CrossMark

Comparison of emergency department crowding scores: a discrete-event simulation approach

Virginia Ahalt[1] · Nilay Tanık Argon[1] · Serhan Ziya[1] · Jeff Strickler[2,3] · Abhi Mehrotra[2]

Abstract According to American College of Emergency Physicians, emergency department (ED) crowding occurs when the identified need for emergency services exceeds available resources for patient care in the ED, hospital, or both. ED crowding is a widely reported problem and several crowding scores are proposed to quantify crowding using hospital and patient data as inputs for assisting healthcare professionals in anticipating imminent crowding problems. Using data from a large academic hospital in North Carolina, we evaluate three crowding scores, namely, EDWIN, NEDOCS, and READI by assessing strengths and weaknesses of each score, particularly their predictive power. We perform these evaluations by first building a discrete-event simulation model of the ED, validating the results of the simulation model against observations at the ED under consideration, and utilizing the model results to investigate each of the three ED crowding scores under normal operating conditions and under two simulated outbreak scenarios in the ED. We conclude that, for this hospital, both EDWIN and NEDOCS prove to be helpful measures of current ED crowdedness, and both scores demonstrate the ability to anticipate impending crowdedness. Utilizing both EDWIN and NEDOCS scores in combination with the threshold values proposed in this work could provide a real-time alert for clinicians to anticipate impending crowding, which could lead to better preparation and eventually better patient care outcomes.

Keywords Emergency department crowding · Discrete-event simulation · Queueing model · Edwin · NEDOCS · READI

1 Introduction

Emergency department (ED) crowding is a widely reported problem and may adversely affect patient care [10, 13]. There is currently no universally accepted definition for an ED being in a "crowded" or "overcrowded" state [4, 15]. The existing definitions are rather vague and do not have any time metrics associated. For example, the American College of Emergency Physicians (http://www.acep.org/Clinical—Practice-Management/Crowding) define ED crowding as an event that occurs when the identified need for emergency services exceeds available resources for patient care in the ED, hospital, or both. There have been several attempts in the literature to provide a quantitative measure for ED crowding. Patient counts has emerged as a basic tool for measuring the non-flow of patients through the ED (i.e., patient crowding) [8], and there have been several more advanced strategies to quantify the crowdedness in an ED by developing crowding scores [16]. These crowding scores provide an assessment of the current crowding level in an ED, and allow healthcare professionals to anticipate imminent crowding problems and make better resource allocation and staffing decisions [5]. Three of the more widely used scores in the U.S. are the National Emergency Department Overcrowding Scale (NEDOCS), the Emergency Department Work Index (EDWIN), and the Real-time Emergency Analysis of Demand Indicator (READI) [9, 11]. The formula for each of

✉ Nilay Tanık Argon
nilay@unc.edu

1 Department of Statistics and Operations Research, University of North Carolina, Chapel Hill, NC, USA

2 Department of Emergency Medicine, University of North Carolina School of Medicine, Chapel Hill, NC, USA

3 University of North Carolina Health Care, Chapel Hill, NC, USA

these scores takes into account various patient, ED, and hospital information, and yields a numerical value that indicates whether an ED is running smoothly, is crowded but effective, or is overcrowded at a given point in time.

Since each crowding score takes in specific hospital and patient characteristics and is judged against its own corresponding measurement scale, there are general scalability concerns of the various scoring systems [8, 9]. For example, a scoring system that performs well at a rural hospital with a small ED may not perform well at a large, urban hospital system with an extremely busy ED. Regardless of scalability concerns, a "successful" crowding score should assist medical professionals through the measurement of current ED conditions, detection of abnormal ED operations, and anticipation of increased crowding levels. Measuring current crowding and alerting the ED to predicted overcrowding could assist medical professionals to make real-time operational changes thereby improving patient access to care [6].

In this study, we consider a large academic hospital in North Carolina, where the EDWIN score was used to assess the level of crowdedness in the ED during the time this study was conducted. Our main objective is to explore if NEDOCS or READI would be a more informative crowding score at this North Carolina hospital and hospitals alike and if so how they should be utilized to measure the crowding most effectively. For this purpose, we built a discrete-event simulation (DES) model of the ED using hospital and patient data collected during January 7 to February 3, 2013. As established in [1], emergency department crowding can be analyzed using an input-throughput-output approach, where factors contributing to ED crowding are either inputs, throughputs, or outputs. This view of the ED and ED crowding is the most widely accepted [8], and discrete-event simulation can use this approach to model the operations at an emergency department as a sequence of discrete events. Discrete-event simulation has been also well established as a modeling approach for ED patient flow and crowding research, see, e.g., [1, 3], [7, 18]. Use of simulation is especially suitable to answer what-if questions which cannot be answered using historical data or the real system.

