Version 7.002 1999.

35 Considine J, LeVasseur SA, Villanueva E. The Australasian Triage Scale: examining emergency department nurses' performance using computer and paper scenarios. Annals of Emergency Medicine. 2004;44:516-523.

36 Manos D, Petrie DA, Beveridge RC, Walter S, Ducharme J. Inter-observer agreement using the Canadian Emergency Department Triage and Acuity Scale. Canadian Journal of Emergency Medicine. 2002;4:16-22.

37 Bergeron S, Gouin S, Bailey B, Patel H. Comparison of triage assessments among pediatric registered nurses and pediatric emergency physicians. Academic Emergency Medicine. 2002;9:1397-1401.

38 Bergeron S, Gouin S, Bailey B, Amre DK, Patel H. Agreement among pediatric health care professionals with the pediatric Canadian triage and acuity scale guidelines. Pediatric Emergency Care. 2004;20:514-518.

39 Beveridge R, Ducharme J, Janes L, Beaulieu S, Walter S. Reliability of the Canadian emergency department triage and acuity scale: interrater agreement. Annals of Emergency Medicine. 1999;34:155-159.

40 Considine J, Ung L, Thomas S. Triage nurses' decisions using the National Triage Scale for Australian emergency departments. Accident and Emergency Nursing. 2000;8:201-209.

41 Considine J, Ung L, Thomas S. Clinical decisions using the National Triage Scale: how important is postgraduate education? Accident and Emergency Nursing. 2001;9:101-108.

42 Maldonado T, Avner JR. Triage of the pediatric patient in the emergency department: are we all in agreement? Pediatrics. 2004;114:356-360.

43 Wollaston A, Fahey P, McKay M, Hegney D, Miller P, Wollaston J. Reliability and validity of the Toowoomba adult trauma triage tool: a Queensland, Australia study. Accident and Emergency Nursing. 2004;12:230-237.

44 Doherty S. Application of the National Triage Scale is not uniform. Emergency Nursing Journal. 1996;1:26.

45 Durani Y, Brecher D, Walmsley D, Attia MW, Loiselle JM. The Emergency Severity Index (Version 4): Reliability in pediatric patients. Academic Emergency Medicine. 2007;14(Suppl 1):S95-S96.

46 Göransson K, Ehrenberg A, Marklund B, Ehnfors M. Accuracy and concordance of nurses in emergency department triage. Scandinavian Journal of Caring Sciences. 2005;19:432-438.

47 Göransson KE, Ehrenberg A, Marklund B, Ehnfors M. Emergency department triage: is there a link between nurses' personal characteristics and accuracy in triage decisions? Accident and Emergency Nursing. 2006;14:83-88.

48 Gravel J, Gouin S, Bailey B, Roy M, Bergeron S, Amre D. Reliability of a computerized version of the Pediatric Canadian Triage and Acuity Scale. Academic Emergency Medicine. 2007;14:864-869.

49 Worster A, Gilboy N, Fernandes CM, Eitel D, Eva K, Geisler R, et al. Assessment of inter-observer reliability of two five-level triage and acuity scales: a randomized controlled trial. Canadian Journal of Emergency Medicine. 2004;6:240-245.

50 Wuerz RC, Milne LW, Eitel DR, Travers D, Gilboy N. Reliability and validity of a new five-level triage instrument. Academic Emergency Medicine. 2000;7:236-242.

51 Graham P, Jackson R. The analysis of ordinal agreement data: beyond weighted kappa. Journal of Clinical Epidemiology. 1993;46:1055-1062.

52 Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33:159-174.

53 Cooke MW, Jinks S. Does the Manchester triage system detect the critically ill? Journal of Accident and Emergency Medicine. 1999;16:179-181.

54 Grafstein E, Innes G, Westman J, Christenson J, Thorne A. Inter-rater reliability of a computerized presenting-complaint-linked triage system in an urban emergency department. Canadian Journal of Emergency Medicine. 2003;5:323-329.

55 Kim TG, Cho JK, Kim SH, Lee HS, Gu HD, Chung SW. Reliability and validity of the modified Emergency Severity Index-2 ad a triage tool. Journal of the Korean Society of Emergency Medicine. 2006;17:154-164.

56 Toulson K, Laskowski-Jones L, McConnell LA. Implementation of the five-level emergency severity index in a level I trauma center emergency department with a three-tiered triage scheme. Journal of Emergency Nursing. 2005;31:259-264.

57 George S, Read S, Westlake L, Williams B, Fraser-Moodie A, Pritty P. Evaluation of nurse triage in a British accident and emergency department. British Medical Journal. 1992; 4;304(6831):876-878.

58 Nakagawa J, Ouk S, Schwartz B, Schriger DL. Interobserver agreement in emergency department triage. Annals of Emergency Medicine. 2003;41:191-195.

59 Wong TW, Tseng G, Lee LW. Report of an audit of nurse triage in an accident and emergency department. Journal of Accident and Emergency Medicine. 1994;11:91-95.

60 Goodacre SW, Gillett M, Harris RD, Houlihan KP. Consistency of retrospective triage decisions as a standardized instrument for audit. Journal of Accident and Emergency Medicine. 1999;16:322-324.

5

Sample size calculation methods for reliability studies

Based on:
van der Wulp I, van Stel H.F. A review of sample size calculation methods for reliability studies revealed no appropriate methods for ordinal scales.
Submitted.

ABSTRACT

Objective: to review which methods for calculating sample sizes in reliability studies are described in the literature and to study the extent to which these methods were applied in triage reliability studies.

Methods: a systematic search of the literature was conducted in four online databases and a search engine. Papers were included when describing methods for calculating sample sizes in reliability studies which report the reliability with a kappa statistic. Moreover, a literature search from a previous study was updated and papers describing reliability studies to emergency department triage systems which reported a kappa statistic were included. In these papers we checked which sample size calculation methods were applied.

Results: Sixteen papers describing sample size calculation methods were included in the study. Two types of sample size calculation methods were found for reliability studies which estimate the precision of kappa or test hypotheses. Forty-two papers describing triage reliability studies were included, of which thirty-five did not report information about sample size calculations.

Conclusions: Most sample size calculation methods described in the literature were focussed on two raters and binary scales. Only one method was extended to multiple raters who use a nominal scale. Currently, no methods exist for ordinal scales. When planning a triage reliability study we suggest to account for the behaviour of kappa by calculating sample sizes with smaller confidence interval widths or to derive the null and alternative hypotheses from the maximum weighted kappa.

INTRODUCTION

Triage systems are methods often used in healthcare settings with the purpose to allocate patients on the basis of the urgency of their complaint[1]. In the late eighties and early nineties of the twentieth century, emergency departments (EDs) mainly used triage systems with three or four categories. In recent years, triage systems with five categories were developed and implemented[2]. Since the development of five-level triage systems, evaluations of the reliability of these systems in the literature became more common.

Reliability is a reflection of the amount of systematic and random error in measurements. A scale is reliable when only variability between raters exist[3]. When measuring the reliability of a categorical measurement scale one can calculate a kappa statistic. The kappa statistic is a measure of agreement between raters or methods which corrects for agreement that exist due to chance[4]. However, the kappa statistic depends heavily on the distribution of ratings, which could lead to high agreement between raters but low chance corrected agreement. This paradox was first described by Feinstein et al.[5], followed by Lantz et al.[6] and Byrt et al.[7]. As a result, Lantz et al[6] created algorithms to calculate the minimum, normal, and maximum obtainable kappa. The minimum kappa is obtained when the proportion of agreement is centred to one cell of the contingency table while disagreement is equally distributed over the two cells that are located within one category from agreement. The maximum kappa is obtained when agreement is equally distributed and disagreement is located to one cell of the contingency table. The normal kappa is obtained when agreement and disagreement are equally distributed. At the same time Byrt et al.[7] described the 'prevalence and bias adjusted' kappa (PABAK), which is equal to the normal kappa. They propose to use the PABAK kappa when comparisons are made between several reliability studies. In a previously conducted study, the work of Lantz et al.[6] was extended for measurement scales with three, four, or five categories, and for weighted kappa[8]. These methods were subsequently applied and used to plot the results of previously conducted triage reliability studies.

The studies of Lantz et al.[6], Byrt et al.[7] and our study provided more insights into the interpretation of the kappa statistic. The interpretation of a kappa statistic is important when calculating sample sizes for preparing a reliability study. Sample sizes are calculated because of cost constraints, a rare prevalence of the event under study, or to increase the statistical power of the study[9]. Furthermore, reliability studies can be aggravating for study participants, especially when they have to combine their participation with their usual activities in, for example, the ED. Because of these time constraints, as well as to keep participants enthusiastic in participating in research, sample size calculations are necessary. However, it is unknown whether current sample size calculation methods account for influences on kappa due to the distribution of data. The objective of this study was to review sample size calculation methods for studies reporting the reliability with a kappa statistic. Furthermore, the frequency and type of applied sample size calculations in previously conducted triage reliability studies was studied. Finally, suggestions are proposed on how to apply these sample size calculation methods when planning a triage reliability study.

METHODS

Literature search

A systematic search of the literature was conducted to find papers describing sample size calculation methods for reliability studies reporting the reliability with a kappa statistic. This search was conducted in Zentralblatt MATH (mathematical literature), PsycINFO, Pubmed, Embase, and Google Scholar. The search terms are presented in table 1. From the included studies the reference lists were reviewed for additional relevant articles. Papers not written in English or Dutch were excluded from the study.

A second search was conducted to find studies reporting about the reliability of ED triage systems. This search was used in two previous studies and updated to August 2009[8;10]. Papers describing triage reliability studies in which the reliability was reported with a kappa statistic, were included in the study. From the included papers, it was studied which triage reliability studies calculated a sample size prior to the data collection and which methods were applied. Studies calculating sample sizes prior to the data collection were compared with studies that did not, by calculating the median number of raters and cases in both groups. A Mann Whitney test was conducted to test for statistically significant differences between the number of raters and cases in these groups. These analyses were performed with SPSS for Windows (version 15.0).

Table 1. Search strategy to find studies describing sample size calculation methods

Database	Search terms (hits)	Results
Zentralblatt MATH	(“Sample size” OR “statistical power” OR “precision”) AND (“kappa statistic” OR “kappa coefficient” OR “reliability” OR “interrater agreement” OR “interobserver agreement” OR “chance corrected agreement” | “kappa”) (366)	4
PsycINFO	“Sample size.mp” OR “exp Sample size” OR “Statistical power.mp” OR “exp statistical power” AND “Interrater reliability.mp” OR “exp Interrater reliability” OR “Test reliability.mp” OR “exp test reliability” OR “exp rating scales” OR “kappa statistic.mp” OR “interrater agreement.mp” OR “exp statistical reliability” (287)	1
Pubmed	“Sample size (Mesh)” AND “Reproducibility of results (Mesh)” (943)	2
	“Sample size (Mesh)” AND “Statistics as topic (Mesh: no explosion)” (242)	1
	“Sample size (Mesh)” AND “Reproducibility of results (Mesh)” AND “Statistics as topic (Mesh: no explosion)” (32)	0
	“Sample size (Mesh)” AND (“Kappa statistic” OR “Kappa OR “kappa coefficient”) (27)	2
Embase	(‘Sample size (Emtree)’ OR ‘power of a test (Emtree)’) AND (‘Reliability (Emtree)’ OR ‘reproducibility (Emtree)’ OR ‘kappa statistics (Emtree)’) (946)	0
Google Scholar	'Sample size' 'kappa' 'kappa statistic' 'kappa coefficient' 'chance-corrected agreement' 'interobserver agreement' 'interrater agreement' 'reliability' 'reproducibility of results' (26)	0
Hand search		6
Total	2869 abstracts screened	16

RESULTS

Sample size calculation methods in reliability studies

In total, 2869 abstracts were screened which led to the initial inclusion of thirty-two papers and three books. After careful reading, eighteen papers and one book were excluded because: they did not present (new) methods for calculating sample sizes (n=14), presented methods for other outcome measures than kappa (n=4), or the paper was not focussed on reliability studies (n=1). In total, fourteen papers and two books were available for further analysis.

The majority (n=10) of the papers described sample size calculation methods for conducting reliability studies to binary rating scales. Three papers presented methods focussed on nominal scales and one paper presented a general estimate of sample sizes that is required when conducting reliability studies. This latter paper studied the influence of sample size on the widths of the confidence intervals for different types of reliability studies. It was found that the width of the confidence interval increased as the sample size decreased for all types of reliability studies. By plotting the percentage change in confidence interval width as a function of sample size, the author concluded that a minimum sample size of 400 subjects would be needed when conducting a reliability study[11]. However, several algorithms for calculating sample sizes based on the width of the confidence interval are described and provide a more precise estimate (table 2). These methods can be applied in case an estimation of the precision of kappa is to be obtained instead of hypothesis testing[12]. Table 3 presents an overview of sample size calculation methods for studies with the purpose of hypothesis testing. These methods are frequently based on the common correlation model. In this model the probabilities on agreement and disagreement depend on the number of raters and scale categories. Use of the common correlation model assumes an equal distribution of categories between raters and independency between raters[13].

Besides methods for reliability studies to binary and nominal measurement scales, no methods were found that could be applied for reliability studies to ordinal measurement scales.

Table 2.Sample size calculation methods for reliability studies with the purpose of estimating kappa

Method	Type of scale	Fixed nr. raters	Nr of samples	Calculation
Fixed confidence interval width method[9;12;13]	Binary	2	1	$N=4.Z^2\alpha/2/w^2\{(1-ќ)[(1-ќ)(1-2ќ) + (ќ(2-ќ)/(2\pi(1-\pi))]\}$
Lower confidence limit method[12]	Binary	2	1	Only a table with the results of calculations was presented
Hybrid Goodness of fit approach[25]	Binary	2	≥2	Only a table with the results of calculations was presented
Flack maximum standard error approach[18]	Nominal	2	1	$N= 4.Z^2\alpha/2$ (max τ^2 (ќ) // w^2

- π = category prevalence
- ќ = the expected value of kappa (based on previous studies or pilot study)
- w = width of confidence interval
- max τ^2(ќ) = maximum standard error of expected kappa

Table 3. Sample size calculation methods for reliability studies with the purpose of hypothesis testing

Method	Type of scale	Nr. of raters	Hypotheses	Calculation
Donner GOF approach[13-19]	Binary	2	H0:κ= κ0 Ha:κ=κ1	N= λ(1,1-β,α) {[π(1- π) (κ1-κ0)]² / (π² + π(1- π)κ0) + 2[π(1- π) (κ1-κ0)]² / (π (1- π)(1- κ0)) + [π(1- π) (κ1-κ0)]² / ((1- π)² + π(1- π)κ0)}$^{-1}$
	Binary	n	H0:κ=κ0 Ha:κ=κ1	N= λ(1,1-β,α) {Σ (i=0) [Pri(κ1)-Pri(κ0)]² / Pri(κ0)}$^{-1}$
	Nominal	n	H0:κ=κ0 Ha:κ=κ1	N= λ(1,1-β,α) { Σ (i=0)[Pri(π1,..πk,κ1) - Pri(π1,..πk,κ0)] ² / Pri(π1,…. πk,κ0)} $^{-1}$
	Binary	2	H0:κ1=κ2 Ha:κ1≠κ2	N= (Xα/2 + Xβ)² {[π(1- π)]² [1/ (π² + π(1- π) (κ1+κ2/2)) + (2/ π(1- π) (1- (κ1+ κ2/2))) + 1/ (1- π)² + π(1- π)(κ1+κ2/2) . Σ (samples)(κh-(κ1+κ2/2))²}$^{-1}$
	Binary	2	H0:κh=κ Ha:κh≠κ	N= λ(h-1) {[π(1- π)]² [1/ (π² + π(1- π)(κpooled)) + (2/ π(1- π) (1- κpooled)) + (1/ (1- π)² + π(1- π) (κpooled)] .Σ (samples)(κh-pooled)²}$^{-1}$
	Binary	1	H0:κh=κ Ha:κh≠κ	N= λ(2h-2,1-β,α)(1-κ(pooled)) / [π(1- π) {1+(1/κ(pooled)+(1-κ(pooled))²} Σ (samples) th(κh-κ(pooled))²]
Asymptotic power function[20]	Binary	1 or 2	H0:κ=κ0 Ha:κ>κ0	N= {Z(1-α) . V0 $^{1/2}$ + Z(1-β) .V1$^{1/2}$ }² / D2
	Binary	1 or 2	H0:κ=κ0 Ha:κ=κ1	N= λ(1,1-β,α) .V0/ D2
Homogeneity score test [21]	Binary	1 or 2	H0:κh=κ Ha:κh≠κ	N= λ(H-1, -β,α) / {Σ (h=1)th.ch.d²h- (Σ (h=1)th.dh)² / (Σ (h=1)th/ch)}
Fleiss[19;22]	Binary	1	H0:κh=κ Ha:κh≠κ	N= λ(H-1,1-β,α) .c. (1-κ(pooled))² / { Σ (samples) th(κh-κ(pooled)²}
Likelihood score method[19]	Binary	1	H0:κh=κ Ha:κh≠κ	N= λ(H-1,1-β,α)/ [c.{(1+κ(pooled))π(1- π) / (κ(pooled) + (1-κ(pooled))². π(1- π)}. { Σ (samples) th(κh-κ(pooled))²}]
Likelihood ratio test[23]	Nominal	1 or 2	H0:κ1=κ2 Ha:κ1≠κ2	N= λ(f,1-β) / 2(Dfull – Dhom)
Cantor[24]	Binary	2	H0:κ=κ0 Ha:κ=κ1	N= [(Zα√Q0 + Zβ√Q1)/(κ1-κ0)] ²
	Binary	2	H0:κ1=κ2 Ha:κ1≠κ2	N= [(Zα√Q01+Q02) +(Zβ√QA1+ QA2)/(κ1-κ2)] ²
Maximum standard error approach[18]	Nominal	2	H0:κ=κ0 Ha:κ=κ1	N= [Z(1-α) max τ(κ́|κ=κ0) + Z(1-β) max τ(κ́|κ=κ1) / (κ0-κ1)] ²

- κ0= value of kappa under null hypothesis
- κ1= value of kappa under alternative hypothesis
- κh= value of kappa in sample h
- α= type 1 error rate
- β= type 2 error rate
- λ = non centrality parameter of the chi square distribution
- Pri= cell probability (i.e. 1,1,1)
- π = category prevalence
- max τ= the maximum standard error

Sample size calculation methods applied in triage reliability studies

In total, sixty-five papers were initially included in our previous study[8]. Of these, thirty-six papers reported the reliability of an ED triage system with a kappa statistic and were included into the present study. The other papers were excluded because: the reliability of a triage system was not reported with a kappa statistic (n=14), no fulltext or abstract of the paper could be obtained (n=9), or the paper did not report the reliability of a triage system (n=6). The search update resulted in the inclusion of six papers. In total, forty-two papers were available for further analyses.

Information about the calculation of sample size prior to the onset of the study was not reported in thirty-five papers[26-60]. In these studies, a median number of fifty-five patients, patient charts, or case scenarios were triaged by a median number of twenty raters.

