Journal of Clinical Epidemiology 94 (2018) 27−34

Journal of Clinical Epidemiology

Multiple performance measures are needed to evaluate triage systems in the emergency department

Joany M. Zachariasse[a], Daan Nieboer[b], Rianne Oostenbrink[a], Henriëtte A. Moll[a], Ewout W. Steyerberg[b,*,1]

[a]*Department of General Paediatrics, Erasmus MC-Sophia Children's Hospital, University Medical Center Rotterdam, P.O. Box 2040, 3000 CB, Rotterdam, The Netherlands*

[b]*Department of Public Health, Erasmus MC, University Medical Center Rotterdam, P.O. Box 2040, 3000 CA, Rotterdam, The Netherlands*

Accepted 8 November 2017; Published online 14 November 2017

Abstract

Objectives: Emergency department triage systems can be considered prediction rules with an ordinal outcome, where different directions of misclassification have different clinical consequences. We evaluated strategies to compare the performance of triage systems and aimed to propose a set of performance measures that should be used in future studies.

Study Design and Setting: We identified performance measures based on literature review and expert knowledge. Their properties are illustrated in a case study evaluating two triage modifications in a cohort of 14,485 pediatric emergency department visits. Strengths and weaknesses of the performance measures were systematically appraised.

Results: Commonly reported performance measures are measures of statistical association (34/60 studies) and diagnostic accuracy (17/60 studies). The case study illustrates that none of the performance measures fulfills all criteria for triage evaluation. Decision curves are the performance measures with the most attractive features but require dichotomization. In addition, paired diagnostic accuracy measures can be recommended for dichotomized analysis, and the triage-weighted kappa and Nagelkerke's R^2 for ordinal analyses. Other performance measures provide limited additional information.

Conclusion: When comparing modifications of triage systems, decision curves and diagnostic accuracy measures should be used in a dichotomized analysis, and the triage-weighted kappa and Nagelkerke's R^2 in an ordinal approach.

Keywords: Triage; Performance measures; Decision curve; Diagnostic accuracy measures; Kappa; Nagelkerke's R^2

1. Introduction

Emergency departments face large and unpredictable numbers of patients, presenting with a broad spectrum of illnesses and injuries [1,2]. As demand often exceeds the capacity to provide immediate care, most patients have to wait before they can be seen by a health care professional. The vast majority of emergency care settings have triage systems in place to prioritize patients and ensure they are seen in the order of clinical need rather than in the order of attendance [3]. Research on triage systems is important, both to understand the performance of currently used systems and to enable the evaluation of modifications for improvement. Nevertheless, despite the almost universal application of triage systems at the emergency departments, studies about their performance to correctly discriminate between high- and low-urgency patients are limited and hampered by methodological limitations [4−6].

Triage systems can be considered a specific type of prediction model, and their evaluation can be approached accordingly. Some important characteristics of triage systems, however, make their evaluation more challenging. Triage systems typically classify patients into five ordinal categories. There is a vast methodological literature on prediction models with dichotomous outcomes (i.e., the presence or absence of the disease), but evaluation of ordinal prediction models has been less well studied [7,8].

Funding statement: This research did not receive any specific grant from funding agencies in the public, commercial, or non-for-profit sectors.

Conflict of interest statement: There are no conflicts of interest to disclose.

[1] Current address: Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, P.O. Box 9600, 2300 RC Leiden, The Netherlands.

* Corresponding author. Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, P.O. Box 9600, 2300 RC, Leiden, The Netherlands. Tel.: +31 71 526 9700.

E-mail address: e.w.steyerberg@lumc.nl (E.W. Steyerberg).

https://doi.org/10.1016/j.jclinepi.2017.11.004

What is new?

Key findings

- Commonly used performance measures do not take into account the specific features of emergency department triage systems, including their ordinal nature and the different clinical consequences of the different directions of misclassification.
- Decision curves, paired diagnostic accuracy measures, the triage-weighted kappa, and Nagelkerke's R^2 are the performance measures with the most attractive features.

What this adds to what was known?

- We provide a comprehensive evaluation of the performance measures used in the evaluation of triage systems.
- Based on an appraisal of the strengths and weaknesses of each of the performance measures in a case study, we propose a set of performance measures that researchers should report when evaluating triage systems.

What is the implication and what should change now?

- As a minimum, decision curves and diagnostic accuracy measures should be used in a dichotomized analysis, and the triage-weighted kappa and Nagelkerke's R^2 in an ordinal approach, when comparing triage systems.

Furthermore, different types of misclassification by a triage system have different clinical consequences. Classifying critically ill patients to a too low urgency level (''undertriage'') leads to delays in treatment with immediate clinical consequences. Classifying nonurgent patients to a too high urgency level (''overtriage'') does not have a direct effect on the patient but decreases the efficiency of the system, ultimately leading to increased waiting times for severely ill patients correctly classified as high urgent. Commonly used performance measures typically do not take into account the ordinal nature of triage systems or the different weights of the different directions of misclassification.

In this study, we aim to evaluate currently available strategies to compare the performance of triage systems. We will illustrate the challenges when assessing triage systems and the properties of several performance measures with a case study that evaluates two modifications of a commonly used triage system, one aimed to reduce overtriage and one aimed to reduce undertriage. Furthermore, we aim to propose a set of performance measures that should be used in future studies.

2. Review of performance measures

We first considered performance measures based on a review of the current literature. We used the search selection from a previously conducted systematic review including EMBASE, Medline, OvidSP, Cochrane central, Web-of-science, and CINAHL databases from 1980 till 2013 to identify studies that assessed the performance of a triage system in emergency care with a predefined reference standard [9]. After an update in May 2017, a total of 60 studies were included, published between 1996 and 2017.

Of the 60 included studies, 35 (58%) used measures of statistical association to describe the performance of triage systems. The most commonly reported measures were Pearson's *chi-square test*, *t*-test, ANOVA, and their nonparametric equivalent. We also found several types of correlation coefficients (7 studies) and regression coefficients or odds ratios (9 studies). Seventeen studies (28%) reported some type of diagnostic accuracy measure, including sensitivity, specificity, predictive values, and likelihood ratios, whereas 10 reported the area under the curve, and 15 studies described some other type of performance measures such as the kappa statistic (Appendix A).

For our case study, we selected the performance measures that were used in at least two different studies, thereby excluding RIDIT analysis, univariate optimal discriminant analysis, the Kolmogorov−Smirnov test, the Wald−Wolfowitz test, Friedman's test, Cox-Stuart trend test, and Net reclassification improvement. In addition to these measures selected from the literature, methodologic experts (D.N. and E.W.S.) identified a number of performance measures commonly used in diagnostic and prognostic research, and several recently developed performance measures (Table 1). These included Nagelkerke's R^2, the ordinal C-statistic and decision curve analysis [7,10,11]. Besides the unweighted kappa, we calculated a weighted kappa using quadratic weights and a ''triage-weighted kappa''. The triage-weighted kappa has been proposed as an alternative weighting scheme specifically adapted to the practice of triage [12] (Appendix B).

In this article, we focus on the measures of overall performance, diagnostic accuracy, discrimination, clinical usefulness, and agreement that can be used to compare triage systems' performance to correctly discriminate between high- and low-urgency patients.

We used the following criteria to evaluate the selected performance measures for their ability to evaluate and compare modifications of triage systems:

1. Can the performance measure be applied to ordinal data without requiring dichotomization? (+, if the performance measure can be applied to ordinal data without dichotomization; −, if not)
2. Does the performance measure take into account the weights of different types of misclassification (overtriage or undertriage)? (+, if the performance measure