Assessment of inter-observer reliability of two five-level triage and acuity scales: a randomized controlled trial

Andrew Worster, MD, MSc;[*§] Nicki Gilboy, RN;[†] Christopher M. Fernandes, MD;[*]
David Eitel, MD, MBA;[‡] Kevin Eva, PhD;[§] Rose Geisler, RN;[*] Paula Tanabe, RN, PhD[¶]

SEE ALSO COMMENTARY, PAGE 288.

ABSTRACT

Introduction: The Emergency Severity Index (ESI©) is an initial measure of patient assessment in the emergency department (ED). It rates patients based on acuity and predicted resource intensity from Level 1 (most ill) to Level 5 (least resource intensive). Already implemented and evaluated in several US hospitals, ESI has yet to be evaluated in a Canadian setting or compared with the five-level *Canadian Emergency Department Triage and Acuity Scale* (CTAS).

Objective: To compare the inter-observer reliability of 2 five-level triage and acuity scales.

Methods: Ten triage nurses, who had all been trained in the use of CTAS, from 4 urban, academic Canadian EDs were randomly assigned either to training in ESI version 3 (ESI v.3) or to refresher training in CTAS. They independently assigned triage scores to 200 emergency cases, unaware of the rating by the other nurses.

Results: Number of years of nursing practice was the only significant demographic difference found between the 2 groups ($p = 0.014$). A quadratically weighted kappa to measure the inter-observer reliability of the CTAS group was 0.91 (0.90, 0.99) and not significantly different from that of the ESI group 0.89 (0.88, 0.99). An inter-test generalizability (G) study performed on the variance components derived from an analysis of variance (ANOVA) revealed G(5) = 0.90 (0.82, 0.99).

Conclusions: After 3 hours of training, experienced triage nurses were able to perform triage assessments using ESI v.3 with the same inter-observer reliability as those with experience and refresher training in using the CTAS.

Key words: triage; acuity; emergency department; randomized controlled trial

RÉSUMÉ

Intoduction : L'Emergency Severity Index (ESI©) est une mesure initiale dans le cadre de l'évaluation des patients à l'urgence. Cet indice permet de classer les patients selon la gravité de leur état,

*Department of Emergency Medicine, Hamilton Health Sciences, Hamilton, Ont.
†Brigham and Womens Hospital, Boston, Mass.
‡Department of Emergency Medicine, York Hospital, York, Pa.
§Department of Clinical Epidemiology, McMaster University, Hamilton, Ont.
¶Feinberg School of Medicine, Northwestern University, Chicago, Ill.

Received: Sept. 15, 2003; final submission: Feb. 9, 2004; accepted: Feb. 27, 2004

This article has been peer reviewed.

This study has been published in abstract form only (*Can J Emerg Med* 2003; 5[3]:183) and presented at the Canadian Association of Emergency Physicians Annual Scientific Assembly, Winnipeg, Man., June 15, 2003. The abstract has also been accepted for presentation at the American College of Emergency Physicians Scientific Assembly, Oct. 13, 2004, Boston, Ma.

Can J Emerg Med 2004;6(4):240-5

https://doi.org/10.1017/S1481803500009192 Published online by Cambridge University Press

la prévision du niveau de ressources nécessaires allant du Niveau 1 (gravité maximale) au Niveau 5 (niveau de ressources le plus bas). Déjà mis en place et évalué dans plusieurs hôpitaux américains, l'ESI n'a pas encore été évalué dans un contexte canadien ni comparé à *l'Échelle canadienne de triage et de gravité pour les départements d'urgence* (ÉTG).

Objectif : Comparer la fiabilité inter-observateurs de deux échelles de triage et de gravité à cinq niveaux.

Méthodes : Dix infirmières de triage de quatre départements d'urgence universitaires canadiens en milieu urbain formées à l'utilisation de l'ÉTG furent assignées au hasard à une formation pour la version 3 de l'ESI (ESI v.3), ou à un recyclage dans l'utilisation de l'ÉTG. Elles assignèrent individuellement des niveaux de triage à 200 cas à l'urgence, sans connaître l'assignation des autres infirmières.