Information about sample size calculations was reported in seven studies[61-67]. Of these, five papers described the method for calculating sample sizes that was applied of which four calculated the sample size with the confidence interval width approach and one with the hypothesis testing approach. Worster et al.[65], Considine et al.[61] and Gravel et al.[63] accepted a standard error of 0.05 while this information was not reported by Travers et al.[64]. In a second study by Gravel et al.[62] they applied a method for hypothesis testing and defined a difference between kappas of 0.10 as clinically significant. However, the exact method that was used in these studies was only reported by Worster et al[65].

From two studies it could not be determined whether sample size calculation methods were applied[66;67]. Eitel et al.[66] reported that the sample size in their study was estimated on the basis of a previous study. Maningas et al.[67] estimated that a minimum of 350 patients were needed who: '*could be triaged during evening shifts while still sufficient to determine interrater reliability*'. Studies which calculated a sample size prior to the onset of the study included a median number of 304 patients, patient charts, or case scenarios and a median number of twenty-four raters. A Mann-Whitney test showed no statistically significant differences ($p>0.05$) in the number of raters and cases between studies that calculated sample sizes prior to the onset of the study and studies that did not.

Example

Tables 2 and 3 show that most sample size calculation methods can be applied when assessing the reliability of a scale with two raters. However, the previous section showed that it is more common in triage reliability studies to assess the reliability with more than two raters. Furthermore, sample size calculation methods for studies with the purpose of estimating the precision of kappa are not extended to reliability studies in which nominal or ordinal scales are evaluated with more than two raters. In addition, only one method for studies with the purpose of hypothesis testing is extended to nominal scales. This method was applied in an example, derived from our previously conducted triage reliability study[50]. In 2007, the reliability of the Manchester Triage System was studied in two EDs. Forty-eight triage nurses allocated fifty patient vignettes into one out of five triage categories, resulting in a quadratically weighted kappa of 0.62. Shortly after this study, a second edition of the system became available in Dutch. As an example, we want to repeat the study from 2007 to find out whether the revision of the Manchester Triage System improved the system's reliability. The number of patient vignettes that is needed to test H0: $\kappa=\kappa 0$ and Ha: $\kappa=\kappa 1$ is calculated by applying the method of Altaye et al.[15]. For these calculations, the number of raters (n) is set to five, ten, twenty-five and fifty, $\alpha=0.05$, $\beta=0.20$, $\kappa 0=0.62$ and $\kappa 1=0.70$, 0.80 and 0.90. The expected category prevalence for the five categories of the Manchester Triage System is $\pi 1=0.011$, $\pi 2=0.134$, $\pi 3=0.462$, $\pi 4=0.376$ &

$\pi 5=0.016$. The category prevalences and the null hypothesis: $\kappa 0=0.62$, have been derived from the results of the previous study[50]. Table 4 presents the results of these calculations for which a sample size calculator was developed. This sample size calculator was developed in Microsoft Excel 2003 for Windows and can be ordered for free at i.vanderwulp@hotmail.com. It can be used to calculate sample sizes in future studies. With this calculator, one can set the number of raters (≤ 100), π, α, β, $\kappa 0$, $\kappa 1$, the non centrality parameter (λ), and the sample size is calculated immediately.

Table 4: Sample sizes based on hypothesis testing for evaluating a scale with five categories

Number of raters	K1=0.70	K1=0.80	K1=0.90
5	227	43	17
10	160	29	11
25	120	20	7
50	104	16	5

- Sample sizes are rounded up to nearest integer
- K0 is in all calculations 0.62
- $\beta=0.80$, $\alpha=0.05$

DISCUSSION

We studied which methods for calculating sample sizes in reliability studies reporting a kappa statistic are described in the literature. Two types of methods were found: sample size calculation methods for studies with the purpose to estimate the precision of kappa or to test hypotheses. Most methods were designed for binary measurement scales. Three methods were found that were extended to nominal scales, of which only the Goodness of fit method by Altaye and Donner[15] is suitable when more than two raters participate in the study. No methods were found which could be applied in reliability studies evaluating ordinal scales such as triage systems. Because ordinal scales often have a different distribution of categories compared to nominal scales, extension of the current sample size calculation methods to ordinal scales is necessary. Until these methods are developed, methods for nominal scales should be applied which could result in excessive sample sizes.
Moreover, we reviewed the literature to the extent that sample size calculations were applied in previously conducted triage reliability studies. It appeared that the majority of the papers did not report information about sample size calculations. Of the few studies which reported information about the sample size, the majority of the applied methods were based on confidence interval widths. A possible explanation for not calculating sample sizes prior to triage reliability studies is that several triage systems were studied almost simultaneously, so no prior information about confidence intervals, standard errors or estimated kappa's was available. Another possibility is that triage reliability studies are usually non invasive for participating patients and therefore researchers probably considered it unnecessary to calculate sample sizes. In these studies, reliability is often measured with patient cases, retrospective review of patient records or simultaneous triage. A third possibility is that the methods described in the literature are difficult to apply and sensitive to errors. Furthermore, studies not reporting information on sample size calculations differed from studies that did report this information in that a smaller number of patients or patient cases that was rated. Although no statistically significant differences were found between studies that calculated sample sizes and studies that did not, it is possible that several triage reliability studies did not have a sufficient amount of statistical power. This questions the 'true' reliability of ED triage systems. It is recommended for

future triage reliability studies to calculate a priori a sample size and report this information, including the applied methods.
Variables needed to calculate a sample size such as the estimated kappa under the null and alternative hypotheses and confidence interval widths require a different approach when measuring the reliability of ED triage systems compared to other measurement scales. In ED triage systems it is important that misclassifications or disagreement between raters remains within one category from agreement because (excessive) mistriage can have serious consequences for patients. This approach results in a kappa that will reach the maximum weighted kappa which ranges between 0.0 and 1.0(8). In the literature, sample size calculations for reliability studies with the purpose of hypothesis testing, often estimated a kappa under the null and alternative hypothesis that is derived from an interpretation scheme by Landis and Koch[68]. According to this interpretation scheme, the reliability of a measurement scale should minimally obtain a kappa of 0.40 which represents moderate agreement. A kappa of 0.60 (substantial agreement) is usually defined as the alternative hypothesis. This interpretation is probably based on obtaining a normal kappa because the corresponding percentages observed agreement range between 50% and 65%. In the minimum kappa these values cannot be obtained and in the maximum kappa these values correspond to 10% and 15% observed agreement. We suggest using different kappa values when calculating sample sizes. These values can correspond to the maximum kappa values obtained with 50% and 65% agreement. For sample size calculations with a hypotheses testing approach this results in a null hypothesis of $\kappa 0=0.85$ and an alternative hypothesis of $\kappa a=0.90$.
In case the sample size is calculated with confidence intervals widths, one must be cautious with setting the confidence interval width. A width of 0.20 which was frequently used in the literature seems too large as the population kappa still can range within one category of the interpretation scheme. We suggest using a maximum width of 0.10 which was previously used in triage literature[62]. These suggestions will result in larger sample sizes.
Most sample size calculation methods were developed when the reliability is measured with two raters. Triage reliability studies however, are mostly conducted with more than two raters. Because triage is usually performed by several triage nurses working in shifts it is difficult to test the reliability between two raters who are representative for all ED nurses. It is therefore recommended to extent the existing methods to the case of multiple raters.
In conclusion, methods for calculating sample sizes in reliability studies reporting a kappa statistic are mostly developed for the case of two raters and binary scales. No methods were found that could be applied for ordinal scales and development of these methods is necessary. When measuring the reliability of a triage system, a different approach compared to other scales is needed concerning sample size calculations. In case of hypotheses testing the null and alternative hypotheses should approach the maximum kappa. Sample size calculation methods using confidence interval widths, should set the width of the confidence interval to a maximum of 0.10. As a result, the sample sizes in triage reliability studies will increase.

REFERENCES

1 Iserson KV, Moskop JC. Triage in medicine, part I: Concept, history, and types. Annals of Emergency Medicine. 2007;49:275-281.

2 Fernandes CM, Tanabe P, Gilboy N, Johnson LA, McNair RS, Rosenau AM, et al. Five-level triage: a report from the ACEP/ENA Five-level Triage Task Force. Journal of Emergency Nursing. 2005;31:39-50.

3 Streiner DL, Norman GR. Health Measurement Scales. 3 ed. Oxfort University Press; 2007.

4 Cohen J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement. 1960;20:37-46.

5 Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. Journal of Clinical Epidemiology. 1990;43:543-549.

6 Lantz CA, Nebenzahl E. Behavior and interpretation of the kappa statistic: resolution of the two paradoxes. Journal of Clinical Epidemiology. 1996;49:431-434.

7 Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. Journal of Clinical Epidemiology. 1993;46:423-429.

8 van der Wulp I, van Stel HF. Calculating kappas from adjusted data improved the comparability of the reliability of triage systems: a comparative study. *[Unpublished work]*.

9 Shoukri MM. Sample size requirements for the design of a reliability study. Measurement of interobserver agreement. Chapman & Hall/ CRC; 2003. p. 85-102.

10 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. Journal of Clinical Epidemiology. 2009;62:1196-1201.

11 Charter RA. Sample size requirements for precise estimates of reliability, generalizability, and validity coefficients. Journal of Clinical and Experimental Neuropsychology. 1999;21:559-566.

12 Donner A. Sample size requirements for interval estimation of the intraclass kappa statistic. Communications in Statistics - Simulation and Computation. 1999;28: 415-429.

13 Donner A, Eliasziw M. A goodness-of-fit approach to inference procedures for the kappa statistic: confidence interval construction, significance-testing and sample size estimation. Stat Med. 1992;11:1511-1519.

14 Donner A. Sample size requirements for the comparison of two or more coefficients of inter-observer agreement. Statistics in Medicine. 1998;17:1157-1168.

15 Altaye M, Donner A, Eliasziw M. A general goodness-of-fit approach for inference procedures concerning the kappa statistic. Statistics in Medicine. 2001;20:2479-2488.

16 Bartfay E, Donner A. The effect of collapsing multinomial data when assessing agreement. International Journal of Epidemiology. 2000;29:1070-1075.

17 Altaye M, Donner A, Klar N. Inference procedures for assessing interobserver agreement among multiple raters. Biometrics. 2001;57:584-588.

18 Flack VF, Afifi AA, Lachenbruch PA. Sample size determinations for the two rater kappa statistic. Psychometrika. 1988;53:321-325.

19 Nam JM. Assessment on homogeneity tests for kappa statistics under equal prevalence across studies in reliability. Statistics in Medicine. 2006;25:1521-1531.

20 Nam JM. Testing the intraclass version of the kappa coefficient of agreement with binary scale and sample size determination. Biometrical Journal. 2002;44:558-570.

21 Nam JM. Homogeneity score test for the intraclass version of the kappa statistics and sample-size determination in multiple or stratified studies. Biometrics. 2003;59:1027-1035.

22 Fleiss JL. Determining sample sizes needed to detect a difference between two proportions. Statistical methods for rates and proportions. 2 ed. John Wiley & Sons; 1981. p.33-48.

23 Roberts C. Modelling patterns of agreement for nominal scales. Statistics in Medicine. 2008;27:810-830.

24 Cantor AB. Sample size calculations for Cohen's kappa. Psychological Methods. 1996;1:150-153.

25 Donner A, Zou G. Interval estimation for a difference between intraclass kappa statistics. Biometrics. 2002;58:209-215.

26 Baumann MR, Strout TD. Evaluation of the Emergency Severity Index (version 3) triage algorithm in pediatric patients. Academic Emergency Medicine. 2005;12:219-224.

27 Bergeron S, Gouin S, Bailey B, Patel H. Comparison of triage assessments among pediatric registered nurses and pediatric emergency physicians. Academic Emergency Medicine. 2002;9:1397-1401.

28 Bergeron S, Gouin S, Bailey B, Amre DK, Patel H. Agreement among pediatric health care professionals with the pediatric Canadian triage and acuity scale guidelines. Pediatric Emergency Care. 2004;20:514-518.

29 Beveridge R, Ducharme J, Janes L, Beaulieu S, Walter S. Reliability of the Canadian emergency department triage and acuity scale: interrater agreement. Annals of Emergency Medicine. 1999;34:155-159.

30 Brillman JC, Doezema D, Tandberg D, Sklar DP, Davis KD, Simms S, et al. Triage: limitations in predicting need for emergent care and hospital admission. Annals of Emergency Medicine. 1996;27:493-500.

31 Brillman JC, Doezema D, Tandberg D, Sklar DP, Skipper BJ. Does a physician visual assessment change triage? American Journal of Emergency Medicine. 1997;15:29-33.

32 Dong SL, Bullard MJ, Meurer DP, Colman I, Blitz S, Holroyd BR, et al. Emergency triage: comparing a novel computer triage program with standard triage. Academic Emergency Medicine. 2005;12:502-507.

33 Dong SL, Bullard MJ, Meurer DP, Blitz S, Ohinmaa A, Holroyd BR, et al. Reliability of computerized emergency triage. Academic Emergency Medicine. 2006;13:269-275.

34 Dong SL, Bullard MJ, Meurer DP, Blitz S, Holroyd BR, Rowe BH. The effect of training on nurse agreement using an electronic triage system. Canadian Journal of Emergency Medicine. 2007;9:260-266.

35 Fernandes CM, Wuerz R, Clark S, Djurdjev O. How reliable is emergency department triage? Annals of Emergency Medicine. 1999;34:141-147.

36 George S, Read S, Westlake L, Fraser-Moodie A, Pritty P, Williams B. Differences in priorities assigned to patients by triage nurses and by consultant physicians in accident and emergency departments. Journal of Epidemiology and Community Health. 1993;47:312-315.

37 Gerdtz MF, Bucknall TK. Influence of task properties and subjectivity on consistency of triage: a simulation study. Journal of Advanced Nursing. 2007;58:180-190.

38 Gill JM, Reese CL, Diamond JJ. Disagreement among health care professionals about the urgent care needs of emergency department patients. Annals of Emergency Medicine. 1996;28:474-479.

39 Goodacre SW, Gillett M, Harris RD, Houlihan KP. Consistency of retrospective triage decisions as a standardised instrument for audit. Journal of Accident and Emergency Medicine. 1999;16:322-324.

40 Goransson K, Ehrenberg A, Marklund B, Ehnfors M. Accuracy and concordance of nurses in emergency department triage. Scandinavian Journal of Caring Sciences. 2005;19:432-438.

41 Grafstein E, Innes G, Westman J, Christenson J, Thorne A. Inter-rater reliability of a computerized presenting-complaint-linked triage system in an urban emergency department. Canadian Journal of Emergency Medicine. 2003;5:323-329.

42 Hay E, Bekerman L, Rosenberg G, Peled R. Quality assurance of nurse triage: consistency of results over three years. American Journal of Emergency Medicine. 2001;19:113-117.

43 Maldonado T, Avner JR. Triage of the pediatric patient in the emergency department: are we all in agreement? Pediatrics. 2004;114:356-360.

44 Maningas PA, Hime DA, Parker DE. The use of the Soterion Rapid Triage System in children presenting to the Emergency Department. Journal of Emergency Medicine. 2006;31:353-359.

45 Manos D, Petrie DA, Beveridge RC, Walter S, Ducharme J. Inter-observer agreement using the Canadian Emergency Department Triage and Acuity Scale. Canadian Journal of Emergency Medicine. 2002;4:16-22.

46 Nakagawa J, Ouk S, Schwartz B, Schriger DL. Interobserver agreement in emergency department triage. Annals of Emergency Medicine. 2003;41:191-195.

47 Reilly BM, Evans AT, Schaider JJ, Wang Y. Triage of patients with chest pain in the emergency department: a comparative study of physicians' decisions. The American Journal of Medicine. 2002;112:95-103.

48 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. Journal of Clinical Epidemiology. 2006;59:615-621.

49 Tanabe P, Gimbel R, Yarnold PR, Kyriacou DN, Adams JG. Reliability and validity of scores on The Emergency Severity Index version 3. Academic Emergency Medicine. 2004;11:59-65.

50 van der Wulp I, van Baar ME, Schrijvers AJP. Reliability and validity of the Manchester Triage System in a general emergency department patient population in the Netherlands: Results of a simulation study. Emergency Medicine Journal. 2008;25:431-434.

51 Wollaston A, Fahey P, McKay M, Hegney D, Miller P, Wollaston J. Reliability and validity of the Toowoomba adult trauma triage tool: a Queensland, Australia study. Accident and Emergency Nursing. 2004;12:230-237.

52 Wuerz R, Fernandes CM, Alarcon J. Inconsistency of emergency department triage. Emergency Department Operations Research Working Group. Annals of Emergency Medicine. 1998;32:431-435.

53 Wuerz RC, Milne LW, Eitel DR, Travers D, Gilboy N. Reliability and validity of a new five-level triage instrument. Academic Emergency Medicine. 2000;7:236-242.

54 Wuerz RC, Travers D, Gilboy N, Eitel DR, Rosenau A, Yazhari R. Implementation and refinement of the emergency severity index. Academic Emergency Medicine. 2001;8:170-176.

55 Durani Y, Brecher D, Walmsley D, Attia MW, Loiselle JM. The Emergency Severity Index Version 4: Reliability in Pediatric Patients. Journal of Emergency Medicine. 2007;33:333.

56 Grouse AI, Bishop RO, Bannon AM. The Manchester Triage System provides good reliability in an Australian emergency department. Emergency Medicine Journal. 2009;26:484-486.

57 Olofsson P, Gellerstedt M, Carlstrom ED. Manchester Triage in Sweden - interrater reliability and accuracy. International Emergency Nursing. 2009;17:143-148.

58 Parenti N, Ferrara L, Bacchi Reggiani ML, Sangiorgi D, Lenzi T. Reliability and validity of two four-level emergency triage systems. European Journal of Emergency Medicine. 2009;16:115-120.

59 Storm-Versloot MN, Ubbink DT, Choi V, Luitse JS. Observer agreement of the Manchester Triage System and the Emergency Severity Index: a simulation study. Emergency Medicine Journal. 2009;26:556-560.

60 Taboulet P, Moreira V, Haas L, Porcher R, Braganca A, Fontaine JP, et al. Triage with the French Emergency Nurses Classification in Hospital scale: reliability and validity. European Journal of Emergency Medicine. 2009;16:61-67.

61 Considine J, LeVasseur SA, Villanueva E. The Australasian Triage Scale: examining emergency department nurses' performance using computer and paper scenarios. Annals of Emergency Medicine. 2004;44:516-523.

62 Gravel J, Gouin S, Bailey B, Roy M, Bergeron S, Amre D. Reliability of a computerized version of the Pediatric Canadian Triage and Acuity Scale. Academic Emergency Medicine. 2007;14:864-869.

63 Gravel J, Gouin S, Manzano S, Arsenault M, Amre D. Interrater Agreement between Nurses for the Pediatric Canadian Triage and Acuity Scale in a Tertiary Care Center. Academic Emergency Medicine. 2008;15:1262-1267.

64 Travers DA, Waller AE, Bowling JM, Flowers D, Tintinalli J. Five-level triage system more effective than three-level in tertiary emergency department. Journal of Emergency Nursing. 2002;28:395-400.