After building our simulation model, we first used visual tools and a statistical test to validate it as a reasonable model that yields output similar to the actual system. Using both the actual and simulated data, we then observed how each score behaves during the course of a typical day with normal operating conditions. Later, we simulate the ED to test the predictive power of the three ED crowding scores under two hypothetical scenarios where the ED faces a higher-than-usual patient demand over the course of four days, e.g., due to a short cold/flu outbreak. In particular, we compare each score in terms of the probability that the score will detect the presence of an unusual load. We also propose that each ED score should be compared to a high percentile threshold based on historical data from the ED under consideration to detect an overcrowding event instead of comparing the score to a fixed value as proposed in the literature. Using historical data from the ED, this proposed approach takes into account the characteristics of different EDs and hospitals using the same score. Our simulation experiments show that when used together with our proposed threshold, EDWIN and NEDOCS perform well in detecting abnormal patient loads.

The outline of the article is as follows. In Section 2, we provide a background and literature review on the three crowding scores under consideration. In Section 3, we discuss the data and the ED operations in the hospital under study. Section 4 introduces our simulation model and discusses the statistical analysis of the input data. In Section 5, we present results on the validation of our simulation model and provide observations on how the three crowding scores behave under normal operating conditions. In Section 6, we utilize the model results and statistical analysis to evaluate how well EDWIN, NEDOCS, and READI detect the onset of a hypothetical event that results in an increased rate of arrivals. Finally, in Section 7 we discuss our conclusions and limitations of this study.

2 Background on ED crowding scores

We start by defining the three crowding scores of interest. The variables used in each formula are described in detail in Table 1.

The formula for calculating the EDWIN score is given as follows [2]:

$$EDWIN = \frac{\sum_{i=1}^{5} n_i t_i}{N(B_T - B_A)}$$

where the numerator represents a weighted sum of the number of patients from all five triage categories present in the ED. Here, severity level t_i is the reverse of the ith level of severity as determined by the Emergency Severity Index (ESI), i.e., $t_i = 6 - i$. (ESI is one of the most widely used triage systems in US hospitals, see, e.g., [14].) According to ESI, a patient who receives a smaller index level is more urgent. Hence, for the score value to increase with increased severity of illness, t_i is set to $(6 - i)$. Once an ED patient has been admitted to the hospital, i.e., a patient who is occupying an ED bed while awaiting an open inpatient bed, he or she is considered a "hold" and is no longer included in the numerator of the EDWIN score. The denominator of the EDWIN score, on the other hand, represents the total ED capacity as a product of the number of attending physicians (N) and the total number of available beds (excluding the "hold" beds). It is suggested in the literature that an EDWIN score of less than 1.5 indicates an "active but manageable" ED, 1.5 to 2 represents a busy ED, and a score of over 2 corresponds to an overcrowded ED [2].

Table 1 Variables utilized in crowding scores

Details		Used in		
Variable	Description	EDWIN	NEDOCS	READI
i	ESI level of severity	✓		✓
n_i	Number of patients in the ED that fall into the ith ESI level	✓		✓
t_i	Reversed severity level, i.e., $t_i = 6 - i$	✓		✓
N	Total number of attending physicians working in the ED	✓		
M	Total number of patients in the ED, i.e., $\sum_{i=1}^{5} n_i$		✓	
B_T	Total number of treatment beds in the ED	✓	✓	✓
B_A	Patient holds, i.e., number of patients who have been admitted to the hospital but are occupying an ED bed while awaiting an open inpatient bed	✓	✓	
B_h	Total number of hospital (non-ED) beds		✓	
W	Most recent waiting time for an ED Bed		✓	
A_{time}	Longest current wait of admitted patients for an inpatient bed to open		✓	
R_h	Number of ED patients on a ventilator		✓	
β	A real number from 13.4 to 20, which is set by the user		✓	
P_A	Predicted arrivals for each hour of the day			✓
P_D	Predicted departures for each hour of the day			✓
P_{bed}	Number of patients in ED beds			✓
A_h	Number of arrivals per hour			✓
H_j	Number of patients seen per hour by physician j, as determined by historical ED data			✓
BR	Bed ratio, $BR = (P_{bed} + P_A - P_D)/B_T$			✓
PR	Provider ratio, $PR = A_h / \sum_j H_j$			✓
AR	Acuity ratio, $AR = \sum_i n_i t_i / P_{bed}$			✓