Résultats : Le nombre d'années d'expérience des infirmières était la seule différence démographique rencontrée entre les deux groupes (p = 0,014). Une valeur kappa quadratique pondérée pour mesurer la fiabilité inter-observateurs du groupe ÉTG était de 0,91 (0,90, 0,99) et n'était pas statistiquement différente de celle du groupe ESI qui était de 0,89 (0,88, 0,99). Une étude de généralisabilité inter-tests (G) effectuée sur les composantes de la variance dérivées d'une analyse de variance (ANOVA) révéla G(5) = 0,90 (0,82, 0,99).

Conclusions : Après trois heures de formation, des infirmières de triage expérimentées furent en mesure d'effectuer des évaluations de triage à l'aide de l'ESI avec la même fiabilité inter-observateurs que les infirmières avec de l'expérience et un recyclage dans l'utilisation de l'ÉTG.

Introduction

Triage is typically the first step in the evaluation of a patient presenting to an emergency department (ED). This process involves a brief assessment that focuses on the patient's clinical needs and priority for care.[1,2] The triage nurse then assigns the patient a place in queue and to an appropriate treatment area of the ED based on that assessment.[2]

A variety of ED triage methods are in use. These usually address the issue of acuity and range from 3 to 5 levels. Canada, the United Kingdom and Australia have each adopted a 5-level triage and acuity scale. The *Canadian Emergency Department Triage and Acuity Scale* (CTAS)[3] is based on the Australasian Triage Scale (now called the National Triage Scale [NTS]).[1,4,5] A new triage tool, the Emergency Severity Index (ESI©), has been developed and subsequently modified (ESI v.3) as an initial measure of patient assessment in the ED.[6–8] It is an algorithm with ratings ranging from Level 1 (Most ill) to Level 5 (Least resource intensive). The ESI approach is novel because it incorporates both patient acuity and resource utilization prediction to arrive at a triage level. ESI has been implemented and evaluated in several ED settings in the United States but has yet to be evaluated outside of the US or directly compared with any other triage method.[6,7]

Both NTS and CTAS have been shown, to varying degrees, to be reliable measures of patient acuity.[1,9,10] The 2 published studies measuring the inter-observer reliability of CTAS were based on written summaries of actual ED cases (n = 50, n = 41) and involved groups of 20 subjects unfamiliar with CTAS.[1,11] The former employed equal numbers of emergency physicians (EPs) and emergency nurses (RNs), and the latter used 4 equal groups: EPs, RNs, and 2 groups of paramedics with different levels of training. Using weighted kappas, the inter-observer reliabilities of each study were reported as 0.80 and 0.77 respectively. The inter-observer reliability of ESI was measured by an EP reviewing the triage record of 351 previously ESI-triaged cases and assigning ESI scores while blinded to the original RNs' ESI assessments.[7] The authors reported a weighted kappa of 0.80 for the inter-observer reliability.

To date, there have been no studies comparing the inter-observer reliabilities of 2 or more triage tools. Therefore, the objective of this study is to compare the inter-observer reliability of ESI v.3 to CTAS, using Canadian ED cases.

Methods

This was a randomized controlled trial in which 10 nurses from 4 tertiary-care hospitals were assigned to undergo 3 hours of didactic training in either ESI v.3 or CTAS. The participants were selected by their respective nurse managers according to their availability to participate in this 2-day project. All were trained and experienced in the use of CTAS prior to the study, but none were familiar with ESI. Each group of 5 RNs underwent a 3-hour training (or, in the case of the CTAS group, review) session of the system they were to use.

Randomization was conducted in a stratified manner

https://doi.org/10.1017/S1481803500009192 Published online by Cambridge University Press

such that RNs from each of the 4 hospitals were randomly assigned to 1 of the 2 training groups. The assignment to training group was revealed to the participants after completion of a questionnaire that included demographics, education and work history data. Participants could not be blinded due to past exposure to CTAS.

After completion of training or review, the RNs in each group independently assigned triage levels using their assigned triage tool to 200 case scenarios abstracted from prospectively collected, local ED cases (Fig. 1). Each case included the patient's age and gender, the presenting complaint, and a brief case scenario with vital signs and pain score as documented on the original ED triage sheet. To prevent communication between participants, the groups were observed by a proctor unaware of group assignment.

The data were entered onto a spreadsheet by a staff member who was blind to the study objective. During both data entry and analysis, the group assignment remained coded and concealed.