65 Worster A, Gilboy N, Fernandes CM, Eitel D, Eva K, Geisler R, et al. Assessment of inter-observer reliability of two five-level triage and acuity scales: a randomized controlled trial. Canadian Journal of Emergency Medicine. 2004;6:240-245.

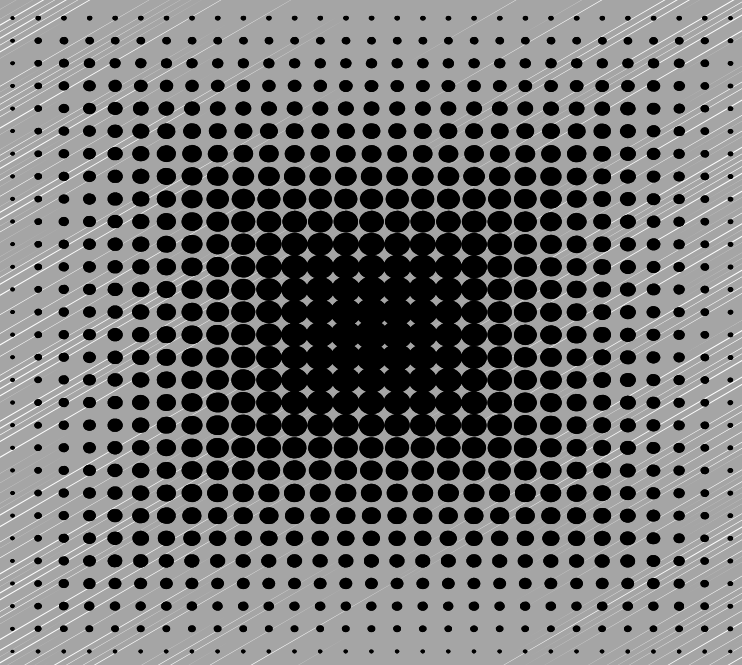
66 Eitel DR, Travers DA, Rosenau AM, Gilboy N, Wuerz RC. The emergency severity index triage algorithm version 2 is reliable and valid. Academic Emergency Medicine. 2003;10:1070-1080.

67 Maningas PA, Hime DA, Parker DE, McMurry TA. The Soterion Rapid Triage System: evaluation of inter-rater reliability and validity. Journal of Emergency Medicine. 2006;30:461-469.

68 Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33:159-174.

Part 2

Validity of emergency department triage systems



Construct validity of the Emergency Severity Index and Manchester Triage System

Based on:
van der Wulp I, Schrijvers A.J.P, van Stel H.F. Predicting admission and mortality with the Emergency Severity Index and the Manchester Triage System: a retrospective observational study. Emergency Medicine Journal. 2009;26:506-509.

ABSTRACT

Objective: To compare the degree to which the Emergency Severity Index (ESI) and the Manchester Triage System (MTS) predict admission and mortality.

Methods: A retrospective observational study of four emergency department (ED) databases was conducted. Patients who presented to the ED between the 1st of January and the 18th of July 2006, and were triaged with the ESI or MTS, were included in the study.

Results: 37974 patients triaged with the ESI and 34258 patients triaged with the MTS were included. The likelihood of admission decreased significantly with urgency categories in both populations. However, this association was larger for patients triaged with the ESI than with the MTS. Mortality rates were small in both populations. Most patients who died in the ED were triaged in the most urgent triage categories of both systems.

Conclusions: Both, the ESI and MTS, predicted admission well. The ESI was a better predictor of admission than the MTS. Mortality was associated with urgency categories of both triage systems.

INTRODUCTION

In recent years, triage systems have frequently been used in emergency departments (EDs). In the Netherlands, the Dutch Institute for Healthcare Improvement (CBO) developed a guideline in 2004 that advised all EDs in the Netherlands to implement the Manchester Triage System (MTS). Creators of the guideline preferred the MTS because it is not diagnosis-based and therefore is particularly applicable for use by nurses[1]. After publication of the guideline, EDs started implementing the MTS. At the end of 2006, 87% of the hospitals that used a triage system, had implemented the MTS, while 13% used the Emergency Severity Index (ESI)[2].

The MTS was developed by the Manchester Triage Group and consists of fifty-two flowcharts, each presenting a complaint. Urgency of the patient's complaint is assessed by discriminators presented in each flowchart. Every patient is classified in one of five urgency categories: red (needs to see a doctor immediately); orange (can wait ten minutes); yellow (can wait one hour); green (can wait two hours); and blue (can wait four hours)[3].

In contrast to the MTS, the ESI consists of one flowchart and patient urgency is assessed with four conceptual key questions. Answers to these four questions will lead to one of five urgency categories: 1 (requires immediate live saving intervention); 2 (a high risk situation); 3 (patient needs two or more resources); 4 (patient needs one resource); and 5 (no resources are needed). The system defines resources as, laboratory tests, radiology, intravenous fluids, specialty consultation, a simple or complex procedure, and intravenous, intramuscular, or nebulised medications[4].

A major question for triage systems is validity. Previously conducted studies were mainly concentrated on relationships between urgency and certain outcome variables such as admission, length of stay, and mortality. These relationships were studied in an attempt to find evidence for the construct validity of triage systems. It is not possible to assess criterion validity as a gold standard for triage urgency is absent[5].

Associations between urgency and admission have been reported in eight studies, of which six studied the ESI[6-11], one the MTS[12] and one the Australasian Triage Scale[13]. Associations with length of stay in the ED and resource use were only found in the ESI[6;8;9;11;14;15]. Associations between urgency categories of triage systems and hospital length of stay, mortality, and survival have been studied less frequently. Hospital length of stay was not associated with the ESI, but six-month survival was[15;16]. Furthermore, the ESI and Australasian Triage Scale showed associations with mortality[10;17]. One study measured the predictive ability of two triage systems, the ESI and the Canadian Triage and Acuity Scale. Admitted patients and patients who died were allocated to a significantly higher urgency category compared to patients who were not admitted or did not die. There were no significant differences between the ability of the two triage systems to predicting admission and in-hospital mortality[18].

In our opinion, the above-mentioned studies are limited by the fact that a large number of studies have focussed on specific subgroups of patients (e.g. elderly, self-referred patients, children or patients below the age of fourteen years). Also, none of these studies adjusted for variables that could have influenced the outcomes. A study was therefore conducted to compare the construct validity of the ESI and the MTS by studying the associations of urgency categories of both systems with hospital admission and mortality in EDs in the Netherlands.

METHODS

Study design
The present study is a retrospective observational study of ED databases. The protocol has been approved by the Medical Ethical Committee of the University Medical Center Utrecht.

Study setting and population
The study was conducted in four hospitals in the Netherlands. The Onze Lieve Vrouwe Gasthuis (Amsterdam) and Sint Elisabeth hospital (Tilburg) implemented the ESI in 2003 and 2004, respectively and the Haaglanden Medical Center (The Hague) and Meander Medical Center (Amersfoort) implemented the MTS in 2004 and 2005, respectively. The EDs of the participating hospitals have an annual number of patients ranging from 30.000 to 47.000. All nurses who triage with the ESI were trained in the use of the system; nurses at the Onze Lieve Vrouwe Gasthuis were trained and tested by one of the founders of the system[19] and, at the Sint Elisabeth Hospital, nurses followed a 1-day course of training by an expert. Nurses who triage with the MTS were trained by in-company training sessions following the guidelines[1]. In 2007, a reliability study was conducted in two hospitals that triage with the MTS, including the Meander Medical Center. Interrater agreement in this study was substantial (quadratically weighted kappa: 0.62)[20].
All patients who presented at the ED and were triaged with the ESI or MTS between the 1st of January and the 18th of July 2006, were included in the study.

Data collection
The ED databases of the four participating hospitals were used to collect information about triage category, gender, age, admission, and mortality of all patients. Hospitalised patients included patients who were admitted to a hospital ward or who were transferred to a ward of another hospital. Mortality was defined as all patients who died in the ED. Patients who were pronounced dead on arrival were excluded from the study.

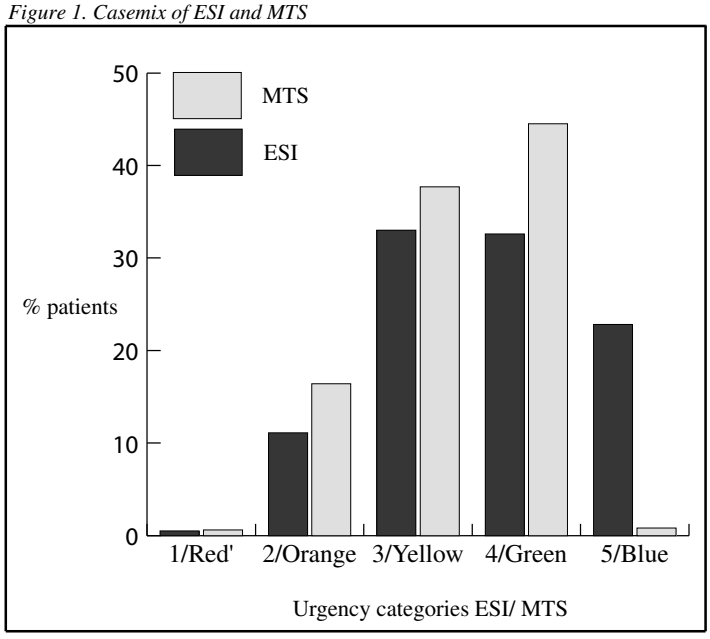
Data analysis
The influence of the ESI and MTS on the prediction of hospital admission was assessed by binary logistic regression analyses. Univariate logistic regression analyses were performed on age, gender, hospital, and urgency for each triage system. All variables which significantly ($p<0.05$) predicted hospital admission were selected for multivariate analyses.
Because mortality rarely occurred in the ED, associations were assessed with the Fisher exact test. This test measures whether the probability of dying is equal in all triage categories. Data were analysed using SPSS for Windows version 14.

RESULTS

In total, 38330 patients presented at the ED of hospitals triaging with the ESI, and 46537 patients presented to the hospitals using the MTS. Of these, 356 patients (0.9%) were excluded because of missing ESI urgency categories and 12279 patients (26.4%) because of missing MTS urgency categories. A total of 37974 patients triaged with the ESI and 34258 patients triaged with the MTS were included in the study.
The mean age (SD) of patients triaged with the ESI and MTS was 38.7 (SD:23.6) years and 42.4 (SD:23.5) years, respectively. In both patient populations, more than half of the patients who presented to the ED were male (56.0% ESI and 51.6% MTS).

Most of the patients triaged with the ESI were triaged in the third (33.0%) or fourth (32.6%) category (figure 1); relatively few patients were triaged in categories 1 (0.5%) and 2 (11.1%). Patients triaged with the MTS were mostly triaged to the yellow (37.7%) or green (44.5%) categories. Categories red and blue rarely occurred (0.6% and 0.8%, respectively).
In both triage systems, triaged patients were admitted more often than patients with missing triage scores ($p<0.05$). There was no significant difference in the mortality rates between these two groups ($p=0.59$). Patients with missing triage scores in the hospitals triaging with the MTS were significantly younger ($p<0.01$) and were more often male ($p<0.01$). In one hospital that triaged with the MTS, patients with missing triage scores were significantly ($p<0.01$) more often referred by the general practitioner or ambulance services. In the other hospitals, this information was not available.

Figure 1. Casemix of ESI and MTS



Admission
The frequency of hospital admission was almost equal in both populations. Of patients triaged with the ESI, 21.6% were admitted compared with 21.0% triaged with the MTS. In both systems, the majority of the admitted patients were triaged in the middle category (table 1). Patients in the more urgent categories were more likely to be admitted (figure 2). Remarkably, 158 patients (1.9%) triaged with the ESI were admitted, while it was estimated that no resources were needed for these patients. In patients triaged with the MTS, ten patients (0.1%) were admitted while it was estimated that they could wait up to four hours in the waiting room.
In univariate logistic regression analyses, age, and hospital were significant predictors for hospital admission ($p<0.01$). Gender was only a significant predictor in the ESI population,

with women having a greater likelihood of admission to hospital ($p<0.01$). The results of the multivariate logistic regression analyses are presented in table 2. The likelihood of admission decreased substantially with urgency in both populations. Also, for each category in both triage systems, no overlap was found in 95% confidence intervals. The likelihood of hospital admission for patients triaged with the ESI was larger compared to patients triaged with the MTS. In the fifth triage category, the likelihood of admission for patients triaged with the ESI was almost three times lower than the reference category.
The same effect was observed in the MTS, but this result appeared to be non-significant, possibly because only ten patients were admitted. In both triage systems, substantial differences were found between hospitals to the degree they predicted hospital admission.

Table 1. Casemix of admitted patients in ESI and MTS

Triage categories	**ESI** % Admission (N=8186)	**MTS** % Admission (N=7063)
ESI 1/ MTS Red	1.8	1.9
ESI 2/ MTS Orange	25.8	35.5
ESI 3/ MTS Yellow	58.0	48.9
ESI 4/ MTS Green	12.3	13.5
ESI 5/ MTS Blue	1.9	0.1
Total	100	100

Table 2. Prediction of admission by ESI and MTS

Independent variables	**ESI** Adjusted OR (95% CI)	**MTS** Adjusted OR (95% CI)
ESI 1/ MTS Red	34.03 (23.83 – 48.59)	21.42 (15.54 – 29.52)
ESI 2/ MTS Orange	13.78 (12.52 – 15.18)	8.56 (7.83 – 9.36)
ESI 3/ MTS Yellow	5.91 (5.47 – 6.39)	4.77 (4.40 – 5.17)
ESI 5/ MTS Blue	0.32 (0.27 – 0.38)	0.54 (0.28 – 1.07)*
Hospital	0.35 (0.33 – 0.37)	0.49 (0.46 – 0.52)
Age	1.02 (1.02 – 1.03)	1.03 (1.03 – 1.04)
Gender (male)	1.06 (1.00 – 1.13)	*

** Not statistically significant (p>0,05)*
*** Reference category 4 is by definition 1*

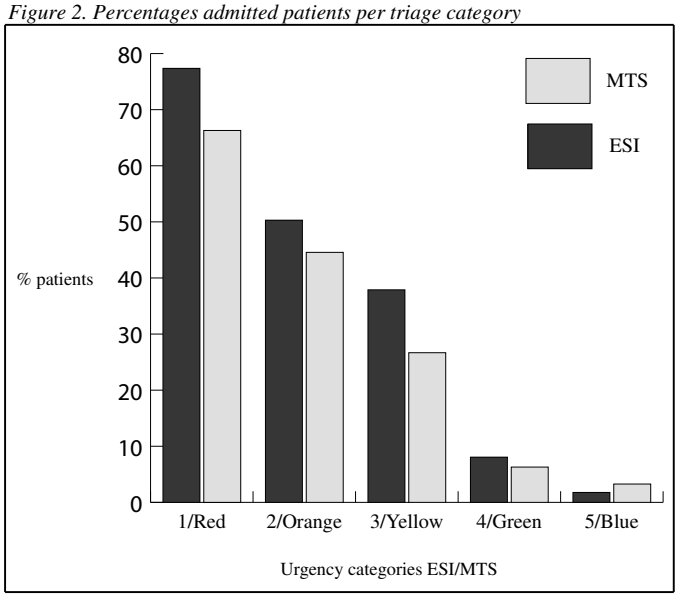
ED mortality

Twenty-eight patients triaged with the ESI and twenty-nine patients triaged with the MTS died in the ED. Of the patients who died, twenty-three (82.1%) were triaged in ESI category 1 and twenty-two (75.9%) were triaged in MTS category red (table 3). A Fischer exact test revealed significant associations ($p<0.01$) between the urgency categories of both triage systems and ED mortality. Patients in the more urgent categories were at increased risk of dying. However, because of the small mortality rates, it is not possible to calculate the strength of the associations.

Table 3.Casemix of patients triaged with the ESI and MTS who died in the ED

Triage categories	**% Mortality ESI**	**% Mortality MTS**
ESI 1/ MTS Red	82.1	75.9
ESI 2/ MTS Orange	10.7	20.7
ESI 3/ MTS Yellow	7.1	3.4
ESI 4/ MTS Green	0.0	0.0
ESI 5/ MTS Blue	0.0	0.0
Total	100	100

Figure 2. Percentages admitted patients per triage category



DISCUSSION

The results of this study showed that urgency, as defined in both triage systems, was a very good predictor of hospital admission and proved to be closely related to ED mortality. A limiting factor for triage research is the fact that there is no gold standard available, for measuring triage urgency. We therefore studied the predictive ability of urgency categories of the ESI and MTS for hospital admission and ED mortality as an indicator of construct validity. By studying this relationship, it was expected that the urgency categories of the ESI and MTS would be associated with these patient outcomes. However, this relation will not be perfect as some patient categories, such as patients with severe pain, will be triaged in the more urgent categories, with or without subsequent hospitalisation. In addition, the patient's condition may deteriorate in the period between triage assessment and medical treatment, which can result in unexpected patient outcomes.

It was found that hospital admission decreased with urgency. The confidence intervals of the urgency categories predicting hospital admission did not overlap, which adds support to the discriminative power of the two systems. The ESI seems to be a better predictor of hospital admission than the MTS, possibly because triage nurses using the ESI estimate the number of resources the patient needs. The number of resources is therefore likely to be associated with hospital admission.

To our knowledge, this is the first study to measure the predictive capacity of two triage systems by means of logistic regression analyses. In both triage systems, significant differences were found between hospitals to the degree they predict admission. It is possible that different admission policies may have accounted for these differences. It is therefore important to conduct future studies in several hospitals so that adjustments can be made for these influences. Previous studies have not adjusted for these factors and have mainly reported the percentage of admitted patients per triage category. Roukema et al.[12]

found an increased rate of hospital admissions with increased urgency in the MTS. Patients triaged in the blue category were less likely to be admitted (0.9%) than those triaged in the red category (53.5%). Although the percentage of patients admitted per triage category of the MTS was higher in the present study than that reported by Roukema et al, the present study showed the same effect. Differences in study populations could explain these differences, because age was a significant predictor of urgency. The percentage of patients admitted per triage category in the ESI was also comparable to other studies. Tanabe et al.[7] and Wuerz et al.[9;16] reported that 80% to 92% of the patients in the first category were admitted and 0% to 5% in the last category. Again, the differences could be explained by differences in the study populations and study setting.
The probability of dying in the ED is very small, but was significantly associated with urgency categories in both triage systems. Most patients who died in the ED were triaged in the highest urgency category. None of the patients who died were triaged in the two least urgent categories. A comparable effect has been reported by Wuerz et al.[16] however, they included a follow-up period of six months after an ED visit. In that study, 32% of the patients triaged in the first category died within six months after their ED visit. No patients died in the ED. The difference in design could explain the difference in mortality rate in the present study.
Overall, compared with the present study, several other studies have reported more patients in the second triage category of the ESI and fewer patients in the fifth category[6;9;11;15]. The above mentioned explanations, such as differences between hospitals, could also explain differences in reported casemix of the ESI.