The NEDOCS score, which is based on a linear regression model, is given as follows [17]:

$$NEDOCS = 85.8 \times \left(\frac{M}{B_T}\right) + 600 \times \left(\frac{B_A}{B_h}\right) + 5.64 \times W$$
$$+ 0.93 \times A_{time} + R_h \times \beta$$

where the first variable $\left(\frac{M}{B_T}\right)$ is the ratio of total numbers of ED patients and beds, the second variable $\left(\frac{B_A}{B_h}\right)$ is the ratio of the numbers of "hold" and inpatient beds, the third variable (W) is the most recent waiting time for an ED bed, the fourth variable (A_{time}) is the longest current boarding time, and the fifth variable (R_h) is the number of ED patients on a ventilator. (The weight for the number of patients on a ventilator (β) is left as a flexible parameter that could be set to a value between 13.4 and 20 depending on the importance of this variable for the ED under consideration.) The NEDOCS score is evaluated on a scale from 1 to 200, where a higher score indicates a higher congestion level. In particular, [17] suggests that a score of 100 or larger indicates an overcrowded ED.

Finally, the READI score is defined as follows [12]:

$$READI = (BR + PR) \times AR$$

where the bed ratio (BR) is the ratio of predicted number of ED patients at the top of the next hour and the total number of ED beds, the provider ratio (PR) is the ratio of number of arrivals per hour and the number of patients seen per hour by the ED physicians, and the acuity ratio (AR) is given by the ratio of a weighted sum of the number of patients from all five triage categories at the ED (the same as the numerator of the EDWIN score) and the total number of patients at the ED [9, 11]. (The exact calculations for BR, PR, and AR are given in Table 1.) A READI score value of greater than 7 is said to indicate an overcrowded ED [12].

A common criticism of these three ED crowding scores is that they are difficult to assess for accuracy. In order to evaluate their abilities [9, 11] compared clinicians' perceptions of crowding to the results reported by the crowding scores, where clinician perspectives have been generally gathered through survey. A more objective standard of assessment for these scores is not available [11]. In [16], the EDWIN and NEDOCS scores were calculated every two hours, and compared against physicians and nurses ratings of crowding as measured by a visual assessment scale. EDWIN and NEDOCS demonstrated a high correlation with these clinicians' perceptions [16], but in other studies, READI does not seem to provide reliable results which agree with

clinicians' perceptions of crowding [9, 11]. Additionally, the published threshold values for each crowding score may not align with clinician' perceptions of crowding. However, NEDOCS, and to a slightly lesser extent EDWIN, have demonstrated strong predictive power for ED crowding, particularly through the correlation with ambulance diversion as an indicator of increased ED crowding [9, 16].

3 Data

The data for this study came from a large public, academic medical center that provides tertiary care. It is one of only six Level I Trauma Centers in the state of North Carolina. The hospital system has an active residency program and approximately 800 inpatient beds. The ED saw approximately 68,000 patients in calendar year 2013, during which the data used in this study was collected, and sees some of the highest acuity patients in the state.

The ED at this institution is divided into several sections based on operating hours, patient type, and patient severity. The two main sections, A and B, remain open during all hours of the day and see primarily acute patients. Less acute patients are seen in two additional sections, C and D, during the regular working hours of the ED (9:00 a.m. to 2:00 a.m.). Pediatric patients (those under the age of 18) are seen in a specific pediatric section during regular working hours. During non-regular hours, all patients are seen in sections A and B. Sections C and D have separate physician and nurse teams from A and B, and C is used primarily for patient holds and patients that are receiving medicine (such as IV medicine) but do not require more resource-intensive care. (This study excluded behavioral health [psychiatric] patients who are seen at a separate area from the main ED sections.) Based on the ED regular working hours, there are between 41 and 65 ED beds available at any given hour of the day.

The hospital provided hourly scheduling data for residents, physicians, and nurses, as well as specific patient data (excluding data for behavioral health patients) for January 7 to February 3, 2013. The data for each patient seen during this time period consisted of the patient's triage category as designated by the Emergency Severity Index (ESI) as in [14] and four key timestamps:

- **arrival time:** patient arrives to the ED, either by ambulance or walk-in;
- **bed time**: patient is assigned to an ED bed or treatment space;
- **disposition decision time:** physician or resident makes the decision to admit the patient to an inpatient hospital bed, or to complete medical care and send the patient home;
- **discharge time**: patient leaves ED, either for admission to an inpatient hospital bed or to depart for home.