Two-sample *t* tests were conducted on continuous variables from the nurses' personal information to compare the 2 groups. Inter-observer reliability for each of the 2 groups was measured using a quadratically weighted kappa.[11]

In determining the sample size for this study, we anticipated a kappa value of approximately 0.8 from the previously cited studies and deemed a standard error of 0.05 to

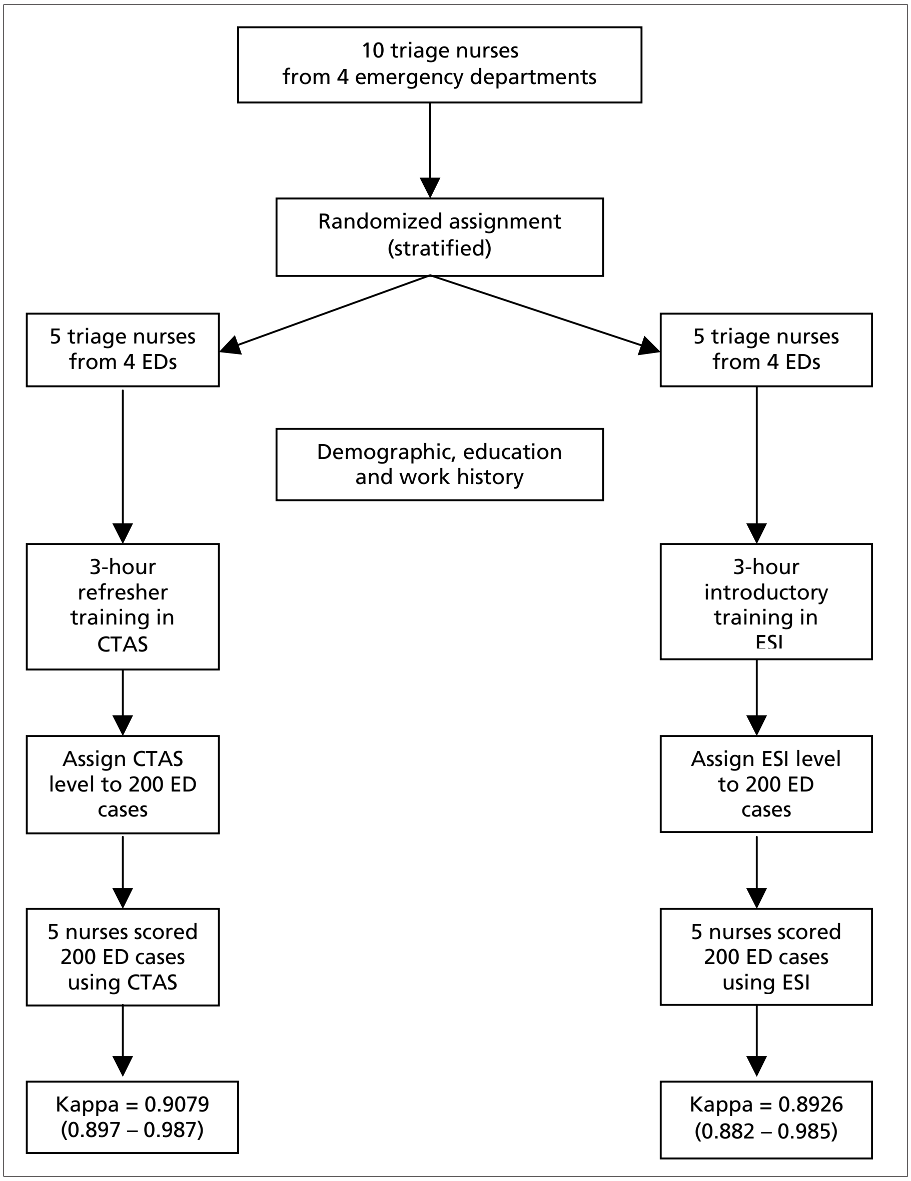

Fig. 1. Reliability study flow chart. CTAS - Canadian ED Triage and Acuity Scale; ESI = Emergency Severity Index

https://doi.org/10.1017/S1481803500009192 Published online by Cambridge University Press

be acceptable. Using the formulae outlined by Streiner and Norman it was determined that approximately 100 cases and 5 observers per case would be required to achieve this small a confidence interval.[12] To ensure the study was sufficiently powered, the number of cases that observers were asked to assess was doubled to 200. Our institution waived the need for institutional review board approval for this study because it did not alter patient care and involved nurse volunteers.