Limitations
Our study has some limitations and therefore the results should be interpreted carefully. First, the retrospective character of the study. Because information was registered for other purposes than this study, it is possible that not all information was entered correctly. This could have influenced our results. To minimise the limitations of retrospective studies, we have selected a study period in which no organisational changes in the ED occurred. A second limitation is the large number of missing triage scores from hospitals that triage with the MTS. It is possible that patients who arrived by ambulance were frequently not triaged because they needed care immediately. Furthermore, during a busy period in the ED, the triage nurse may stop registering triage scores or stop triaging and start helping colleagues. In one hospital, patients with missing triage scores were more often referred. This could indicate that triage nurses do not triage these patients because they had already been seen by a professional. The missing data could have influenced the casemix and therefore affected the associations between the MTS and hospital admission.

Conclusions
Both, the ESI and MTS, strongly predicted hospital admission. The ESI was found to be a better predictor of hospital admission than the MTS. Mortality was associated with urgency categories in both triage systems. However, further study in a different design is needed to measure the strength of the association. Future studies, measuring the association between triage systems and other parameters, need to adjust for influencing factors.

REFERENCES

1 Coenen IGA, Hagemeijer A, de Caluwé R, de Voeght FJ, Jochems PJJ. Guideline Triage on the emergency department. Dutch Institute for Healthcare Improvement; 2004.

2 van Baar ME, Giesen P, Grol R, Schrijvers AJP. Results of a preliminary study to Emergency Medicine. Julius Center for Health Sciences and Primary Care & Center for Quality of Care Research WOK; 2007.

3 Mackway-Jones K. Emergency Triage. 2 ed. London: Manchester Triage Group; 2005.

4 Gilboy N, Tanabe P, Travers D, Rosenau A, Eitel DR. Emergency Severity Index, Version 4: Implementation Handbook. Rockville: Agency for Healthcare Research and Quality; 2005.

5 Streiner DL, Norman GT. Validity. Health Measurement Scales: a practical guide to their development and use. 3th ed. Oxford: Oxford University Press; 2003. p. 172-193.

6 Baumann MR, Strout TD. Triage of geriatric patients in the emergency department: validity and survival with the Emergency Severity Index. Annals of Emergency Medicine. 2007;49:234-240.

7 Tanabe P, Gimbel R, Yarnold PR, Kyriacou DN, Adams JG. Reliability and validity of scores on the Emergency Severity Index version 3. Academic Emergency Medicine. 2004;11:59-65.

8 Wuerz RC, Milne LW, Eitel DR, Travers D, Gilboy N. Reliability and validity of a new five-level triage instrument. Academic Emergency Medicine. 2000;7:236-242.

9 Wuerz RC, Travers D, Gilboy N, Eitel DR, Rosenau A, Yazhari R. Implementation and refinement of the emergency severity index. Academic Emergency Medicine. 2001;8:170-176.

10 Eitel DR, Travers DA, Rosenau AM, Gilboy N, Wuerz RC. The emergency severity index triage algorithm version 2 is reliable and valid. Academic Emergency Medicine. 2003;10:1070-1080.

11 Baumann MR, Strout TD. Evaluation of the Emergency Severity Index (version 3) triage algorithm in pediatric patients. Academic Emergency Medicine. 2005;12:219-224.

12 Roukema J, Steyerberg EW, van Meurs A, Ruige M, van der Lei J, Moll HA. Validity of the Manchester Triage System in paediatric emergency care. Emergency Medicine Journal. 2006;23:906-910.

13 van Gerven R, Delooz H, Sermeus W. Systematic triage in the emergency department using the Australian National Triage Scale: a pilot project. European Journal of Emergency Medicine. 2001;8:3-7.

14 Elshove-Bolk J, Mencl F, van Rijswijck BT, Simons MP, van Vugt AB. Validation of the Emergency Severity Index (ESI) in self-referred patients in a European emergency department. Emergency Medicine Journal. 2007;24:170-174.

15 Tanabe P, Gimbel R, Yarnold PR, Adams JG. The Emergency Severity Index (version 3) 5-level triage system scores predict ED resource consumption. Journal of Emergency Nursing. 2004;30:22-29.

16 Wuerz RC. Emergency severity index triage category is associated with six-month survival. ESI Triage Study Group. Academic Emergency Medicine. 2001;8:61-64.

17 Dent A, Rofe G, Sansom G. Which triage category patients die in hospital after being admitted through the emergency department? A study in one teaching hospital. Emergency Medicine. 1999;11:68-71.

18 Worster A, Fernandes CM, Eva K, Upadhye S. Predictive validity comparison of two five-level triage acuity scales. European Journal of Emergency Medicine. 2007;14:188-192.

19 Gilboy N, Schoenaker E, Ramaker Y, Huysinga H. Implementation of triage on a emergency department in Amsterdam *(Implementatie van triage op een Amsterdamse SEH)*. Triage. 2003;5,16-22.

20 van der Wulp I, van Baar ME, Schrijvers AJP. Reliability and validity of the Manchester Triage System in a general emergency department patient population in the Netherlands: Results of a simulation study. Emergency Medicine Journal. 2008;25:431-434.

7

Content validity of the Emergency Severity Index

Based on:
van der Wulp I, Sturms L.M, Schrijvers A.J.P, van Stel H.F. An observational study of patients triaged in Emergency Severity Index category 5. European Journal of Emergency Medicine. *In press.*

ABSTRACT

Objective: Patients triaged in the fifth category of the Emergency Severity Index (ESI) do not need any resources before discharge from the emergency department (ED). The characteristics of these patients were studied with a focus on those, who were admitted or sent to the outpatient department after their ED visit.

Methods: A retrospective observational study was conducted on 117740 patient presentations. Patients were included in the study when they were triaged with the ESI and presented to one of the two EDs under study between the 1st of September 2004 and the 1st of June 2006.

Results: Overall, 22.2% of the patients were triaged in ESI category 5. Patients aged less than twenty-one years, women, and self-referred patients were most likely triaged in ESI category 5, as well as patients presenting with complaints such as 'check-up appointments at the ED' and 'complaints of the skin and subcutaneous tissue'. Patients triaged in ESI category 5 who were admitted or sent to the outpatient department were most likely elderly (aged above sixty-five years) and referred patients. They were also more likely to present with complaints such as 'post operative complications, wound care problems and plaster problems' and 'complaints of the genitourinary system'.

Conclusions: Although younger patients and women were more likely triaged in ESI category 5, patients within this category who were admitted or sent to the outpatient department were more likely elderly and referred patients. Being admitted or sent to the outpatient department and triaged in ESI category 5 indicates undertriage. Revision of the system's guidelines is required to properly account for these patient groups.

INTRODUCTION

The Emergency Severity Index (ESI) is a five-level triage system that has been implemented in several emergency departments (EDs) in, for example, the United States, Korea, and the Netherlands[1-4]. Patients requiring immediate life-saving interventions such as intubation are triaged in the most urgent triage category of the system (ESI 1). The second category (ESI 2) contains patients who are at high risk for deterioration. In addition, patients in severe pain, distress, or with changes in level of consciousness are triaged in ESI 2. Less severely injured or ill patients are triaged in the remaining three triage categories by interpreting the patient's vital signs or estimating the number of resources a patient needs. In the triage system, resources are defined as: specialty consultation, radiology, several simple and complex procedures, laboratory tests, intravenous fluids and medications, intramuscular and nebulized medications[3]. The ESI is the first system in which nonurgent patients are triaged by estimating the number of resources they require. Other frequently used five-level triage systems such as the Manchester Triage System (MTS), the Canadian Triage and Acuity Scale (CTAS) and the Australasian Triage Scale (ATS) estimate the maximum waiting time to see a doctor[5-7].

A number of studies have evaluated the construct validity and reliability of the ESI. These studies found associations among urgency categories and hospital admission, ED length of stay, mortality, resource use and survival time after an ED visit[1;8-15]. Associations between hospital admission and mortality were also found in the MTS, the CTAS and the ATS[12;16-19]. The reliability of the ESI has been reported to be moderate to almost perfect[8;10;13;14;20-23]. This is comparable with the reliability of the CTAS, which was reported to be fair to almost perfect[24-33]. The reliability of the ATS and MTS ranged between fair and substantial. However, the number of studies that reported the reliability of these systems with a kappa statistic is small[18;34-36].

Despite the fact that previous studies have indicated that the ESI is a reliable and valid triage system for use in the ED, several studies reported that up to 6.2% of the patients triaged in ESI 5 are admitted to the hospital[1;4;10;14;20]. Possibly these patients were undertriaged because the nurse estimated that they did not require any resources for discharge from the ED, while admission to the hospital minimally requires a specialty consultation. These patients seem to belong in the ED, whereas the ESI implementation handbook describes that patients triaged in the fourth and fifth category of the ESI can be served in, for example, a fast track area[3]. This also applies for patients who were sent to the outpatient department after their ED visit. No literature was found that studied the frequency of this event. The objective of this study was first to describe the patient characteristics that were associated with being triaged in ESI 5, and second, within this triage category to focus on patients who were admitted or sent to the outpatient department. The results of this study were used to propose suggestions for further refinement of the ESI.

METHODS

Study design

This study is a retrospective observational study of ED databases. The protocol has been approved by the Medical Ethical Committee of the University Medical Center Utrecht.

Study setting and population

The study was conducted in two hospitals in the Netherlands. Both EDs implemented the ESI in 2004. The nurses of both EDs have been trained into the use of the system[2].

One of the hospitals was the first to implement the ESI in the Netherlands. They were trained by one of the founders of the system and developed a Dutch training program. The ESI training consists of six hours of theory combined with practical training in the ED.

One of the hospitals is a designated trauma center. Severely injured patients in the province are transported to the ED of this hospital. Both hospitals have an annual number of ED visitors between 27.000 and 42.000. In this study, all patients who presented to the ED and were triaged with the ESI between the 1st of September 2004 and the 1st of June 2006 were included.

Data collection

ED databases from both hospitals were used to collect information on: triage category, gender, age, discharge destination, patient complaint, type of referrer, and time of presentation. Discharge destination included admission to the hospital, discharge home with or without an appointment to the outpatient department, transfer to another healthcare institution, and a remaining category. This remaining category consisted of patients who were already admitted and returned to their hospital ward and patients who died in the ED. Referrer type consisted of patients referred by the ambulance service, a general practitioner, patients without a referral, and a remaining category. Patients in the remaining category were referred by a specialist or the police department, or were recalled by the ED because their diagnosis was missed during an earlier visit.

Patient complaints were coded according to the chapters of the International Classification of Diseases, 10th edition (ICD10)[37]. To fit the data, some modifications to this classification were made. The category diseases of the ear and mastoid process, was adjusted to diseases of the ear, nose, and throat. Furthermore, we added the following three categories: 'non traumatic bleedings', 'check-up appointment at the ED' and 'post operative complications, wound care problems and plaster problems' (Appendix B). This latter category contained patients who, for example, had had minor surgery and were discharged from the hospital the same day.

Data analysis

To describe patient characteristics in ESI category 5, an ordinal logistic regression analysis should be performed. As parametric assumptions were not met ($p<0.05$), the data was analyzed with a multinomial logistic regression analysis, which is an extension of binary logistic regression analysis. Multinomial logistic regression analysis produces odds ratios for n-1 categories of the dependent variable with one category as reference. Univariate multinomial logistic regression analysis was performed on the variables age, gender, time of arrival, complaint, hospital, and type of referrer. In these analyses, ESI categories 1 and 2 were merged because of the small number of patients triaged in ESI category 1 (n=635), and were used as reference category (n=14373) in the analyses. All variables which significantly predicted urgency ($p<0.05$) were selected for multivariate analysis.

The characteristics of ESI category 5 patients who were admitted to hospital or sent to the outpatient department were assessed by binary logistic regression analysis. Univariate logistic regression analysis was performed on the variables age, gender, hospital, patient complaint, type of referrer, and time of arrival in the ED. In these analyses, patients in ESI category 5 who were not admitted to hospital or who did not receive an appointment to the

outpatient department were taken as reference category. Variables that significantly ($p<0.05$) predicted hospital admission and outpatient department follow-up were selected for multivariate analysis. Data were analyzed using SPSS for Windows version 14.

RESULTS

In total, 119219 patients presented to the EDs of both hospitals. As a result of missing urgency categories, 1479 patients (1.2%) were excluded from the study, leaving 117740 patients for analyses. Excluded patients were slightly younger (median age: 33.0 vs. 37.0 years, $p<0.01$) and were more often male (58.6% vs. 55.8%, $p= 0.03$).
Patient characteristics and casemix are described in table 1. The majority of the patients referred themselves to the ED (61.3%), of which 32.3% were triaged in ESI category 5. In total, almost a quarter (22.2%) of the patients were triaged in ESI category 5. These patients were significantly younger compared to patients triaged in other triage categories (median age: 28.0 vs. 40.0 years, $p<0.001$). The most frequently occurring complaints in ESI category 5 were: 'injuries and certain other consequences of external causes' (58.2%); 'symptoms, signs, and abnormal clinical and laboratory findings not elsewhere classified' (6.6%); and 'complaints of the ear, nose, and throat' (5.4%). In univariate multinomial logistic regression analyses, the variables age, gender, time of presentation, complaint, hospital, and referrer were significant ($p<0.01$) predictors of ESI category 5. Results of the multivariate multinomial logistic regression analysis are presented in table 2. Patients below the age of twenty-one years, women, patients presenting at the ED during daytime (9:00 a.m. to 4:00 p.m.), and self-referred patients were more likely triaged in ESI category 5. Patient complaint showed the strongest associations with urgency. Patient complaints with regard to 'check up appointments at the ED', 'complaints of the skin and subcutaneous tissue', and 'complaints of the ear, nose, and throat' were most likely triaged in ESI category 5. In contrast, complaints such as 'complaints of the circulatory system' and 'diseases of blood and blood forming organs' were more likely to be triaged in the urgent triage categories (ESI categories 1 and 2).

Table 1. Patient characteristics by triage category

N=117740	ESI 1	ESI 2	ESI 3	ESI 4	ESI 5
Casemix	635	13738	38413	38823	26131
Age (median)	57.0	47.0	49.0	31.0	28.0
% male	68.8	57.4	52.2	56.9	58.5
% referred*	87.4	54.3	55.6	23.0	6.8
% admission	74.9	49.9	38.7	8.3	2.0
% outpatient department	0.5	16.0	27.4	26.8	10.3

** Patients referred by a general practitioner or ambulance service*

Table 2. Associations between urgency categories and patient characteristics

Predictor	OR Category 5 (95% CI)
Age (years)	
0-12	1.09 (1.01 – 1.18)
13-20	1.20 (1.09 – 1.32)
21-39	Reference
40-64	0.61 (0.58 – 0.65)
65+	0.33 (0.30 – 0.36)
Gender	
Male	0.91 (0.86 – 0.95)
Female	Reference
Time of presentation	
1:00 AM – 8:00 AM	0.84 (0.78 – 0.90)
9:00 AM – 4:00 PM	1.38 (1.31 – 1.45)
5:00 PM – 12:00 PM	Reference
Complaint	
Check up appointment at the ED	11.29 (8.12 – 15.70)
Complaints of the skin and subcutaneous tissue	5.86 (3.79 – 9.05)
Complaints of the ear, nose and throat	3.22 (2.60 – 3.98)
Post operative complications, wound care problems and plaster problems	2.84 (1.98 – 4.08)
Complaints of the eye	2.44 (2.01 – 2.95)
Complaints of the musculoskeletal system and connective tissue	1.56 (1.37 – 1.77)
Non traumatic bleedings	0.62 (0.47 – 0.83)
Symptoms, signs and abnormal clinical and laboratory findings not elsewhere classified	0.62 (0.57 – 0.68)
Complaints of the nervous system	0.52 (0.47 – 0.57)
Oncology related complaints	0.35 (0.13 – 0.92)
Complaints of the digestive system	0.29 (0.27 – 0.32)
Complaints of the respiratory system	0.18 (0.16 – 0.20)
Complaints of the genitourinary system	0.17 (0.14 – 0.20)
Mental and behavioral disorders	0.16 (0.12 – 0.21)
Diseases of blood and blood forming organs	0.07 (0.02 – 0.32)
Complaints of the circulatory system	0.04 (0.04 – 0.05)
Injuries and certain other consequences of external causes	Reference
Hospital	1.25 (1.18 – 1.32)
Referrer type	
Referred	0.09 (0.08 – 0.09)
Remaining	0.39 (0.34 – 0.44)
Self referred	Reference

Variables are statistically significant (p<0.05);
Nagelkerke R square: 0.40;
ESI 1 and 2 are overall reference category

Outcomes in ESI 5

Of the 26131 patients triaged in ESI category 5, the majority (83.9%) was sent home, 10.0% received an appointment for the outpatient department, and 1.9% were admitted to the hospital. Within the group of patients triaged in ESI category 5, patients who were admitted to the hospital or sent to the outpatient department were slightly older compared to other patients triaged in ESI category 5 (median age: 33.0 vs. 28.0 years, $p<0.01$). They also attended the ED more often during the daytime (9:00 a.m. to 4:00 p.m.) (56.8% vs. 46.5%, $p<0.01$) and most frequently presented with 'injuries and certain other consequences of external causes' (35.9%) or 'check-up appointments at the ED' (15.7%).

In univariate logistic regression analyses, the variables age, hospital, patient complaint, referrer type, and time of presentation were significant predictors ($p<0.05$) for being admitted to the hospital or sent to the outpatient department while triaged in ESI category 5. Results of the multivariate logistic regression analysis are presented in table 3. Variables

referrer type, complaint, and hospital were highly associated with hospital admission and outpatient department follow-up. It seemed that patients referred by a healthcare professional were most likely to leave the ED with these outcomes. In addition, patients presenting with 'post operative complications, wound care problems, and plaster problems', 'complaints of the genitourinary system', and 'check-up appointments at the ED' were more likely to be admitted to the hospital or sent to the outpatient department.
Comparisons of tables 2 and 3 show some interesting findings. Patients presenting with complaints such as 'complaints of the genitourinary system', 'complaints of the digestive system', and 'symptoms, signs, and abnormal clinical and laboratory findings', were more likely to be triaged in the urgent triage categories of the ESI. Furthermore, patients presenting with these complaints and who were subsequently triaged in ESI category 5 were more likely to be admitted to the hospital or sent to the outpatient department. The same trend was observed for the variables age and referrer type. Patients below the age of forty years and self-referred patients were more likely to be triaged in ESI category 5, while older (aged above sixty-five years) and referred patients triaged in ESI category 5 were most likely to be admitted to the hospital or sent to the outpatient department.