The patient flow through the ED, as well as the associated timestamps, are shown in Fig. 1. For each section of the ED, the corresponding number of available beds is given in parentheses.

According to the provided data, the ED saw 5100 patients during the chosen time period of January 7 to February 3, 2013. From this data, approximately 4 % were excluded as bad data. Possible reasons for exclusions included missing timestamps, invalid acuity scores, data suggesting departure before arrival or other out-of-order timestamps, and unrealistic total service times (one minute from arrival to discharge, for example). We used the cleaned data to estimate distributions and parameters that are needed in the simulation model and also in the calculation of the crowding scores. We found that 1 %, 13 %, 60 %, 22 %, and 4 % of all adult patients fell into ESI categories 1, 2, 3, 4, and 5, respectively, whereas the respective percentages for pediatric patients were 0.5 %, 11 %, 42 %, 40 %, and 6.5 %. About 20 % of all incoming patients were estimated to be pediatric patients. We also found that almost all ESI 1–2 and none of ESI 5 adult patients were admitted to the hospital, whereas 30 % and 3 % of ESI 3 and 4 adult patients, respectively, needed inpatient care and hence were admitted to a
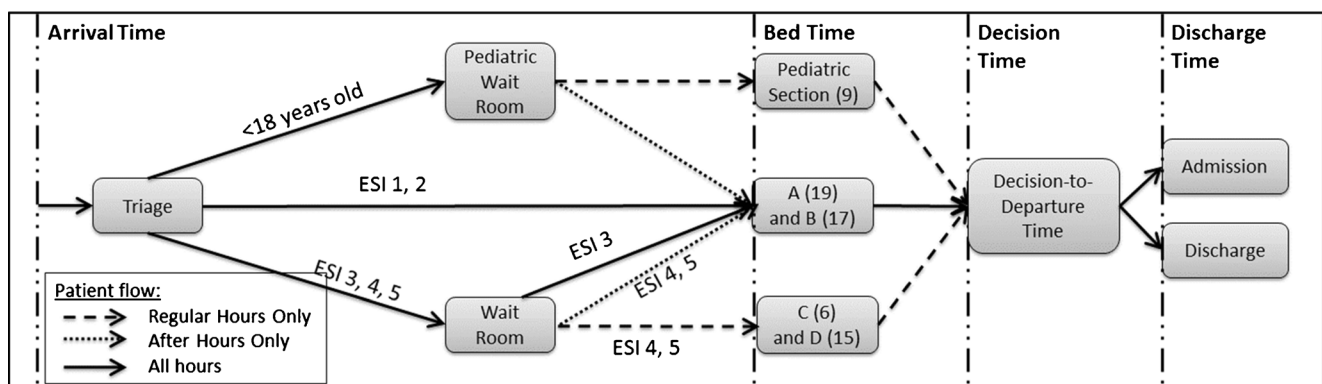

Fig. 1 Patient flow through emergency department

hospital ward. For pediatric patients, the admission rates were similar except that only about 18 % of pediatric ESI 3 patients were admitted to an inpatient hospital ward.

The existing patient tracking system at this hospital did not provide explicit data on the use of ventilators, which is needed in the NEDOCS formula. Therefore, we used the average number of patients who are later admitted to an ICU as a proxy for the average number of patients who were on a ventilator. From the data, we found that around 2 % of all patients were later admitted to an ICU and on average there were around 66 patients in the ED. Based on these figures, we estimated the average number of patients who occupy the ED and are later admitted to an ICU as 1.3. This is consistent with the ED management's expectation of having on average one patient on a ventilator at all times. In addition, in the NEDOCS for-mula, we set β (the weight for the number of patients on a ventilator) to 13.4, the lowest possible value, because the number of patients on a ventilator does not seem to contribute drastically to crowding at this ED. Finally, based on the ED management's observations, on average a physician sees two to three patients per hour at this ED. (To be conservative, we set this provider productivity rate to two patients per hour in READI score calculations, which is the only place it is needed in this study.) We provide further details on the input data analysis in Section 4.