Results

There were no significant differences found between the 2 groups with respect to: age ($p = 0.053$); years in the ED ($p = 0.13$); hours of prior CTAS training ($p = 0.57$); or years of experience in triage ($p = 0.61$). Nine of the 10 RNs had their nursing diploma, and the 10th nurse had a BScN degree. One of the 9 diploma-level RNs also had an Emergency Nursing Certificate. Table 1 shows that there was a significant difference in the number of years in nursing practice (CTAS, 14.4 v. ESI, 25.2, $p = 0.01$) between the 2 groups.

The inter-observer reliability of the CTAS group as measured by the quadratically weighted kappa was 0.91 (0.90, 0.99), similar to that of the ESI group 0.89 (0.88, 0.99).

The 2 triage scales appear to be in moderate agreement with one another, as indicated by an inter-test generalizability of 0.58. An average score assigned by all 5 RNs in each group was 0.90 (0.81, 0.99). The relationship between ESI and CTAS is illustrated in Figure 2.

Discussion

ESI v.3 is a triage-specific, flowchart-based algorithm that asks not only "Who should be seen first?" but also "What does this patient require to reach a disposition?"[7] The first step of the algorithm is to determine if the patient is intubated, apneic, pulseless or unresponsive. These patients are at the highest level of the acuity scale (Level 1). Step 2 of the algorithm addresses all other patients, where they are assessed to determine if they are in a "high-risk situation" – confused, lethargic, disoriented, in severe pain or distress. A patient meeting any one of these criteria is triaged as Level 2 acuity.

The algorithm next focuses on expected ED resources required for a patient. Patients requiring 2 or more distinct ED resources are triaged as a Level 3. However, if their age-adjusted vital signs (heart rate, respiratory rate and oxygen saturation) are not within the described ranges, these patients can be "up-triaged" to Level 2. Patients requiring only 1 ED resource for disposition are triaged as Level 4, and those not requiring any of the listed resources are Level 5.

CTAS is a 5-level triage and acuity scale that categorizes patients as follows: Level I, Resuscitation; Level II, Emergent; Level III, Urgent; Level IV, Less Urgent; and Level V, Non Urgent. It prioritizes patient care requirements by sorting patients according to the type and severity of their presenting signs and symptoms, without consideration of resource utilization. Both ESI and CTAS utilize a descending numeric order of acuity level (i.e., level 1, most acute; level 5, least acute).

Measurement of inter-observer reliability is typically reported using a kappa score for dichotomous variables (agree/disagree). This reports the proportion of agreement between 2 observers given the probability of agreement by chance alone (usually 0.5). When more than 2 choices are available to the observer, such as with these 5-level triage

Table 1. Demographic comparison of the two triage nurse groups

	Group; mean (and SD)		
Variable	CTAS	ESI	*p* value
Age, yr.	38.4 (3.78)	45.8 (5.76)	0.053
Total no. of years of nursing practice	14.4 (4.72)	25.2 (5.72)	0.014
No. of years in ED	8.0 (4.69)	13.8 (5.93)	0.570
No. of years of experience in triage	5.8 (3.35)	7.4 (5.68)	0.610
No. of hours of prior CTAS training	4.8 (3.35)	3.6 (2.97)	0.570

CTAS = Canadian ED Triage and Acuity Scale; ESI = Emergency Severity Index; SD = standard deviation; ED = emergency department

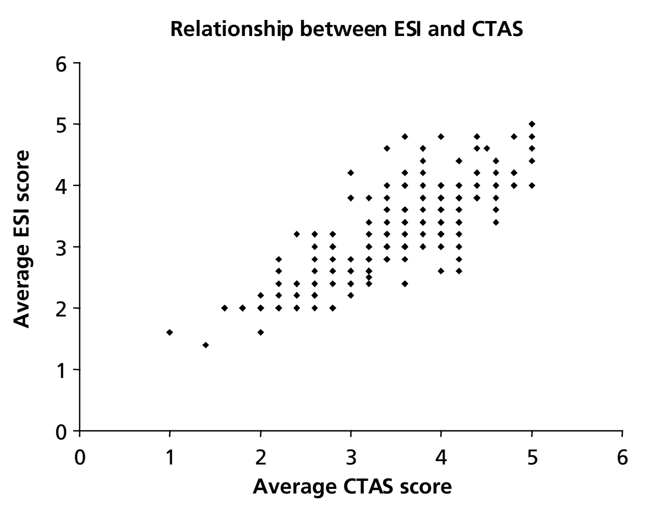

Fig. 2. Relationship between the Emergency Severity Index (ESI) and the Canadian ED Triage and Acuity Scale (CTAS).