Table 3. Characteristics of ESI category 5 patients admitted to hospital or sent to the outpatient department

Predictor	**Admitted/outpatient department follow up OR (95% CI)**
Age (years)	
0-12	0.99 (0.87 – 1.12)**
13-20	0.96 (0.83 – 1.11)**
21-39	Reference
40-64	1.40 (1.26 – 1.56)
65+	1.90 (1.61 – 2.25)
Time of presentation	
1:00 AM – 8:00 AM	0.99 (0.85 – 1.14)**
9:00 AM – 4:00 PM	1.13 (1.03 – 1.24)
5:00 PM – 12:00 PM	Reference
Complaint	
Post operative complications, wound care problems and plaster problems	4.69 (3.51 – 6.28)
Complaints of the genitourinary system	3.72 (2.58 – 5.37)
Check up appointment at the ED	3.40 (2.86 – 4.05)
Complaints of the digestive system	2.77 (2.30 – 3.34)
Symptoms, signs and abnormal clinical and laboratory findings not elsewhere classified	2.59 (2.22 – 3.03)
Non traumatic bleedings	2.51 (1.56 – 4.05)
Complaints of the musculoskeletal system and connective tissue	2.20 (1.86 – 2.60)
Complaints of the skin	1.92 (1.41 – 2.61)
Complaints of the eye	1.56 (1.27 – 1.92)
Complaints of the nervous system	1.54 (1.27 – 1.88)
Complaints of the ear, nose and throat	1.45 (1.17 – 1.79)
Complaints of the respiratory system	1.35 (1.00 – 1.83)
Injuries and certain other consequences of external causes	Reference
Hospital	3.51 (3.20 – 3.87)
Referrer type	
Self referred	Reference
Referred	3.87 (3.42 – 4.39)
Remaining	5.17 (4.35 – 6.16)

Variables are statistically significant ($p<0.05$)
*** Not statistically significant ($p>0.05$)*

DISCUSSION

In this study, the characteristics of patients triaged in the fifth category of the ESI were described. Almost a quarter of the total ED population was triaged in this category, which is higher than reported in previous studies. These reported percentages between 5% and 14.1%[4;8;10;14;21;38]. In the present study it was found that self-referral to the ED was strongly associated with ESI category 5. A comparable association was reported by Elshove-Bolk et al.[1]. They reported that 54% of the self-referred patients were triaged in ESI category 5. In the present study, the majority of the population referred themselves to the ED, which resulted in a relatively high percentage of patients triaged in ESI category 5. Other patient characteristics associated with ESI category 5 were women, younger patients, and patients presenting with complaints such as 'check-up appointments at the ED', 'complaints of the skin and subcutaneous tissue', and 'complaints of the ear, nose, and throat'.

Within the group of patients triaged in category 5 of the ESI, we also studied the characteristics of patients who were admitted to the hospital or sent to the outpatient department for follow-up. This situation occurred in 11.9% of the patients triaged in ESI category 5. It was found that these patients were older and more likely referred by a general practitioner or ambulance service compared to other patients triaged in ESI category 5. They were also more likely to present with complaints such as 'post operative complications, wound care problems, and plaster problems', 'complaints of the genitourinary system', and 'check-up appointments at the ED'.

Although it has been reported that the ESI is valid in the elderly population, the present study shows that triage of the elderly can be difficult[9]. This was also found in two earlier studies on the MTS and a four-level triage system[36;39]. Possible explanations according to the authors were that elderly frequently present with multiple and or atypical complaints, which can make it more difficult to estimate the patient's urgency. In addition, the authors mentioned practical difficulties at triage, such as communication problems because of cognitive impairment or hearing disorders. Except for the elderly, the ESI also seems to underestimate the urgency of referred patients, and patients with check-up appointments, genitourinary complaints, or complaints related to an earlier visit to the hospital or ED. A possible solution for this could be to add new rules to the system's guidelines and to focus on these patient groups in training sessions. This will make the triage nurse more alert to these groups of patients. For example, patients referred by a general practitioner or ambulance service should at least be triaged in ESI category 4, because the urgency of these patients has already been assessed by a health professional. In contrast, one must be cautious with interpreting these outcomes because the patient's condition can deteriorate between triage and receiving treatment. Therefore, further study in a prospective design is required.

The present study indicates that a substantial number of patients were undertriaged, because several patient characteristics in ESI category 5 were strongly associated with hospital admission and outpatient department follow-up. Undertriage in the ED can lead to adverse outcomes for admitted patients during their stay in the hospital. Parkhe et al.[40] reported that patients who were initially admitted to a hospital ward and later transferred to the intensive care unit were more likely to have been triaged in the less urgent triage categories of the ATS. As a result, these patients had an increased risk of mortality within thirty days of admission to the intensive care unit compared with patients who were directly sent from the ED to the intensive care unit.

One limitation of this study was its retrospective character. It is possible that not all information was entered correctly in the database, which could have influenced the results. A second limitation is that patient complaints could only be coded according to the chapters of the ICD 10, which is rather generic. Therefore, it is not possible to determine more precisely which complaints occurred in ESI category 5, or which complaints caused undertriage in this category. As a consequence, it is difficult to determine the exact shortcomings of the ESI. Another limitation is neither hospital originally coded patient complaints into the ICD 10. Therefore, the most likely ICD 10 code derived from the complaint descriptions in the data was chosen. Finally, 1.2% of the triage scores were missing. Although the number of patients missing is small, it could have biased our findings. It is possible that these patients comprised urgent patients (ESI categories 1 or 2), because they are treated immediately and the registration of urgency was forgotten. However, these patients could as well have been triaged in ESI category 5. Whether an overestimation or underestimation of the results is to be expected cannot be determined because the characteristics (gender and age) of the missing patients do not give an indication of the severity of the urgency.

In conclusion, patients aged below twenty-one years, women, patients who visited the ED for check-up appointments, complaints of the skin and subcutaneous tissue, and complaints of the ear, nose, and throat were more likely triaged in ESI category 5. Of the patients initially triaged in ESI category 5, 11.9% were admitted to the hospital or sent to the outpatient department. Patient characteristics associated with these outcomes in ESI category 5 were elderly patients, referred patients, and patients presenting with 'post operative complications, wound care problems and plaster problems', 'check-up appointments at the ED', and 'complaints of the genitourinary system'. This indicates a structural flaw of the system. As a result, the system's guidelines need revisions in a manner that accounts for these patient groups. In addition, training sessions should focus on these groups to make triages nurses more aware of these patients.

REFERENCES

1 Elshove-Bolk J, Mencl F, van Rijswijck BT, Simons MP, van Vugt AB. Validation of the Emergency Severity Index (ESI) in self-referred patients in a European emergency department. Emergency Medicine Journal. 2007;24:170-174.

2 Gilboy N, Schoenaker E, Ramaker Y, Huysinga H. Implementation of triage on a emergency department in Amsterdam *(Implementatie van triage op een Amsterdamse SEH)*. Triage. 2003;5:16-22.

3 Gilboy N, Tanabe P, Travers D, Rosenau A, Eitel DR. Emergency Severity Index, Version 4: Implementation Handbook. Rockville: Agency for Healthcare Research and Quality; 2005.

4 Kim TG, Cho JK, Kim SH, Lee HS, Gu HD, Chung SW. Reliability and validity of the modified Emergency Severity Index-2 ad a triage tool. Journal of the Korean Society of Emergency Medicine. 2006;17:154-164.

5 Australasian College for Emergency Medicine. Policy Document: The Australasian Triage Scale; 2000.

6 Beveridge R, Clarke B, Janes L, Savage N, Thompson J, Dodd G, et al. Guidelines for the Canadian Emergency Department Triage & Acuity Scale (CTAS); 1998.

7 Mackway-Jones K. Emergency Triage. 2 ed. London: Manchester Triage Group; 2005.

8 Baumann MR, Strout TD. Evaluation of the Emergency Severity Index (version 3) triage algorithm in pediatric patients. Academic Emergency Medicine. 2005;12:219-224.

9 Baumann MR, Strout TD. Triage of geriatric patients in the emergency department: validity and survival with the Emergency Severity Index. Annals of Emergency Medicine. 2007;49:234-240.

10 Tanabe P, Gimbel R, Yarnold PR, Kyriacou DN, Adams JG. Reliability and validity of scores on The Emergency Severity Index version 3. Academic Emergency Medicine. 2004;11:59-65.

11 Tanabe P, Gimbel R, Yarnold PR, Adams JG. The Emergency Severity Index (version 3) 5-level triage system scores predict ED resource consumption. Journal of Emergency Nursing. 2004;30:22-29.

12 van der Wulp I, Schrijvers AJP, van Stel HF. Predicting admission and mortality with the Emergency Severity Index and the Manchester Triage System: a retrospective observational study. Emergency Medicine Journal. 2009;26:506-509.

13 Wuerz RC, Milne LW, Eitel DR, Travers D, Gilboy N. Reliability and validity of a new five-level triage instrument. Academic Emergency Medicine. 2000;7:236-242.

14 Wuerz RC, Travers D, Gilboy N, Eitel DR, Rosenau A, Yazhari R. Implementation and refinement of the emergency severity index. Academic Emergency Medicine. 2001;8:170-176.

15 Wuerz RC. Emergency severity index triage category is associated with six-month survival. ESI Triage Study Group. Academic Emergency Medicine. 2001;8:61-64.

16 Martins HM, Cuna LM, Freitas P. Is Manchester (MTS) more than a triage system? A study of its association with mortality and admission to a large Portuguese hospital. Emergency Medicine Journal. 2009;26:183-186.

17 Roukema J, Steyerberg EW, van Meurs A, Ruige M, van der Lei J, Moll HA. Validity of the Manchester Triage System in paediatric emergency care. Emergency Medicine Journal. 2006;23:906-910.

18 van Gerven R, Delooz H, Sermeus W. Systematic triage in the emergency department using the Australian National Triage Scale: a pilot project. European Journal of Emergency Medicine. 2001;8:3-7.

19 Worster A, Fernandes CM, Eva K, Upadhye S. Predictive validity comparison of two five-level triage acuity scales. European Journal of Emergency Medicine. 2007;14:188-192.

20 Eitel DR, Travers DA, Rosenau AM, Gilboy N, Wuerz RC. The emergency severity index triage algorithm version 2 is reliable and valid. Academic Emergency Medicine. 2003;10:1070-1080.

21 Travers DA, Waller AE, Bowling JM, Flowers D, Tintinalli J. Five-level triage system more effective than three-level in tertiary emergency department. Journal of Emergency Nursing. 2002;28:395-400.

22 Vance J, Sprivulis P. Triage nurses validly and reliably estimate emergency department patient complexity. Emergency Medicine Australasia. 2005;17:382-386.

23 Worster A, Sardo A, Eva K, Fernandes CM, Upadhye S. Triage Tool Inter-rater Reliability: A Comparison of Live Versus Paper Case Scenarios. Journal of Emergency Nursing. 2007;33:319-323.

24 Bergeron S, Gouin S, Bailey B, Amre DK, Patel H. Agreement among pediatric health care professionals with the pediatric Canadian triage and acuity scale guidelines. Pediatric Emergency Care. 2004;20:514-518.

25 Beveridge R, Ducharme J, Janes L, Beaulieu S, Walter S. Reliability of the Canadian emergency department triage and acuity scale: interrater agreement. Annals of Emergency Medicine. 1999;34:155-159.

26 Dong SL, Bullard MJ, Meurer DP, Colman I, Blitz S, Holroyd BR, et al. Emergency triage: comparing a novel computer triage program with standard triage. Academic Emergency Medicine. 2005;12:502-507.

27 Dong SL, Bullard MJ, Meurer DP, Blitz S, Ohinmaa A, Holroyd BR, et al. Reliability of computerized emergency triage. Academic Emergency Medicine. 2006;13:269-275.

28 Goransson K, Ehrenberg A, Marklund B, Ehnfors M. Accuracy and concordance of nurses in emergency department triage. Scandinavian Journal of Caring Sciences. 2005;19:432-438.

29 Grafstein E, Innes G, Westman J, Christenson J, Thorne A. Inter-rater reliability of a computerized presenting-complaint-linked triage system in an urban emergency department. Canadian Journal of Emergency Medicine. 2003;5:323-329.

30 Gravel J, Gouin S, Bailey B, Roy M, Bergeron S, Amre D. Reliability of a computerized version of the Pediatric Canadian Triage and Acuity Scale. Academic Emergency Medicine. 2007;14:864-869.

31 Gravel J, Gouin S, Manzano S, Arsenault M, Amre D. Interrater Agreement between Nurses for the Pediatric Canadian Triage and Acuity Scale in a Tertiary Care Center. Academic Emergency Medicine. 2008;15:1262-1267.

32 Manos D, Petrie DA, Beveridge RC, Walter S, Ducharme J. Inter-observer agreement using the Canadian Emergency Department Triage and Acuity Scale. Canadian Journal of Emergency Medicine. 2002;4:16-22.

33 Worster A, Gilboy N, Fernandes CM, Eitel D, Eva K, Geisler R, et al. Assessment of inter-observer reliability of two five-level triage and acuity scales: a randomized controlled trial. Canadian Journal of Emergency Medicine. 2004;6:240-245.

34 Considine J, LeVasseur SA, Villanueva E. The Australasian Triage Scale: examining emergency department nurses' performance using computer and paper scenarios. Annals of Emergency Medicine. 2004;44:516-523.

35 Gerdtz MF, Bucknall TK. Influence of task properties and subjectivity on consistency of triage: a simulation study. Journal of Advanced Nursing. 2007;58:180-190.

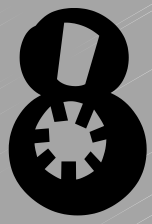
36 van der Wulp I, van Baar ME, Schrijvers AJP. Reliability and validity of the Manchester Triage System in a general emergency department patient population in the Netherlands: Results of a simulation study. Emergency Medicine Journal. 2008;25:431-434.

37 World Health Organization. International statistical classification of diseases and health related problems, tenth revision. World Health Organization; 2005.

38 Daniels JH. Outcomes of Emergency Severity Index five-level triage implementation. Advanced Emergency Nursing Journal. 2007;29:58-67.

39 Rutschmann OT, Chevalley T, Zumwald C, Luthy C, Vermeulen B, Sarasin FP. Pitfalls in the emergency department triage of frail elderly patients without specific complaints. Swiss Medical Weekly. 2005;135:145-150.

40 Parkhe M, Myles PS, Leach DS, Maclean AV. Outcome of emergency department patients with delayed admission to an intensive care unit. Emergency Medicine (Fremantle). 2002;14:50-57.



Content validity of the Manchester Triage System

Based on:
van der Wulp I, Sturms L.M, de Jong A, Schot-Balfoort M, Schrijvers A.J.P, van Stel H.F. Pain assessments at triage with the Manchester Triage System: a prospective observational study. *Submitted*

ABSTRACT

Objective: Pain is an important discriminator of the Manchester Triage System (MTS). We studied the frequency of pain assessments conducted at triage with the MTS and patient, nurse, and triage characteristics associated with pain assessments. Also, nurses' reasons for not assessing pain at triage were studied.

Methods: Nurses from two emergency departments (EDs) registered the process of triage for every presenting patient during one week in May 2009. Nurse characteristics were registered on a second registration form. Thirteen nurses were interviewed about reasons for not assessing pain at triage.

Results: According to the MTS guidelines, pain assessments should have been conducted in 86.1% of the patient presentations. It was only assessed in 32.2% of these patients. Characteristics associated with conducting pain assessments were children below twelve years of age, patients referred by others than a general practitioner or the ambulance service, intake of medication prior to ED visit, experience of the nurse with the MTS, and the duration of triage. Reasons for not assessing pain according to the guidelines included the feeling that pain assessments result in overtriage.

Conclusions: Pain assessments at triage are conducted infrequently because of a lack of clarity in the MTS guidelines, insufficient education, and conducting activities at triage that are not necessary for estimating urgency. Changes in these areas are necessary to improve the reliability and validity of pain assessments and the MTS.

INTRODUCTION

The Manchester Triage System (MTS) has been implemented in emergency departments (EDs) since 1996[1]. It is predominantly used in EDs in the United Kingdom, Portugal and the Netherlands[1-3]. The urgency of patients presenting to the ED is assessed by a triage nurse who selects a flowchart that represents the patient's complaint. The triage system consists of fifty-two flowcharts with which patients are allocated on the basis of discriminators into one out of five triage categories. Each triage category represents a maximum waiting time for a patient to see a doctor: red (immediately), orange (ten minutes), yellow (one hour), green (two hours) and blue (four hours). A predominant discriminator of the triage system, occurring in forty-three flowcharts, is pain[1]. The MTS prescribes to assess pain with the pain ruler (figure 1). According to the guidelines, pain should be interpreted by taking into account the patient's judgement of the pain combined with the nurse's perception of the patient's pain. This approach to pain assessments is necessary because patients can have reasons for reporting their pain to be severe as a justification of their attendance, or to claim faster treatment[1]. On the other hand, a single interpretation of the patient's pain from the triage nurse can lead to an underestimation of pain, especially in children[4].
Most studies conducted to pain assessments at triage were focused on the provision of analgesia at triage[5-9]. However, before providing analgesia a triage nurse should obtain information about the severity of the patient's pain. Only one study was found that measured the frequency of pain assessments at triage[10]. It appeared that no pain scores were registered. Furthermore, after developing and implementing a pain scale only twelve out of twenty-five patients had pain scores recorded. From this study it is unknown whether pain assessments were not conducted by triage nurses, or were not registered. Because pain is a predominant discriminator in the MTS, it affects the systems' reliability and validity when it is not assessed at triage. Besides the reliability and validity of the MTS, pain assessments are an important aspect of the quality of patient care and patient satisfaction in the ED[11].
The objective of this study was to measure the frequency of pain assessments that were conducted in accordance with the MTS guidelines. Furthermore, we studied patient, nurse, and triage characteristics associated with pain assessments at triage and, examined nurses' reasons for not conducting pain assessments at triage.

METHODS

Study design

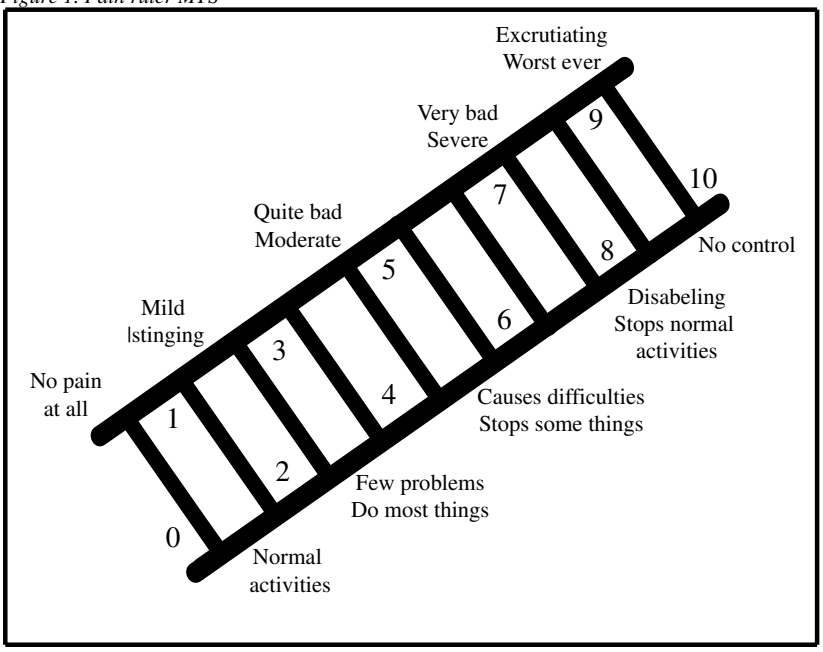
A prospective observational study was conducted in two EDs in the Netherlands. The protocol has been approved by the Medical Ethical Committee of the participating hospitals.