4 Discrete-event simulation model and input analysis

Using the provided hospital, ED, and patient data, we first created a discrete-event simulation model of the ED using Arena simulation software. In particular, we modeled the ED operations as a queuing model with multiple classes of pa-tients that seek service from multiple, and possibly non-iden-tical, resources (beds). Arrivals to this queueing system are non-stationary, i.e., the rate of arrivals depend on the hour of the day, and service times consist of two phases: the time interval between bed time and disposition decision time, and the time interval between disposition decision time and dis-charge time. By simulating this model, we produce the ESI levels and four critical timestamps of each simulated patient, which is then incorporated with the remaining staffing and hospital data, and used to calculate the crowding score values for analysis.

In our simulation model, a simulated patient encounters the following processes and decisions in the given order: Arrival to the system; decision made for allocation of appropriate beds based on age and ESI level; join queue for an ED bed; enter bed and incur bed-to-decision service time; decision made for admit or discharge; incur decision-to-discharge service time; and depart ED. We used the raw patient data to fit distributions to the arrival process and the service time distributions, and incorporate the results into the model. Goodness of fit for

testing various distribution functions is assessed using the Kolmogorov-Smirnov (K-S) test and the squared-error value. In the remainder of this section, we provide details on how we fit distributions to the interarrival and service times.

4.1 Arrival process

The interarrival times (to the nearest minute) are calculated using sample means of all patient arrivals (without separating according to ESI, gender, etc.). A distribution is later fit to this interarrival data by evaluating hourly groupings of interarrival times. We found that an exponential distribution provided a good fit for the interarrival data with with p-values of greater than 0.13 using the K-S test and squared-error values of less than 0.003. (Note that a large p-value for the K-S test and a small squared-error value are indicative of a good fit.) Therefore, we concluded that the arrival process can be ap-proximated by a Poisson process. The hourly rates for the patient arrival schedule in the Arena model were determined by using the average hourly arrival rate seen in the month of patient data, and can be seen in Table 2 and also visually observed in Fig. 3.

4.2 Service time distribution

Based on the preliminary analysis of the data and also taking into account the operational structure at the ED, we model the total time a patient spends in an ED bed as the sum of two separate service times. Part 1 of service begins when the pa-tient is assigned a bed (i.e., bed time) and ends at the point in time which a patient receives a decision regarding their admis-sion to inpatient care (i.e., disposition decision time). Part 2 of service begins at disposition time and ends when the patient departs the system, either to go home or to transfer to an inpatient bed. This division of the total service requires the fitting of two separate distributions in order to describe the service process.

Table 2 Patient arrival rates (patients/h)

Hour	Rate	Hour	Rate
2 a.m. – 3 a.m.	3.0	2 p.m. – 3 p.m.	10.2
3 a.m. – 4 a.m.	2.9	3 p.m. – 4 p.m.	12.0
4 a.m. – 5 a.m.	2.3	4 p.m. – 5 p.m.	11.1
5 a.m. – 6 a.m.	2.3	5 p.m. – 6 p.m.	11.1
6 a.m. – 7 a.m.	2.2	6 p.m. – 7 p.m.	10.5
7 a.m. – 8 a.m.	3.6	7 p.m. – 8 p.m.	9.7
8 a.m. – 9 a.m.	6.5	8 p.m. – 9 p.m.	9.0
9 a.m. – 10 a.m.	9.0	9 p.m. – 10 p.m.	6.4
10 a.m. – 11 a.m.	10.9	10 p.m. – 11 p.m.	6.1
11 a.m. – 12 p.m.	11.3	11 p.m. – 12 p.m.	6.0
12 p.m. – 1 p.m.	11.8	12 a.m. – 1 a.m.	4.4
1 p.m. – 2 p.m.	11.5	1 a.m. – 2 a.m.	3.7
Shading indicates regular business hours			

It is reasonable to assume that patients with different severity levels may require different distributions of service times. Additionally, the availability of resources (in this case beds, physicians/residents, and nurses) also affects service times, and varies throughout the day. For these reasons, we tested several combinations of hourly and ESI divisions of the service data for goodness of fit under a number of distribution functions (namely, Beta, Erlang, Exponential, Gamma, Johnson, Lognormal, Normal, Triangular, Uniform, and Weibull). The results of the best fit are presented in Table 3, which we used to establish Part 1 and Part 2 service times in the simulation model.