https://doi.org/10.1017/S1481803500009192 Published online by Cambridge University Press

scales, the degree to which the observers disagree must also be considered. For example, the difference between one observer triaging a patient as a level 5 and a second observer triaging the same patient as a level 4 is small compared to the difference of the second observer triaging that patient as a level 1. The quadratically weighted kappa measures this degree of disagreement by employing a standard weighting scheme based on the square of the amount of the disagreement.

After 3 hours of training, nurses experienced in triage were able to perform triage assessments using ESI v.3 with the same inter-observer reliability as those with experience and refresher training in CTAS. The higher inter-observer reliability scores reported in this study for both CTAS and ESI (0.91 and 0.89 respectively) compared to the previously published studies (0.8 and 0.77) might be due to the heterogeneity of the participant population of the latter, which compared nurses to physicians (and paramedics), while this study was restricted to nurses.[7,11] Unlike previous CTAS reliability studies, the participants in this study were experienced in using a 5-level triage tool (i.e., CTAS).

Limitations

There are limitations to this study. It was conducted with paper scenarios, not with real patients. This is no different from the previously published inter-observer reliability studies.[1,11] Furthermore, conducting simultaneous, live triage assessments on a single subject with a statistically necessary number of observers is not feasible in most ED settings. The 200 cases used in this study were prospectively and randomly collected specifically for the purpose of the study. Another limitation was that there was only a single CTAS and ESI level 1 case scenario out of the 200. Level 1 cases, (typically, patient is intubated and/or receiving CPR) are immediately recognized by even lay persons as being critical and, therefore, no triage skill is required to determine that the patient is in need of immediate medical attention. Critical, near death patients often do not undergo formal triage but are transferred immediately to a critical care area. Such cases have been excluded previously in triage research.[13] Given that level 1 patients are easily recognized, the inclusion of more such cases (in which the kappa of the inter-observer reliability would approach 1.0) could only increase the final kappa in each triage group without altering any differences between the 2 groups.

As with previously reported reliability studies, the small number of observers in this study might be interpreted as a limitation. However, this number was predetermined with a sample size calculation using appropriate parameters. Although the sample size calculation for this type of analysis determines both the number of cases and the number of observers, the size of the former has a greater impact on the confidence intervals around the summary measures than the latter. The small confidence intervals around the summary measure found in this study are evidence of an adequate sample size. Inclusion of observers from 4 EDs in each group should increase their heterogeneity. The significant difference in number of years of nursing practice found between the 2 groups is a reflection of the small sample of observers (n = 5/group), and it is unlikely that any selection method could have created 2 perfectly equal groups with respect to all 5 of the demographic criteria assessed.

This study was conducted using ED staff, albeit from 4 different hospitals, and cases from a specific setting: tertiary care hospital, urban geographic location, and unique population characteristics of one city in Canada. Although this study may not necessarily apply to other settings, locations or populations, we believe that it can be easily translated.

A final concern is that the previous CTAS training and experience of the participants provided those in the CTAS group with an advantage. Any future study might employ participants unfamiliar with either triage method.

Determining reliability “is a necessary step in establishing the usefulness of a measure” and a reliable triage scale is fundamental to any attempt at performance measurement in emergency medicine.[14] It must also be demonstrated that the tool is valid (i.e., it measures what it is intended to measure and allows us to draw inferences from the scores of the measure with a high degree of confidence).[14] However, validity cannot be assessed until adequate reliability has been proven. A valid measure of ED case-mix data would allow more reasonable assessment of the relationship of such ED case-mix data with DRG (diagnosis-related groups/case-mix groups) information, and other metrics such as that related to operational efficiency, quality indicators and standards, utilization review, outcome effectiveness, patient satisfaction, costs and the like. Sun and colleagues, for instance, have correlated triage and acuity case-mix data with patient satisfaction, and Fernandes and coworkers have linked triage and acuity data with operational efficiency and quality indicators.[15,16]

There have been no other studies comparing 5-level triage instruments. However, in the past decade, at least 4 such tools have been developed or implemented on large scales, and each is based on different strategies. Comparison studies are, therefore, necessary to determine the relative merits of each in different populations. We have chosen to compare the 2 most utilized 5-level triage instruments in North America. The first step in this comparison is an assessment of reliability for both scales. Our

https://doi.org/10.1017/S1481803500009192 Published online by Cambridge University Press

study suggests a comparable level of inter-observer reliability. We believe that this study should be repeated in other settings and with other triage scales.