Study setting and population

Two EDs in the province of Utrecht participated in the study. These EDs receive between 23000 and 25000 patients annually. The MTS was implemented in 2005 and 2007. Before implementation, the nurses of both departments attended a course on how to triage with the MTS. ED trainee nurses received a course in their educational program and were trained in the ED by an experienced triage nurse.
ED trainee nurses in one ED were allowed to triage under supervision of a certified ED nurse while in the other ED only certified ED nurses performed triage. In this latter ED,

patients were triaged between 11:00 a.m. and 7:00 p.m. (peek hours) while in the other ED patients were triaged 24/7. In this study, all patients were included who presented to the participating EDs during triage hours between the 11th of May and the 18th of May 2009. All nurses who triaged patients during this study period were included in the first part of the study. In the second part of the study, all nurses and ED trainee nurses were eligible for inclusion.

Figure 1. Pain ruler MTS [1]



Data collection

During the study period, a data collection form was completed by the triage nurse for every patient presenting to the ED. This form registered patient characteristics (gender, date of birth, ethnicity, referrer to the ED, and intake of medications prior to the ED visit) and information about the process of triage (flowchart used, discriminator used, measurements, urgency category, medication provided, destination after triage and the time that the triage conversation started and stopped). Measurements consisted of temperature, pain assessments, peak expiratory flow rate, pulse rate, oxygen saturation, blood pressure, or other measurements.

Each nurse also completed another form that registered their own date of birth, gender, whether they worked as an ED trainee nurse, the number of years of experience as an ED nurse, and their experience with the MTS. To minimize information bias, the nurses were informed that the focus of this study was on the quality of triage rather than on pain assessments. After the study period the results were presented in a plenary session. In the six weeks following, nurses were asked about reasons for not assessing pain according to the MTS guidelines. These semi-structured interviews were recorded on a digital voice recorder.

Data analysis

The frequency of pain assessments conducted at triage was reported with descriptive statistics. The degree to which patient, nurse, and triage characteristics could predict pain assessments was analyzed with binary logistic regression analyses. Assessing pain according to the MTS guidelines (yes or no) was the dependent variable. Independent variables were: patient gender, age and ethnicity, referral to the ED, medications taken prior to ED presentation, flowchart, triage category, duration of triage conversation, nurses' age and gender, ED trainee nurse, years of experience as ED nurse, years of experience with the MTS and hospital. These variables were tested in univariate logistic regression analyses. Furthermore, because ED trainee nurses have no experience as an ED nurse and are less experienced with the MTS, it was tested whether these interactions were associated with pain assessments. All variables and interactions which significantly ($p<0.05$) predicted pain assessments were selected for multivariate analysis. In these analyses, patients who were triaged with one out of the nine flowcharts which do not require pain assessments were excluded.

Missing data was analyzed with SPSS missing value analysis and Little's test for data missing completely at random (MCAR). The null hypothesis that the missing data were a random subset of the population was rejected because the test score was statistically significant ($p<0.01$). This indicates that the missing data are either missing at random (MAR) or missing not at random (MNAR)[12]. Whether missing data is MAR or MNAR cannot be tested but is concluded on the basis of reasoning[13]. Further analysis of the missing data showed that it occurred more frequently within categories of the variables referral type (ambulance patients), triage category (urgent categories), and hospital. It was assumed that the missing data occurred randomly conditional on referrer type, urgency category, and hospital. Therefore, the missing data were considered MAR. The missing data were imputed with multivariate linear regression analysis (single imputation) and category prevalence's, means, and standard errors were checked for differences compared to the data before imputation. Because no substantial differences occurred, it was decided not to conduct a multiple imputation procedure. All analyses were performed with SPSS for Windows (version 15).

The semi structured interviews were transcribed for further analysis with Weft QDA (an online qualitative data analysis tool)[14]. With this program text parts of the interviews were labeled and ordered.

RESULTS

In total, 734 patients presented to both EDs. In 38.9% of these presentations the data were registered incompletely. Table 1 presents the characteristics of the patients by which the nurse conducted pain assessments or not. In both groups, the majority of the patients presenting to the ED were male, self referred, or general practitioner referred, and triaged in category green of the MTS. Table 2 presents the characteristics of the triage nurses. The majority of the triage nurses was female and had more than 2.0 years of experience with the MTS.

Table 1. Characteristics of patients (n=734)

Characteristic	Pain assessments conducted* (n=222)	Pain assessments not conducted* (n=512)
Age (median)	31.5	43.0
Gender (% male)	50.9	57.2
Casemix (%):		
Red	0.5	0.8
Orange	8.6	10.5
Yellow	27.0	28.1
Green	62.2	58.6
Blue	1.8	2.0
Referred to ED (%):		
Self referred	50.9	40.2
General practitioner	27.5	37.1
Ambulance service	11.7	16.4
Other	9.9	6.3

**Including patients who do not require pain assessments*

*Table 2. Characteristics of triage nurses (n=22)**

Characteristic	Frequency
Age (median)	45.0
Gender (% male)	36.4
Years of experience on ED (median)	6.5
Years of experience with MTS (median)	2.0
Patients triaged by trainee nurses (%)	20.3

**Including ED trainee nurses*

Pain assessments

From the allocated flowcharts and selected urgency categories, it was determined in how many patient presentations pain should have been assessed according to the MTS guidelines. This appeared to be necessary in 632 patient presentations (86.1%) while in 203 (32.1%) of these presentations, nurses registered that pain assessments were conducted. Remarkably, 188 patients were triaged on the basis of the discriminator pain (mild, moderate, or severe pain) while in only 71 patients (37.8%) the triage nurse registered that a pain assessment was conducted.

Associations with pain measurements

In univariate logistic regression analyses the variables: hospital, patient age, and ethnicity, type of referrer, intake of medication prior to ED visit, duration of triage conversation, nurse gender, ED trainee nurse, experience of the nurse with the MTS, and the interaction ED trainee nurse by experience with the MTS, were significantly associated with conducting pain assessments. The results of the multivariate logistic regression analysis are presented in table 3. In multivariate analysis the variables hospital, ethnicity, and nurse gender were no longer significantly ($p>0.05$) associated with conducting pain assessments. Children below twelve years of age, intake of medication prior to ED visit, and nurses' experience with the MTS showed the strongest association with conducting pain assessments. Also the interaction ED trainee nurse by experience with the MTS showed a strong association, ED trainee nurses were more likely to conduct pain assessments compared to certified ED nurses.

Table 3. Variables associated with pain assessments

Predictor	OR (95% CI)
Age	
< 12 yrs	2.96 (1.49 – 5.90)
13 – 20 yrs	1.76 (0.86 – 3.59)*
21 – 39 yrs	1.81 (0.93 – 3.51)*
40 – 64 yrs	1.48 (0.81 – 2.71)*
> 65 yrs	Reference
Type of referrer	
General practitioner	1.27 (0.78 – 2.06)*
Ambulance service	0.98 (0.48 – 1.99)*
Others**	2.46 (1.22 – 4.96)
Self referral	Reference
Intake medication before ED visit	
Yes	2.48 (1.34 – 4.59)
No	Reference
Duration of triage in minutes	1.14 (1.07 – 1.22)
Years of experience with MTS	1.33 (1.03 – 1.72)
ED trainee x experience with MTS	2.57 (1.32 – 5.02)

Nagelkerke r^2: 0.27
** Not statistically significant (p>0.05)*
*** Patients referred by a specialist from the hospital or by the police department*

Reasons for not measuring pain at triage

Thirteen nurses were interviewed six weeks after the results of the first part of the study were presented. Of these, eight nurses triaged four years with the MTS, four triaged between 1 and 1.5 years and one triaged since a few months. Eight nurses felt the results of the first part of the study represented daily practice while four disagreed. We asked what reasons nurses have for not conducting pain assessments according to the MTS guidelines. Seven reasons were mentioned by the nurses. It was mentioned that nurses estimate the patient's pain themselves without taking into account the patient's judgement because they think the patient's judgement of pain often results in overtriage. Nurses interprete the patient's pain by observing visual symptoms of pain such as paleness, perspiration, patient behaviour, facial expressions, or movements, by asking patients whether they need painkillers or have already taken painkillers, by interpreting vital signs such as tachycardia, and by interpreting the complaint of the patient as painful or not. Another reason mentioned was that nurses experienced time constraints at triage because other activities are conducted such as, taking blood samples for laboratory testing. Some nurses mentioned that in certain cases it is not necessary to conduct pain assessments for example, if patients have taken pain killers prior to their ED visit, if they present themselves calmly, or when the complaint existed for at least five days. Two nurses mentioned that pain assessments are often forgotten either because the guidelines are not applied or because they are applied after the triage conversation. Another reason mentioned was unfamiliarity with the MTS pain ruler and the guidelines concerning pain assessments. In several interviews this reason was confirmed because according to some nurses the guidelines state that only the patient's judgement of the pain should be used in pain assessments. Moreover, some nurses thought pain assessments were not conducted because the pain ruler is difficult to use or to interprete. One nurse mentioned that a lack of education probably resulted in a lack of pain assessments.

DISCUSSION

This study examined the frequency to which pain assessments at triage were conducted according to the MTS guidelines. It appeared that pain was assessed in about one third of the patient presentations who required pain assessments according to the guidelines. Children, intake of medication prior to ED visit, duration of triage, ED trainee nurses, and experience with the MTS were associated with conducting pain assessments.

It was found that a substantial number of patients were triaged on the discriminator pain while pain was not assessed according to the MTS guidelines. This indicates that nurses do consider pain when allocating a patient into a triage category. However, it also indicates that they interprete the patient's pain solely without taking into account the patient's judgement. In the interviews this was confirmed by several nurses who explained that the patient's judgement often results in overtriage. Other reasons for not assessing pain according to the guidelines were: time constraints, forgetting to assess pain, unfamiliarity with the pain ruler, difficulties with interpreting pain, a lack of education in how to assess pain, and the feeling it is unnecessary to assess pain. The reported reasons indicate several problems which may cause the small number of pain assessments at triage. First, a lack of clarity in the MTS guidelines. These guidelines do not prescribe how nurses should interprete pain, and how their interpretation of the patient's pain influences the triage decision. To increase nurses' confidence in assessing and interpreting pain, and to improve the reliability of pain assessments, characteristics of pain such as behavior, paleness, or vital signs could be added to the MTS guidelines. However, caution to adding these criteria is needed as some characteristics may not be strongly associated with pain intensity[15]. Moreover, the guidelines should describe the importance of both nurses' and patients' pain judgements for making a triage decision, for example, by means of adding weights to both judgements. These suggestions will create more uniformity in pain assessments between nurses. Uniformity among patients is impossible to achieve because of differences in pain experiences. The second problem is that nurses are unfamiliar with the pain ruler or the MTS guidelines for pain assessments. MTS (refresher) courses should therefore focus on pain assessments more extensively. The third and final problem is related to the organization of triage in the ED. It is recommended to implement a computer program so that nurses can triage in front of the patient and are less likely to forget pain assessments or other measures. Also activities such as taking blood samples for laboratory testing should be planned after triage because they are not necessary for assessing the urgency of the patient's complaint. This will limit time pressures on triage nurses. These changes may increase the reliability and validity of pain assessments and the MTS in total.

As mentioned in the introduction of the paper, only one study was found that studied pain assessments at triage[10]. In both studies pain assessments were not observed in practice and therefore assumptions were made. Evans et al.[10] studied registrations of pain assessments at triage retrospectively and assumed that each assessment was registered. Although the data in the present study was collected prospectively, the same assumption was made. Moreover, it was assumed that if pain was registered, the nurse assessed it in accordance with the MTS guidelines. Because of these assumptions both studies only estimated the frequency of pain assessments which could partly explain the differences between the reported frequencies, 48% vs. 32%.

Pain assessments at triage are important, not only for making a triage decision but also for relieving the patient's pain sufficiently. This latter will increase patient satisfaction while nurses are able to retriage patients in a less urgent triage category. However, pain

management in the ED has been reported to be inadequate in several studies[16-20] which can be related to poor pain assessments. These studies and the present study emphasize the importance of, the use of a reliable and valid pain tool, educated nurses in assessing pain, and uniform assessments and registrations of pain.

Limitations

This study has some limitations. First, the occurrence of missing data which was probably caused by the fact that nurses had to fill in a form for every triaged patient next to their normal triage activities. Furthermore, several missing data patterns were found. During the study it appeared that several nurses did not triage patients who arrived by ambulance because they were already triaged by the ambulance professional. Also information about patients triaged in the more urgent categories was more often missing, possibly because nurses had less time to complete a form. The differences between hospitals may have occurred because the main researcher had attended one ED more frequently during data collection than the other. By imputation of the missing data the influences on the precision of the estimates were decreased.

A second limitation is that information bias may have occurred because nurses were aware of the study. This could have caused an overestimation of the exact number of pain assessments. A more precise estimate of the frequency of pain assessments according to the MTS guidelines could be obtained by observing triage conversations in the ED for example by camera. Another advantage of triage observations by camera is that no assumptions have to be made concerning registrations or the application of guidelines. This will increase the precision of the estimated frequency. However, the interviews showed that most nurses recognized the results of the study to what they experience in practice and that reasons exist for not assessing pain at triage.

Conclusions

Pain assessments according to the MTS guidelines were conducted in nearly one third of the patient population. Reasons for not conducting pain assessments indicate a lack of clarity in the MTS guidelines, insufficient education in assessing pain and organizational difficulties in the ED. Pain assessments are likely to increase when adding physiological and behavioral characteristics of pain to the MTS guidelines, paying more attention to pain assessments in (refresher) courses, by implementing a computer program for triage, and to skip activities at triage that are not necessary for urgency estimation. Further study is needed to assess the influence of these suggestions to the reliability and validity of the pain tool and the MTS in general.

REFERENCES

1 Mackway-Jones K, Marsden J, Windle J. Emergency Triage Second Edition. 2 ed. London: Manchester Triage Group; 2005.

2 Lipley N. Foreign exchange. Emergency Nurse. 2005;13:5.

3 van Baar ME, Giesen P, Grol R, Schrijvers AJP. Reporting results of a preliminary study to Emergency Medicine. Julius Center for Health Sciences and Primary Care & Center for Quality of Care Research WOK; 2007.

4 Shavit I, Kofman M, Leder M, Hod T, Kozer E. Observational pain assessment versus self-report in paediatric triage. Emergency Medicine Journal. 2008;25:552-555.

5 Ducharme J, Tanabe P, Homel P, Miner JR, Chang AK, Lee J, et al. The influence of triage systems and triage scores on timeliness of ED analgesic administration. American Journal of Emergency Medicine. 2008;26:867-873.

6 Fosnocht DE, Swanson ER. Use of a triage pain protocol in the ED. American Journal of Emergency Medicine. 2007;25:791-793.

7 Seguin D. A nurse-initiated pain management advanced triage protocol for ED patients with an extremity injury at a level I trauma center. Journal of Emergency Nursing. 2004;30:330-335.

8 Singer AJ, Garra G, Chohan JK, Dalmedo C, Thode Jr HC. Triage pain scores and the desire for and use of analgesics. Annals of Emergency Medicine. 2008;52:689-695.

9 Williams J, Sen A. Transcribing in triage: the Wrexham experience. Accident and Emergency Nursing. 2000;8:241-248.

10 Evans C, Hawkes J, Tebbit L. Managing pain. Emergency Nurse. 2008;16:28-33.

11 Bible D. Pain assessment at nurse triage: a literature review. Emergency Nurse. 2006;14:26-29.

12 Rubin DB. Inferences and missing data. Biometrika. 1976;63:581-590.

13 Schafer JL, Graham JW. Missing data: our view of the state of the art. Psychological Methods. 2002;7:147-177.

14 Weft QDA [computer program]; 2006. Available at: http://www.pressure.to/qda/.

15 Bossart P, Fosnocht D, Swanson E. Changes in heart rate do not correlate with changes in pain intensity in emergency department patients. Journal of Emergency Medicine. 2007;32:19-22.

16 Stalnikowicz R, Mahamid R, Kaspi S, Brezis M. Undertreatment of acute pain in the emergency department: a challenge. International Journal for Quality in Health Care. 2005;17:173-176.

17 Probst BD, Lyons E, Leonard D, Esposito TJ. Factors affecting emergency department assessment and management of pain in children. Pediatric Emergency Care. 2005;21:298-305.

18 Grant PS. Analgesia delivery in the ED. American Journal of Emergency Medicine. 2006;24:806-809.

19 Karwowski-Soulie F, Lessenot-Tcherny S, Lamarche-Vadel A, Bineau S, Ginsburg C, Meyniard O, et al. Pain in an emergency department: an audit. European Journal of Emergency Medicine. 2006;13:218-224.

20 Pines JM, Hollander JE. Emergency department crowding is associated with poor care for patients with severe pain. Annals of Emergency Medicine. 2008;51:1-5.

General Discussion

INTRODUCTION

This thesis focussed on the reliability and validity of two ED triage systems and methodological aspects in these types of studies. Three study questions were formulated. First, *how reliable and valid are the MTS and the ESI?* Interpretation of the reliability of triage systems is not as straightforward as commonly used interpretation schemes of, for example, Landis & Koch or Cicchetti suggest[1;2]. According to these schemes, the reliability of the MTS by means of consistency was interpreted as 'substantial' or 'good' while according to experts almost one third of the patient population was mistriaged (chapter 2). Because these interpretation schemes do not account for the dependence of kappa on the distribution of ratings, the reliability of triage systems can be misinterpreted. A more precise interpretation is obtained by calculating triage weighted kappa and by plotting the quadratically weighted kappa against the minimum, normal, and maximum weighted kappa (chapters 3 and 4). It appeared that the reliability of the MTS is rather 'moderate', than 'substantial'. In addition, this is more in accordance with the reproducibility of the system which was not exceptionally high (chapter 2). Compared to the MTS, the reliability of the ESI has been studied more frequently. In these studies, triage weighted kappa was calculated on the basis of the reported data and represented 'substantial' reliability (chapter 3). Moreover, the normal kappa in studies measuring the reliability of the ESI showed better reliability and the plotted quadratically weighted kappas showed less extensive mistriage or disagreement compared to the MTS (chapter 4). The ESI therefore seems more reliable compared to the MTS.

The validity of the MTS and ESI was studied by means of criterion, construct, and content validity. The criterion validity was only studied in the MTS and appeared to be moderate, because substantial percentages mistriage were reported and the sensitivity for the urgent triage categories (red and orange) was only 53.2% (chapter 2). Furthermore, the patient vignettes were predominantly undertriaged which, if these patients were real, increases their risk on deterioration. The construct validity of the MTS and the ESI was studied by measuring the associations between urgency categories and, hospital admission, and ED mortality (chapter 6). Although both systems did well, the ESI was more strongly associated with hospital admission and ED mortality than the MTS. However, a small but structural flaw in the ESI required further study to its content validity (chapter 7). While previous studies reported strong associations with resource use[3-9], in this study it appeared that the ESI improperly estimated resources in elderly, self-referred patients and patients presenting with 'post operative complications, wound care and plaster problems', or 'complaints of the genitourinary system'. The content validity of the MTS was measured by studying the frequency with which triage nurses assess pain at triage (chapter 8). Because pain assessments at triage were conducted in only one third of the patient population, the content validity of the MTS concerning pain assessments is poor.