The hours for the Part 1 service distributions were determined based on the staffing levels and hours of the ED sections. From 2:00 a.m. to 9:00 a.m., there are fewer physicians, residents, and nurses working, compared to the regular working hours. Since the regular working hours (9:00 a.m. to 2:00 a.m.) see a variety of patient arrival rates, this large time period was split in half to allow for better fits to the Part 1 service data. Splitting the Part 1 service by ESI in addition to arrival hour also mimics the true behavior of the ED since ESI 1 and 2 patients are sent to different sections than ESI 3, 4, and 5 patients. We see in Table 3 that the best fit for all Part 1 service distributions have small squared-error values and large p-values for the K-S test.

For Part 2 of service, we first split the patients into admitted and non-admitted patients. After the disposition decision has been made, patients who will not be admitted to inpatient care from the ED wait under an hour before they are discharged to home, whereas admitted patients wait on average four to six hours for an inpatient bed to become available. This drastic difference in wait times for admitted versus non-admitted patients implies a natural separation when considering fits for Part 2 of service. The admitted patients are further divided by ESI, as ESI 1 and 2 patients were admitted to inpatient care faster than the less medically severe ESI 3, 4, and 5 patients. For the given distributions in Table 3 that are fit to Part 2 service time data, although the p-values for the K-S test were not high, the squared-error values turned out to be at most 0.009, which supports that these distributions provide a reasonable fit to the data.

5 Validation of the simulation model

After building our simulation model in Arena using the input distributions and parameters from Section 4, we first validated the model as a reasonable recreation of this ED. For this purpose, we used both statistical and visual tools.

A statistical test for validation Using our simulation model, we conducted a single replication of 28 days after a warm-up period of 365 days. (The actual patient data provided to us was collected over 28 days as well.) We then calculated the average length of stay (LOS) for all patients in the ED and the average EDWIN, NEDOCS, and READI scores during the busy hours of 4 pm to midnight each day, using both the actual and simulated data. To reduce the effects of any autocorrelation present in these data sequences, we used observations for every other day, which resulted in 14 data points per sequence. Then, we applied paired t-tests to test whether the differences between the mean performance measures for the actual and simulated systems were zero. We obtained the following 95 % confidence intervals on the mean differences: $(-0.093, 0.663)$ hours for LOS, $(-0.143, 0.087)$ for EDWIN, $(-0.496, 17.254)$ for NEDOCS, and $(-0.432, 0.992)$ for READI; resulting in

Table 3 Service Time Distributions by ESI Level, Hour, Patient Type. [WEIB(α, β), GAMM(α, β), and BETA(α, β) represent a Weibull, Gamma, and Beta distribution, respectively, with shape parameter α and scale parameter β in minutes]

Service Part 1					
Arrival Hour	ESI	Best Fit	p-value	Squared-Error	Sample Size
2:00 a.m. - 9:00 a.m.	1, 2	9 + WEIB(260,1.1)	>0.15	0.018	83
2:00 a.m. - 9:00 a.m.	3, 4, 5	ERLA(123,2)	>0.15	0.003	560
9:00 a.m. - 2:00 p.m.	1, 2	5 + ERLA(133,2)	>0.15	0.008	163
9:00 a.m. - 2:00 p.m.	3, 4, 5	GAMM(104,2.37)	0.145	0.001	1399
2:00 p.m. - 8:00 p.m.	1, 2	GAMM(209,1.46)	0.136	0.014	259
2:00 p.m. - 8:00 p.m.	3, 4, 5	ERLA(91.8,3)	>0.15	0.001	1551
8:00 p.m. - 2:00 a.m.	1, 2	5 + GAMM(238,1.34)	>0.15	0.020	130
8:00 p.m. - 2:00 a.m.	3, 4, 5	GAMM(116,2.38)	>0.15	0.002	782
Service Part 2					
Departure Type	ESI	Best Fit	p-value	Squared-Error	Sample Size
Non-Admitted	All	0.5 + 44(BETA(0.682,1.29))	<0.005	0.007	3016
Admitted	1, 2	45 + WEIB(192,1.01)	0.0234	0.009	476
Admitted	3, 4, 5	45 + GAMM(215, 0.723)	0.0318	0.001	1435

respective p-values of 0.13, 0.61, 0.25, and 0.41. These results provide statistical support that our simulation model is valid in terms of the average LOS and the three crowding scores under consideration during busy hours.