Conclusions

With minimal training, a group of experienced ED triage nurses were able to perform triage assessments on a set of standardized case scenarios using ESI v.3 and obtain the same inter-observer reliability values as a group of CTAS-experienced and refresher-trained nurses using CTAS.

Acknowledgements: This study was generously funded by an unrestricted educational grant from Hoffman–LaRoche Ltd.

We are grateful to Brandon Ellis-Hale and Hoffman–La Roche Ltd. for their generous support of this project and to the skilled triage nurses for their enthusiastic participation, which made this study possible.

Competing interests: Dr. David Eitel is a co-developer of the Emergency Severity Index, 1 of the 2 triage tools assessed in this study. Dr. Eitel does not receive any financial remuneration for developing this tool and was not involved in the data collection process or analysis.

References

1. Beveridge R, Ducharme J, Janes L, Beaulieu S, Walter S. Reliability of the Canadian emergency department triage and acuity scale: interrater agreement. Ann Emerg Med 1999;34(2):155-9.
2. Wuerz R, Fernandes CM, Alarcon J. Inconsistency of emergency department triage. Emergency Department Operations Research Working Group. Ann Emerg Med 1998;32(4):431-5.
3. Beveridge R, Clarke B, Janes L, Savage N, Thompson J, Dodd G, et al. Canadian Emergency Department Triage and Acuity Scale: implementation guidelines. CJEM 1999;1(3 Suppl). Online version available at: www.caep.ca/002.policies/002-02.ctas.htm (accessed 19 May 2004).
4. MacKway-Jones K, editor. Manchester Triage Group. Emergency Triage. London: BMJ Publishing Group; 1997.
5. The Australasian Triage Scale [policy document]. Australasian College for Emergency Medicine. Emerg Med 1994;6:145-6.
6. Wuerz RC, Milne LW, Eitel DR, Travers D, Gilboy N. Reliability and validity of a new five-level triage instrument. Acad Emerg Med 2000;7(3):236-42.
7. Eitel DR, Travers DA, Rosenau AM, Gilboy N, Wuerz RC. The emergency severity index triage algorithm version 2 is reliable and valid. Acad Emerg Med 2003;10(10):1070-80.
8. Tanabe P, Gimbel R, Yarnold PR, Kyriacou DN, Adams JG. Reliability and validity of scores on The Emergency Severity Index version 3. Acad Emerg Med 2004;11(1):59-65.
9. Jelinek GA, Little M. Inter-rater reliability of the National Triage Scale of 11,500 simulated occasions of triage. Emerg Med 1996;8:226-30.
10. Hollis G, Sprivulis P. Reliability of the National Triage Scale with changes in emergency department acuity level. Emerg Med 1996;8:231-4.
11. Manos D, Petrie DA, Beveridge RC, Walter S, Ducharme J. Interobserver agreement using the Canadian Emergency Department Triage and Acuity Scale. Can J Emerg Med 2002;4(1);16-22.
12. Streiner DL, Norman GR. Health measurement scales: a pracical guide to their development and use. 3rd ed. New York: Oxford University Press; 2003. p. 148-51.
13. 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. Ann Emerg Med 1996;27(4):493-500.
14. Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use, 2nd ed. Oxford: Oxford University Press; 1995. p. 144-6.
15. Sun BC, Adams JG, Burstin HR. Validating a model of patient satisfaction with emergency care. Ann Emerg Med 2001;38: 527-32.
16. Fernandes CM, Wuerz R, Clark S, Djurdjev O. How reliable is emergency department triage? Ann Emerg Med 1999;34(5):141-7.

Correspondence to: Dr. Andrew Worster, Research Director, Department of Emergency Medicine, Hamilton Health Sciences, 237 Barton St. E, Hamilton ON L8N 3Z5; 905 527-4322 x46997, fax 905 527-7051, aworster@rogers.com

https://doi.org/10.1017/S1481803500009192 Published online by Cambridge University Press