Second, *what changes to the MTS and ESI are necessary to improve the reliability and validity?* The studies described in this thesis showed that both systems' guidelines need revisions to increase the reliability and validity. In the MTS a lack of clarity in the guidelines concerning pain assessments exists (chapter 8). Revisions to these guidelines should focus on the estimation of pain by triage nurses and should include, adding characteristics or visual signs of pain such as behaviour and perspiration. Besides a revision of the guidelines it is also necessary that MTS courses focus more extensively on pain assessments and that EDs skip activities at triage that are not necessary for estimating urgency. In the ESI, it appeared that the guidelines were not adapted to health care systems

in which patients are referred to the ED by general practitioners or ambulance services (chapter 7). These patients were more frequently admitted to the hospital or sent to the outpatient department for follow-up while originally they were triaged in ESI category 5 as not needing any resources. The guidelines of the ESI therefore should be revised, for example, by adding criteria in which referred patients are at least triaged in ESI category 4 because their urgency is already assessed by a professional. Furthermore, the ESI training courses should focus extensively on elderly patients, patients presenting with 'post operative complications, wound care problems, and plaster problems', and patients presenting with 'complaints of the genitourinary system'.

The third and final question was: *what changes in triage research methods are necessary to improve interpretations of the reliability and validity of triage systems?* Methodological changes appeared to be predominantly necessary in reliability studies of triage systems. A new weighting scheme, triage weighted kappa, was developed because the reliability was often not in proportion to the reported amount of mistriage(chapter 3). Triage weighted kappa accounts for mistriage more strictly compared to existing weighting schemes and therefore reflects the reliability of triage systems more in accordance with clinical practice. Besides a new weighting scheme, a different approach was developed for interpreting the reliability of different triage systems (chapter 4). This is because kappa depends on the distribution of ratings which can result in biased interpretations regarding the reliability of triage systems, and triage weighted kappa does not account for this influence. Therefore, the reliability of triage systems should be interpreted by plotting kappa in relation to the minimum, normal, and maximum kappa. Furthermore, when comparing triage systems, one should also calculate the normal kappa in both scales. In addition, this approach affects sample size calculations in triage reliability studies and required further study (chapter 5). It appeared that currently no sample size calculation methods exist that can be applied in reliability studies to ordinal measurement scales such as triage systems. Moreover, only a few triage reliability studies calculated sample sizes prior to the onset of the study, which could have led to underpowered studies. Until a new sample size calculation method is developed, the sample sizes for ordinal scales should be calculated with methods for nominal scales. Furthermore, to encourage researchers to calculate sample sizes, a sample size calculator was developed. With this calculator one can enter values for category prevalences, number of raters, the null and alternative hypotheses and the non centrality parameter to their own preferences, after which the sample size is calculated immediately.

In studies to the validity of triage systems it appeared that no studies accounted for variables which influenced the association under study (chapter 6). It is important that future studies account for confounding variables. This will increase the precision of the estimated associations and therefore increase the precision of the interpretation of the validity of triage systems.

Implications for clinical practice and public health policy

In this study, the ESI seems more reliable and valid than the MTS. Although the reliability and validity are important aspects, the decision to implement a triage system is more comprehensive. For example, the Dutch triage guideline recommends that triage systems should be reliable and valid, applicable to all presenting ED patients as well as effective and applicable in Dutch EDs[10]. Because this thesis was focused on the reliability and validity of the MTS and the ESI, it creates limited evidence for the decision to implement one of these triage systems. However, the presented methodological changes in triage reliability and validity studies allow for more precise interpretations, which provide support

to decisions concerning the implementation of triage systems. These decisions are upcoming as recently new triage systems have been developed[11;12]. When interpreting the reliability and validity of newly developed triage systems, comparisons should be made with the reliability and validity of existing triage systems such as the MTS and ESI.
The Dutch triage guideline also recommends EDs to implement the MTS or the ESI[10]. This decision was made because in the Netherlands these two systems are predominantly implemented, probably because the first edition of the guideline reported that the ESI, Canadian Triage and Acuity Scale (CTAS) and Australasian Triage Scale (ATS) were not suitable for use in Dutch EDs[10;13]. However, in this thesis it appeared that the reliability of the CTAS and ATS approached the maximum kappa, which indicates these systems are reliable. It is therefore recommended to consider these triage systems for implementation in Dutch EDs. Besides interpreting the reliability and validity of triage systems, this thesis refuted the argument that three- or four-level triage systems are unreliable. This finding can be of importance in revising triage guidelines. For example in the MTS, in which the blue category is barely used, between 0.8% and 2.0%[14-16]. In this case the discriminative ability of this triage category is limited. Therefore, removal of category blue can be considered, especially because it will increase the consistency of the scale but will not affect the content validity. In case the content validity is affected, it is not recommended to remove a category from the system[17]. Another option is then to revise the guidelines so that more patients will be triaged in this category. In addition, this can be useful when triage systems need changes because of innovations such as the integration of general practitioner out of hours services into EDs. These organisational changes require besides an estimation of the urgency of the patient's complaint also an estimation of the preferred provider of care. The second edition of the MTS is already extended to these changes while the ESI is not. As the ESI seems more reliable and valid compared to the MTS, it is recommended to extent the ESI guidelines to this type of practice.
Several triage guidelines describe that the consistency of triage in EDs should be studied regularly to monitor the quality of triage[18-20]. They also describe that these assessments should obtain a kappa of at least 0.6, or percentages agreement of 95% (dependable on the characteristics of the ED). Based on the findings in this thesis, it is not recommended to describe such thresholds for evaluations concerning the reliability in triage guidelines, because they do not provide information about the extent of mistriage. For example, if 5% of the patients are triaged in the middle triage category of a five-level system while they should have been triaged in the most urgent triage category, this will not be recognised in the current guidelines. This can seriously affect patient safety in the ED. Therefore, guidelines should describe that evaluations of triage or triage systems should either obtain a kappa that reaches the maximum kappa or, in case percentages agreement are preferred, they should focus on the extent and characteristics of percentages disagreement.

Implications for research
When conducting a reliability study it is important to calculate sample sizes prior to the data collection of the study for reasons of costs, limited access to study participants, or statistical power[1;21;22]. The sample size calculator developed in chapter 5 is focussed on hypothesis testing and can be applied to all reliability studies evaluating nominal or ordinal measurement scales with five categories. This calculator can be used by experienced and inexperienced researchers conducting a reliability study, because it provides all steps of the calculation process.

In this thesis, it was found that linear and quadratically weighted kappa do not account for differences in mistriage and reward mistriage or disagreement too extensively. Although these weighting schemes were developed for ordinal scales[23], its use in triage reliability studies appeared to be limited. Further study is needed to develop a symmetrical weighting scheme that can be applied to triage reliability studies measuring agreement or reproducibility. Until this weighting scheme is developed, agreement and reproducibility should be calculated with linear weighted, quadratically weighted or unweighted kappa.
In reliability studies of ordinal scales one should question whether linear or quadratically weighted kappa are useful. When measuring consistency, the existing linear and quadratically weighted kappas are not suitable because of the symmetry of weightings. For triage systems therefore triage weighted kappa was developed. One can consider using this weighting scheme in reliability studies to other ordinal scales with five levels, although caution is needed. Furthermore, the approach presented in this thesis for interpreting the reliability of triage systems can be applied in reliability studies evaluating all types of measurement scales. Commonly used interpretation schemes such as of Landis & Koch[2] and Cicchetti[1] do not reflect the amount of disagreement or mistriage between raters. The reliability of a single measurement scale should be interpreted by relating the location of kappa towards the minimum, normal, and maximum kappa. Moreover, when comparing the reliability of different measurement scales, normal kappas should be calculated and interpreted with the before mentioned interpretation schemes[1;2]. To conclude whether one scale is more reliable than another, one should take into account both approaches of interpreting reliability. Dependent of the measurement scale, one can add more weight to one approach or the other. Although the normal kappa has equal chance corrections between systems, from a validity point of view, in triage systems it is important that the reliability approaches the maximum kappa. This is because extensive mistriage can affect patient safety.

Recommendations for future triage reliability and validity studies
In previously conducted triage reliability studies it appeared to be difficult to interprete the reliability of triage systems as well as to compare different studies because of inconsistencies in the reporting of reliability. Future reports on triage reliability studies should therefore report which kappa is calculated and, in case weighted kappa is used, which type of weighting scheme is applied. Furthermore, the percentages observed agreement between raters should be reported as it provides an indication of the direction of kappa. Finally, one should report what sample size calculation method is applied including the values of the parameters in these calculations.
In validity studies it is important to continue studying the construct validity as this is a process of making and testing inferences[17]. Constructs that are previously studied in triage validity studies are associations between urgency categories and: use of resources[8;16], in-hospital and ED mortality[8], ED length of stay[3-5;7;24-27], hospital length of stay[7], survival time after ED visit[4;28], hospital charges[25;27;29], and admission site[30]. All of these constructs are limited for interpreting the validity of triage systems because they ignore the fact that triage is a dynamic process. Future studies should therefore focus on measures that can be obtained directly at triage such as the modified Early Warning Score which consists of systolic blood pressure, heart rate, respiratory rate, temperature and the AVPU score (a score to determine the patient's level of consciousness)[31]. Furthermore, it is important in these studies that one corrects for variables that influence the associations under study. This

will increase the precision of the associations and improve interpretations of the validity of triage systems.

Limitations

One must be cautious with interpreting the results of this study because of several limitations. First, the reliability of the MTS was studied with patient vignettes which are limited because nurses cannot obtain visual information of patients such as behaviour or facial expression. Second, the use of expert opinions as a reference standard can cause differences in study results, especially when different experts are used in different studies. Because of this type of reference standard the consistency of measurement scales is of limited value for interpreting the reliability. The third limitation concerns the interpretation of the validity of triage systems. The construct validity was studied by measuring the associations between urgency categories and hospital admission and ED mortality. As mentioned above, these outcomes are limited for interpreting the validity because time goes by between triage and admission or ED mortality. In addition, it could not be determined from the data whether patients were retriaged and if they were, whether the registered scores represented the first assessment or a second.

The final limitation of this study is missing data which occurred throughout all studies described in this thesis. It is difficult to conduct studies in EDs because of variations in the number of patient presentations and severity of the complaints. However, attempts were undertaken to estimate missing values or to describe the characteristics of missing data and determine its influence on the point estimates.

Conclusion

The ESI seems more reliable and valid compared to the MTS, which should be considered when deciding to implement a triage system. However, both systems have flaws and need to be revised. In the MTS the guidelines concerning pain assessments need to be revised and training courses should focus on this part of triage. The guidelines of the ESI should be adapted to healthcare systems in which patients need a referral from a general practitioner or ambulance service as well as to recent organisational changes in which the urgency and the appropriate provider of care must be estimated. Further study to the effects of these changes on the reliability and validity of the MTS and ESI is required before recommending to implement one system or the other throughout EDs. Moreover, the triage guidelines of the MTS, ESI, ATS, and CTAS need revisions concerning the outcomes of assessments to the quality of triage because the current thresholds do not reflect the extent of mistriage or disagreement. The Dutch triage guideline should consider to implement the CTAS and ATS in Dutch EDs as these systems were more reliable than the MTS and ESI.

Methodologically, symmetrical weighting schemes such as linear, and quadratically weighted kappa are not useful in studies measuring consistency. In addition, they are of limited use in studies measuring agreement or reproducibility. In case consistency is measured, a non symmetrical weighting scheme should be applied such as triage weighted kappa. An interpretation of the reliability can be obtained by plotting (triage weighted) kappa against the minimum, normal, and maximum kappa. In case comparisons of the reliability between measurement scales are made, this approach should be combined with calculating the normal kappa.

Finally, reports about the reliability and validity of triage systems need to be more precise in the type of reliability and validity they assess and in their statistical reports. Sample size

calculations, the type of kappa used, and the type of weighting scheme that is used are crucial to report.

REFERENCES

1 Cicchetti D, Bronen R, Spencer S, Haut S, Berg A, Oliver P, et al. Rating scales, scales of measurement, issues of reliability. Journal of Nervous and Mental Disease. 2006;194:557-564.

2 Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33:159-174.

3 Baumann MR, Strout TD. Evaluation of the Emergency Severity Index (version 3) triage algorithm in pediatric patients. Academic Emergency Medicine. 2005;12:219-224.

4 Baumann MR, Strout TD. Triage of geriatric patients in the emergency department: validity and survival with the Emergency Severity Index. Annals of Emergency Medicine. 2007;49:234-240.

5 Eitel DR, Travers DA, Rosenau AM, Gilboy N, Wuerz RC. The emergency severity index triage algorithm version 2 is reliable and valid. Academic Emergency Medicine. 2003;10:1070-1080.

6 Elshove-Bolk J, Mencl F, van Rijswijck BT, Simons MP, van Vugt AB. Validation of the Emergency Severity Index (ESI) in self-referred patients in a European emergency department. Emergency Medicine Journal. 2007;24:170-174.

7 Tanabe P, Gimbel R, Yarnold PR, Adams JG. The Emergency Severity Index (version 3) 5-level triage system scores predict ED resource consumption. Journal of Emergency Nursing. 2004;30:22-29.

8 Worster A, Fernandes CM, Eva K, Upadhye S. Predictive validity comparison of two five-level triage acuity scales. European Journal of Emergency Medicine. 2007;14:188-192.

9 Wuerz RC, Milne LW, Eitel DR, Travers D, Gilboy N. Reliability and validity of a new five-level triage instrument. Academic Emergency Medicine. 2000;7:236-242.

10 HAN University of applied sciences, the Dutch Emergency Nurses Association (NVSHV), Netherlands Centre of Excellence in Nursing (LEVV). Guideline Triage on the Emergency Department; 2008.

11 Drijver R, Jochems P, Herrmann G, in 't Veld K, ten Wolde W. Netherlands Triage System: towards uniform triage; 2006.

12 Taboulet P, Moreira V, Haas L, Porcher R, Braganca A, Fontaine JP, et al. Triage with the French Emergency Nurses Classification in Hospital scale: reliability and validity. European Journal Emergency Medicine. 2009;16:61-67.

13 Coenen IGA, Hagemeijer A, de Caluwé R, de Voeght FJ, Jochems PJJ. Guideline Triage on the emergency department. Dutch Institute for Healthcare Improvement; 2004.

14 Grouse AI, Bishop RO, Bannon AM. The Manchester Triage System provides good reliability in an Australian emergency department. Emergency Medicine Journal. 2009;26:484-486.

15 Martins HM, Cuna LM, Freitas P. Is Manchester (MTS) more than a triage system? A study of its association with mortality and admission to a large Portuguese hospital. Emergency Medicine Journal. 2009;26:183-186.

16 Roukema J, Steyerberg EW, van Meurs A, Ruige M, van der Lei J, Moll HA. Validity of the Manchester Triage System in paediatric emergency care. Emergency Medicine Journal. 2006;23:906-910.

17 Streiner DL, Norman GR. Health Measurement Scales. 3 ed. Oxfort University Press; 2007.

18 Australasian College for Emergency Medicine. Policy Document: The Australasian Triage Scale; 2000.

19 Gilboy N, Tanabe P, Travers D, Rosenau A, Eitel DR. Emergency Severity Index, Version 4: Implementation Handbook. Rockville: Agency for Healthcare Research and Quality; 2005.

20 Mackway-Jones K. Emergency Triage. 2 ed. London: Manchester Triage Group; 2006.

21 Shoukri MM. Sample size requirements for the design of a reliability study. Measurement of interobserver agreement. Chapman & Hall/ CRC; 2003. p. 85-102.

22 Shoukri MM, Asyali MH, Donner A. Sample size requirements for the design of reliability study: review and new results. Statistical Methods in Medical Research. 2004;13:251-272.

23 Fleiss JL. Determining sample sizes needed to detect a difference between two proportions. Statistical methods for rates and proportions. 2 ed. John Wiley & Sons; 1981. p. 33-48.

24 Kim TG, Cho JK, Kim SH, Lee HS, Gu HD, Chung SW. Reliability and validity of the modified Emergency Severity Index-2 ad a triage tool. Journal of the Korean Society of Emergency Medicine. 2006;17:154-164.

25 Maningas PA, Hime DA, Parker DE. The use of the Soterion Rapid Triage System in children presenting to the Emergency Department. Journal of Emergency Medicine. 2006;31:353-359.

26 Wuerz RC, Travers D, Gilboy N, Eitel DR, Rosenau A, Yazhari R. Implementation and refinement of the emergency severity index. Academic Emergency Medicine. 2001;8:170-176.

27 Maningas PA, Hime DA, Parker DE, McMurry TA. The Soterion Rapid Triage System: evaluation of inter-rater reliability and validity. Journal of Emergency Medicine. 2006;30:461-469.

28 Wuerz RC. Emergency severity index triage category is associated with six-month survival. ESI Triage Study Group. Academic Emergency Medicine. 2001;8:61-64.

29 Travers DA, Waller AE, Bowling JM, Flowers D, Tintinalli J. Five-level triage system more effective than three-level in tertiary emergency department. Journal Emergency Nursing. 2002;28:395-400.

30 Tanabe P, Gimbel R, Yarnold PR, Kyriacou DN, Adams JG. Reliability and validity of scores on The Emergency Severity Index version 3. Academic Emergency Medicine. 2004;11:59-65.

31 Subbe CP, Kruger M, Rutherford P, Gemmel L. Validation of a modified Early Warning Score in medical admissions. QJM. 2001;94:521-526.

Summary

Emergency department (ED) triage is a dynamic process of rapidly and systematically categorizing the patient's severity of illness or injury and the patient's priority for treatment to efficiently use ED resources. Nowadays, several five-level triage systems such as the Emergency Severity Index (ESI), and the Manchester Triage System (MTS), have been implemented in EDs. It is important that triage systems are reliable and valid because this can affect patient safety in the ED. Furthermore, methods in triage reliability and validity studies should reflect clinical practice because often decisions concerning the implementation of triage systems are made on the basis of the results of these studies. In this thesis, the reliability and validity of the MTS and ESI was studied and, methodological aspects in these types of studies.

The first part of this thesis focussed on reliability. In **chapter 2**, the consistency, reproducibility, and criterion validity of the MTS were studied by means of patient vignettes. In that study, nurses of two EDs rated fifty patient vignettes twice. Their ratings were compared with the ratings of two Dutch MTS expert to measure the consistency, which appeared to be 'substantial'. The reproducibility of the MTS appeared to be high. However, substantial percentages mistriage occurred and the sensitivity of the system for detecting urgent patient complaints was, except for children below sixteen years of age, poor.