Visual comparison of output data from the simulation model and actual system We simulated the ED over artificial dates of January 8–15, 2013, after a warm-up period of 24 h, for a total of 100 replications, and used the output to calculate three crowding scores as well as the average LOS of all patients in the ED by every hour of the day for each of the 100 replications. Figure 2 shows the average crowding scores and LOS based on 100 replications of the given week. Solid horizontal lines for each crowding score indicate the corresponding threshold value, as recommended in [11, 14, 17], above which indicates a crowded scenario and under which indicates normal operating conditions. The average length of stay is estimated by dividing the total time spent (from arrival to discharge) for all current patients in the ED by the number of current patients, captured at each hour over the given week. We calculated a 95 % confidence interval on the simulated average crowding scores and LOS but plot only the mean in Fig. 2 since the confidence intervals were small. Figure 2 also plots the crowding scores and LOS from the true patient data corresponding to the given week.

Figure 2 shows that the simulation model provides a reasonable approximation of ED crowding behavior observed during the simulated time period. The real patient data over January 7–13 is significantly less smooth as compared to the simulation model data, as the simulation model shows the average of 100 trials of the given week. However, the simulation output and real data match very well in terms of the mean scores and their general trend.

In addition to providing a visual comparison for the output from the actual and simulated systems, Fig. 2 is also useful for generating insights into the performance of the three crowding scores under consideration. In Fig. 2, the EDWIN score can be observed to abruptly spike at around 2:00 a.m. and drop sharply at around 9:00 a.m. on each day of the given time period in both the real patient data and the simulated data, which agrees with what this hospital had observed from its own EDWIN reporting system. The staffing numbers and ED section hours undergo sharp changes at these hours which are causing these spikes in the EDWIN score. Both the "number of physicians" and the "number of ED beds" terms are in the denominator of the EDWIN score, hence at 2:00 a.m. the drop in physicians and beds inflate the EDWIN score. Similarly, at 9:00 a.m. the number of physicians and beds return to daytime levels, which drops the EDWIN score.

From Fig. 2, we also observe that EDWIN follows a trend that is similar to the average LOS curves, especially regarding the time it peaks every day. Although we know the drastic spiking to be artificial due to staffing changes, EDWIN suggests an increase in crowdedness during the early morning hours of each day. These early morning hours generally do not see a high number of patients. The highest peaks of the NEDOCS score values occur before the highest peaks for the average LOS. The READI score values generated from the true patient data appears far too jagged to draw strong conclusions about the current or impending level of crowdedness. The simulation results show a smoother READI score due to averaging over 100 replications, but the range of score change is still less significant than in EDWIN or NEDOCS. Also, like NEDOCS, READI appears to peak before the average LOS peaks.
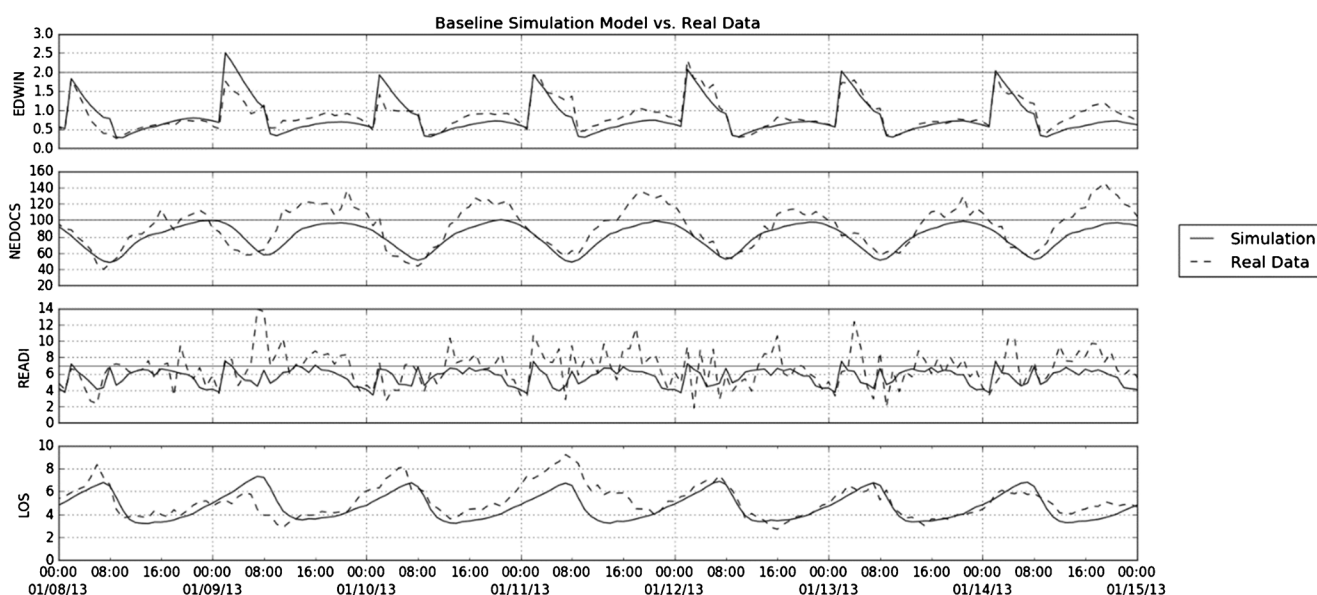