The consistency in chapter 2 was calculated by a weighted kappa statistic which represents chance corrected agreement. In **chapter 3** linear and quadratic weighting schemes for calculating weighted kappa were studied to the degree they reflect clinical practice. It appeared that these weighting schemes reward mistriage too extensively as kappa often represented 'substantial' or 'almost perfect' agreement while substantial percentages of mistriage were reported. Therefore a new weighting scheme, triage weighted kappa, was developed and applied to the reported data of previously conducted triage reliability studies. With this weighting scheme, the reliability of triage systems was smaller than previously reported. Besides shortcomings in the existing weighting schemes, the value of kappa is influenced by the distribution of the data. Because of this influence, interpretations of the reliability of triage systems were often over- or underestimated. In **chapter 4** an approach was developed on how to interprete the reliability of triage systems, and how to compare the reliability of these systems. The reliability of a single triage system should be interpreted by plotting the quadratically weighted kappa to the minimum, normal, and maximum kappa, to obtain information about the skewness of the data. In case of comparing the reliability between systems, also a normal kappa should be calculated to obtain equal chance corrections. Both outcomes should be taken into account when interpreting the reliability of the systems that are compared. The normal kappa in the ESI appeared to be the highest compared to other frequently implemented triage systems. However, the quadratically weighted kappas in studies to the ESI and MTS approached the normal kappa which indicates substantial mistriage. From a validity point of view, the reliability of triage systems should approach the maximum weighted kappa because mistriage or disagreement is centred within one category from agreement. The approach for interpreting the reliability of triage systems affects sample size calculations. Therefore, in **chapter 5** sample size calculation methods for reliability studies reporting a kappa statistic were reviewed. It appeared that currently no sample size calculation methods exist for reliability studies conducted to ordinal measurement scales such as triage systems. Moreover, only a few triage reliability studies calculated sample sizes. Sample size calculations are important to conduct because of time and cost constraints, and because studies are sufficiently powered. To encourage researchers to calculate sample sizes, a

sample size calculator was developed and applied in an example. When calculating sample sizes it is important to take into account the dependence of kappa on the distribution of data by setting the null- and alternative hypothesis to the maximum kappa or to set the confidence interval width to a maximum of 0.10.
The second part of this thesis focussed on validity. In **chapter 6** the construct validity of the MTS and ESI was studied by measuring associations between urgency categories and hospital admission and ED mortality. It appeared that the ESI was more strongly associated with hospital admission than the MTS. Concerning ED mortality, in both systems patients in the more urgent categories were at increased risk of dying. Because of the small number of patients who died in the ED, it was not possible to determine the strength of the association between urgency categories and ED mortality. In this study, it was found that a small number of patients was triaged in ESI category 5 (no resources are needed) but were subsequently admitted to the hospital or sent to the outpatient department for follow-up. In **chapter 7** the characteristics of these patients were studied further (content validity). Patients triaged in ESI category 5 were most likely elderly patients, patients referred by a general practitioner or ambulance service, and patients presenting with 'post operative complications, wound care problems, and plaster problems', or 'complaints of the genitourinary system'. Because of this small but structural flaw of the ESI, the results of this study indicate that the guidelines of the ESI need revisions concerning these patient groups. In **chapter 8** the content validity of the MTS was studied by measuring the frequency that nurses conduct pain assessments at triage. It appeared that in more than four out of the five patient presentations, pain should have been assessed according to the MTS guidelines. Pain was assessed in almost one third of the patient presentations. In a substantial number of patients nurses took pain into account when allocating a patient into a triage category but pain was not assessed. This indicated that nurses estimated the patient's pain solely while according to the MTS guidelines pain should be assessed by taking into account both, the patient's and nurse's judgement of the patient's pain. This was confirmed in the interviews with nurses who mentioned several reasons for not conducting pain assessments according to the MTS guidelines at triage. These reasons included, overtriage because of the patient's judgement, time constraints, and difficulties with interpreting pain. The reported reasons indicated that the guidelines of the MTS concerning pain assessments need to be revised, EDs should skip activities at triage that are not necessary for estimating urgency, and MTS courses should focus on pain assessments more extensively.
The studies described in this thesis indicate that the ESI is more reliable and valid than the MTS. However, both systems' guidelines need revisions. The effects of the proposed revisions should be studied further before recommending to implement the ESI or the MTS throughout EDs. Methodologically, when conducting reliability studies it is important to apply symmetrical weighting schemes in case agreement or reproducibility is studied. In case the consistency is studied, one should apply a non symmetrical weighting scheme such as triage weighted kappa. Furthermore, the reliability should be interpreted by plotting kappa against the minimum, normal, and maximum kappa and by calculating the normal kappa in case comparisons are being made. Compared to previous methods of calculating and interpreting reliability, these suggestions will reflect the reliability more in accordance with clinical practice.

Samenvatting

Triage op de spoedeisende hulp (SEH) is een dynamisch proces waarin patiënten snel en systematisch ingedeeld worden naar de ernst van de aandoening of verwonding en prioriteit op behandeling, met het doel om efficiënt gebruik te maken van SEH middelen. Wereldwijd zijn er verschillende triagesystemen geïmplementeerd op SEHs. De meest frequent geïmplementeerde systemen zijn de Emergency Severity Index (ESI), het Manchester Triage Systeem (MTS), de Canadian Triage and Acuity Scale en de Australasian Triage Scale. Deze systemen bestaan uit vijf triage categorieën. Vanwege de patiëntveiligheid is het belangrijk dat triagesystemen betrouwbaar en valide zijn. Daarnaast is het belangrijk dat onderzoeken naar de betrouwbaarheid en validiteit van triage systemen een goede weergave zijn van de klinische praktijk. De resultaten van deze onderzoeken worden vaak gebruikt om beslissingen omtrent de implementatie van triage systemen te verantwoorden. In dit proefschrift wordt de betrouwbaarheid en validiteit van het MTS en de ESI onderzocht. Tevens worden methodologische aspecten van dit type studies onderzocht.

Het eerste deel van dit proefschrift is gericht op betrouwbaarheid. In **hoofdstuk 2** worden de consistentie, reproduceerbaarheid, en de criterium validiteit van het MTS onderzocht met behulp van patiënt vignetten. Verpleegkundigen van twee SEHs hebben elk tweemaal vijftig patiënt vignetten beoordeeld. Deze beoordelingen werden vergeleken met de beoordelingen van twee Nederlandse MTS experts. De consistentie van de beoordelingen was groot en de reproduceerbaarheid was hoog. Echter van een groot aantal patiënt vignetten werd de urgentie over- of onderschat (over- en ondertriage) en de sensitiviteit voor het herkennen van de urgente triage categorieën was, behalve voor kinderen tot zestien jaar, slecht.
De consistentie in hoofdstuk 2 was berekend met een gewogen kappa, een maat voor overeenstemming gecorrigeerd voor overeenstemming die er op basis van kans al bestaat. **Hoofdstuk 3** bestudeert de mate waarin weegschema's met lineaire en kwadratische gewichten voor het berekenen van een gewogen kappa de klinische praktijk weergeven. Het blijkt dat deze weegschema's over- en ondertriage teveel belonen omdat in een aantal studies de betrouwbaarheid geïnterpreteerd werd als 'bijna perfect' of 'groot', terwijl er grote hoeveelheden over- en ondertriage gerapporteerd werden. Triage weighted kappa is ontwikkeld om over- en ondertriage sterker te bestraffen dan de bestaande weegschema's. Dit weegschema is toegepast op de data van eerder gepubliceerde betrouwbaarheidsstudies. Behalve een te grote beloning voor over- en ondertriage, is de hoogte van kappa afhankelijk van de verdeling van de data. Dit heeft gevolgen voor het interpreteren van de betrouwbaarheid. In **hoofdstuk 4** is een methode ontwikkeld om de betrouwbaarheid van triagesystemen te interpreteren door kappa te plotten ten opzichte van de minimale, normale, en maximaal haalbare kappa. Op deze manier verkrijgt men informatie over de scheefheid van de data. Wanneer de betrouwbaarheid van triage systemen wordt vergeleken, moet men tevens een normale kappa berekenen omdat deze gelijke kanscorrecties hanteert. Van de vier frequent geïmplementeerde triage systemen is de normale kappa het hoogst in de ESI. Echter de kwadratisch gewogen kappas in studies naar de ESI en het MTS benaderen de normale kappa. Dit betekent dat er in deze studies grote hoeveelheden over- en ondertriage voorkomen. Vanuit het oogpunt van validiteit is dit niet wenselijk, en moeten triage systemen de maximaal haalbare kappa benaderen omdat over- en ondertriage zich dan binnen een categorie van perfecte overeenstemming bevinden. Omdat deze benadering ook gevolgen heeft voor het berekenen van de steekproefomvang in betrouwbaarheidsstudies, zijn in **hoofdstuk 5** methoden voor het berekenen van de

steekproefomvang voor dit type studies bestudeerd. Uit deze studie blijkt dat er momenteel geen methoden beschikbaar zijn die geschikt zijn voor betrouwbaarheidsstudies naar ordinale meetschalen zoals triage systemen. Ook hebben weinig betrouwbaarheidsstudies naar triage systemen een steekproefomvang berekend. Het is belangrijk om de steekproefomvang te berekenen vanwege aspecten zoals tijd en kosten, maar ook om een studie op te zetten die voldoende statistische power heeft. Om onderzoekers aan te moedigen om de steekproefomvang te berekenen, is een 'sample size calculator' ontwikkeld. De methode voor het interpreteren van betrouwbaarheid is verwerkt in het berekenen van de steekproefomvang. Dit is gedaan door de nul- en alternatieve hypothese te baseren op de maximaal haalbare kappa, of door de breedte van het betrouwbaarheidsinterval te beperken tot maximaal 0.10.

Het tweede deel van dit proefschrift is gericht op validiteit. In **hoofdstuk 6** is de construct validiteit van het MTS en de ESI onderzocht. Dit is gedaan door de associatie te meten tussen de urgentie categorieën van beide systemen en, opname in het ziekenhuis en SEH mortaliteit. De ESI bleek sterker geassocieerd met ziekenhuisopname dan het MTS. In beide systemen lopen patiënten in de meest urgente categorieën het grootste risico op overlijden. Omdat het aantal patiënten dat overlijd op de SEH erg klein is, is het niet mogelijk de sterkte van de associatie te meten. Een klein percentage patiënten die getriëerd is in ESI categorie 5 (geen hulpmiddelen nodig), bleek aan het einde van het SEH bezoek opgenomen te zijn in het ziekenhuis of verwezen naar een polikliniek. De kenmerken van deze groep zijn nader onderzocht in **hoofdstuk 7** (inhoudsvaliditeit). Oudere patiënten, patiënten verwezen door de huisarts of ambulancedienst en patiënten die zich presenteerden met 'post operatieve complicaties, wondproblemen en gipsproblemen' of 'klachten van het urogenitaal stelsel' liepen de grootste kans om getriëerd te worden in ESI categorie 5 en later opgenomen te worden of verwezen te worden naar de polikliniek. De resultaten van deze studie impliceren een kleine, maar structurele tekortkoming in de ESI die enkele aanpassingen vereisen van de ESI richtlijnen. In **hoofdstuk 8** is de inhoudsvaliditeit onderzocht van het MTS door de frequentie van het meten van pijn door verpleegkundigen te bestuderen. Volgens de richtlijnen van het MTS had pijn bij meer dan vier op de vijf patiënten gemeten moeten worden. Pijn werd gemeten in bijna een derde van de patiënten. Bij een groot aantal patiënten hebben verpleegkundigen wel rekening gehouden hebben met pijn bij het indelen van de patiënt in een urgentie categorie maar is pijn niet gemeten. Dit geeft aan dat verpleegkundigen zelf een inschatting maken van de pijn van de patiënt terwijl de MTS richtlijnen aangeven dat voor het beoordelen van pijn een schatting van beide, de patiënt en de verpleegkundige nodig is. Dit werd door een aantal verpleegkundigen bevestigd in de interviews. Een aantal redenen voor het niet meten van pijn volgens de MTS richtlijnen zijn onder andere, overtriage door het oordeel van de patiënt mee te nemen, tijdsdruk tijdens triage, en moeilijkheden met het interpreteren van pijn. Aanpassingen aan de MTS richtlijnen met betrekking tot het meten van pijn zijn nodig om meer uniformiteit in het meten van pijn tussen verpleegkundigen te krijgen. Bovendien moeten SEHs activiteiten die niet nodig zijn voor het meten van urgentie buiten de triage plannen en moeten MTS cursussen zich nadrukkelijker richten op het meten van pijn.
De in dit proefschrift beschreven onderzoeken geven aan dat de ESI betrouwbaarder en meer valide is dan het MTS. Echter de richtlijnen van beide systemen behoeven aanpassingen. De effecten van de voorgestelde aanpassingen op de betrouwbaarheid en validiteit van het MTS en de ESI moeten nader onderzocht worden alvorens een van beide systemen aan te bevelen voor implementatie in SEHs. Vanuit methodologisch oogpunt is

het belangrijk om bij het berekenen van de betrouwbaarheid gebruik te maken van symmetrische weegschema's als de hoeveelheid overeenstemming of reproduceerbaarheid bestudeerd worden. Wanneer men consistentie bestudeerd is het gebruik van een niet symmetrisch weegschema aanbevolen. Het interpreteren van de betrouwbaarheid gebeurt door het plotten van kappa ten opzichte van de minimale, normale, en maximaal haalbare kappa en door het berekenen van de normale kappa wanneer men systemen wil vergelijken. Deze aanbevelingen geven een betere weergave van de betrouwbaarheid in overeenstemming met de klinische praktijk in vergelijking met bestaande methoden voor het berekenen en interpreteren van betrouwbaarheid.

Appendix A

Creating linear algorithms

1. Calculate ‘a’ by selecting two points (x,y) from the line in figure 1: e.g. (1,1) and (0,-0.25);

2. To obtain ‘a’ divide the difference between y values by the differences in x values: ‘a’= Δ’y’ / Δ’x’= (1.00 + 0.25) / (1.00 – 0.00)= 1.25 / 1.00= 1.25;

3. To obtain ‘b’ fill in the algorithm: y= 1.25x+b; x=0.50 and y=0.40; 0.40=1.25. 0.50+ b; -b=0.225; b=-0.225;

4. Algorithm for five-level triage scales: y=1.25x-0.225.

Appendix B

Chapters of the ICD 10 including modified complaint categories

- Infectious and parasitic diseases
- Oncology related complaints
- Diseases of the blood and blood forming organs
- Mental and behavioral disorders
- Diseases/complaints of the nervous system
- Diseases/complaints of the eye
- Diseases/complaints of the ear, nose and throat
- Diseases/complaints of the circulatory system
- Diseases/complaints of the respiratory system
- Diseases/complaints of the digestive system
- Diseases/complaints of the skin and subcutaneous tissue
- Diseases/complaints of the musculoskeletal system and connective tissue
- Diseases/complaints of the genitourinary system
- Pregnancy, childbirth and the puerperium
- Symptoms, signs and abnormal clinical and laboratory findings not elsewhere classified
- Injury and certain other consequences of external causes
- Endocrine, nutritional and metabolic diseases
- Non traumatic bleedings
- Check up appointment at the ED
- Post operative complications, wound care problems and plaster problems

Dankwoord

The show is over, het is klaar! Ik ben heel erg trots op het resultaat. Graag wil ik een aantal mensen bedanken die, ieder op hun eigen wijze, dit proefschrift mede mogelijk hebben gemaakt:

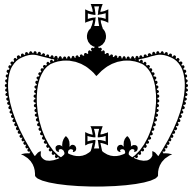
Prof. dr. A.J.P. Schrijvers. Beste **Guus**, je enthousiasme over het onderwerp triage heeft aantstekelijk gewerkt. Bedankt dat je me de tijd en vrijheid hebt gegeven in het bedenken en uitvoeren van de onderzoeken. Dr. H.F. van Stel. **Henk**, bedankt dat je de begeleiding tussentijds hebt overgenomen. Je expertise en enthousiasme op het gebied van de onderzoeksmethodologie kwam erg goed van pas en ik heb er dankbaar gebruik van gemaakt. Dr. L.M. Sturms. **Leontien**, bedankt voor je steun en je kritische vragen. Door jou zijn mijn 'uitvindingen' nog mooier geworden! Dr. M.E. van Baar. **Margriet**, jij was mijn begeleider van het eerste uur. Ik heb heel veel van je geleerd en ik heb met veel plezier met je samengewerkt!

Ik bedank de **verpleging** en het **management** van de deelnemende ziekenhuizen aan de diverse onderzoeken: Sint Antonius Ziekenhuis locaties Nieuwegein en Utrecht Oudenrijn, Meander Medisch Centrum, Medisch Centrum Haaglanden, Sint Elisabeth Ziekenhuis en het Onze Lieve Vrouwe Gasthuis. In het bijzonder bedank ik **Marian Schot-Balfoort,** omdat je ontzettend hard gewerkt hebt voor de studie pijnmetingen in het MTS!

Mijn **kamergenoten** van 'tha livingroom' 6.104. Wat moest ik zonder jullie? Bedankt voor al jullie steun, geduld, adviezen, en luisterende oren. En **Mascha**, ik vind het geweldig dat je mijn paranimf wil zijn!

Joanne, Karin, en **Margit.** Tijdens het promoveren kon ik op elk mogelijk tijdstip bij jullie terecht en dat is heel bijzonder. Ik heb zelfs nog een half jaar bij jullie gewoond. Dit was erg gezellig en maakte het zoeken van een huis in Utrecht en het vele kluswerk wat erop volgde een stuk eenvoudiger. Ook heb ik met jullie (Scharrie Tuttebola en het Kanariepietje) diverse uithoeken van de wereld bekeken. Deze avonturen, en de mini mini momenten, zorgden voor nieuwe inspiratie waardoor ik op volle kracht vooruit kon met het onderzoek.

Tot slot, **pap, mam, Jeroen** en **Eva**. Jullie hebben altijd in mij geloofd en achter mij gestaan, welke uitdaging ik ook aanging. Ook tijdens het promoveren waren jullie er met de worstenbroodjes-sessies, telefoongesprekken, (preventieve) huis- en klusbezoeken en helemaal op het laatst zelfs met het ontwikkelen van de sample size calculator en de voorkant van dit proefschrift. We hebben nog één uitdaging te gaan, de verdediging. Eva, met jou als paranimf gaat dit zeker lukken!



About the Author

Ineke van der Wulp was born on May 11th 1982 in Nuenen, Gerwen and Nederwetten. In 1999, she graduated from the HAVO at the Lorentz Casimir Lyceum in Eindhoven and started a study nursing at the Fontys Hogescholen in Eindhoven. She obtained her Bachelors degree of Nursing in 2003 and started a study Health Sciences at the University of Maastricht. Two years later, in 2005 she obtained her Master of Science degree. Her final thesis described difficulties experienced by adolescents who take care for a chronically ill or handicapped parent. In May 2006 she started her PhD project at the Julius Center for Health Sciences and Primary Care in Utrecht. The project was supervised by Guus Schrijvers, Professor of Public Health, and Henk van Stel, PhD.