Fig. 2 Crowding score results and average length of stay from the data and simulation model

6 Simulation results – under high patient demand

In this section, we adapt our baseline model to simulate "unusual" conditions in the ED, specifically an event that causes a spike in the number of patient arrivals to the ED over several days. An outbreak of a cold or a flu-like illness could cause such an increase in arrival rates and we would like to explore how each crowding score would perform in terms of detecting its occurrence. We consider two hypothetical scenarios – one mild and one extreme – under which the arrival rates were artificially increased.

6.1 Mildly loaded scenario

To simulate a mild outbreak, a gradual increase in the number of patients seen based on a percentage of the normal patient arrivals is used over a period of several days. In particular, we assume that the ED may see its first patient due to this outbreak between 8 and 9 a.m. on the first day of the week, where the patient arrival rate is about 105 % of the normal rate. The patient arrival rates increase to 125 % of the normal hourly arrival rates through days 1–3, and on day 4, the outbreak is "eradicated" and patient arrival rates gradually decrease back to the hourly daily patient flows. The exact arrival rate changes by hour are shown in Table 4 and are also plotted in Fig. 3 for visual comparison to arrival rates under normal operating conditions.

To analyze the success of each crowding score at predicting and detecting the onset of such an outbreak, we calculate an hourly threshold score value for each score that mimics how medical professionals may evaluate the crowding scores in practice. As the relevance of each crowding score's interpretation scale varies by hospital characteristics, in practice, medical professionals compare crowding score values to what they have seen historically to assess the current state of the ED. To provide an objective comparison that captures this historical score assessment analytically, we took an alternative approach and estimated a 90th percentile score value for each hour of the day for each score. (Depending on the preferences of the

ED management, one could also use different upper percentiles instead of the 90th percentile.) For this, we simulated the ED system for six months under normal operating conditions and grouped the resulting values of each score by hour, ranked each hourly grouping in ascending order, and determined, for each hour, the threshold value x under which 90 % of the observations fell. This generated a unique threshold value x for each hour for each of the three scores. In real time, a similar hourly threshold based on historical data could be incorporated into ED monitoring, where a crowding score surpassing the threshold value alerts medical professionals to a potential crowding situation. Figure 4 shows the average simulated crowding scores from 100 trials of the mild outbreak scenario plotted against the calculated 90th percentile hourly threshold, along with the recommended threshold values of 2, 100, and 7 for EDWIN, NEDOCS, and READI, respectively.

Figure 4 shows that the average READI score in the mild outbreak scenario falls consistently under the 90th percentile threshold, not raising an alarm as to the presence of an outbreak. Both the average EDWIN and NEDOCS crowding scores first exceed the 90th percentile threshold in the early morning hours of Day 2, less than 24 h after the true onset of the outbreak. However, EDWIN stays consistently over the threshold starting at 5:00 AM on Day 2, whereas NEDOCS fluctuates above and below the threshold before remaining above the threshold starting at 2:00 AM on Day 3. Additionally, the average EDWIN score comes very close to the 90th percentile threshold value as early as around noon on the first day: only 4 h after the true start of incoming outbreak patients in excess of a normal patient load.

Figure 4 also plots the recommended time-independent threshold values for EDWIN, NEDOCS, and READI. In the EDWIN score, the 90th percentile threshold and the average score under the outbreak scenario only seem to surpass the published threshold value of 2 in the hours 1–8 of each day. More importantly, this may not be regarded as a sign of an outbreak because due to the unique staffing levels at this ED, a similar spiking behavior in the EDWIN score is observed even under normal operating conditions (see Section 5). Hence, an

Table 4 Mildly loaded hourly patient arrival rates as percentage of normal hourly patient arrival rates

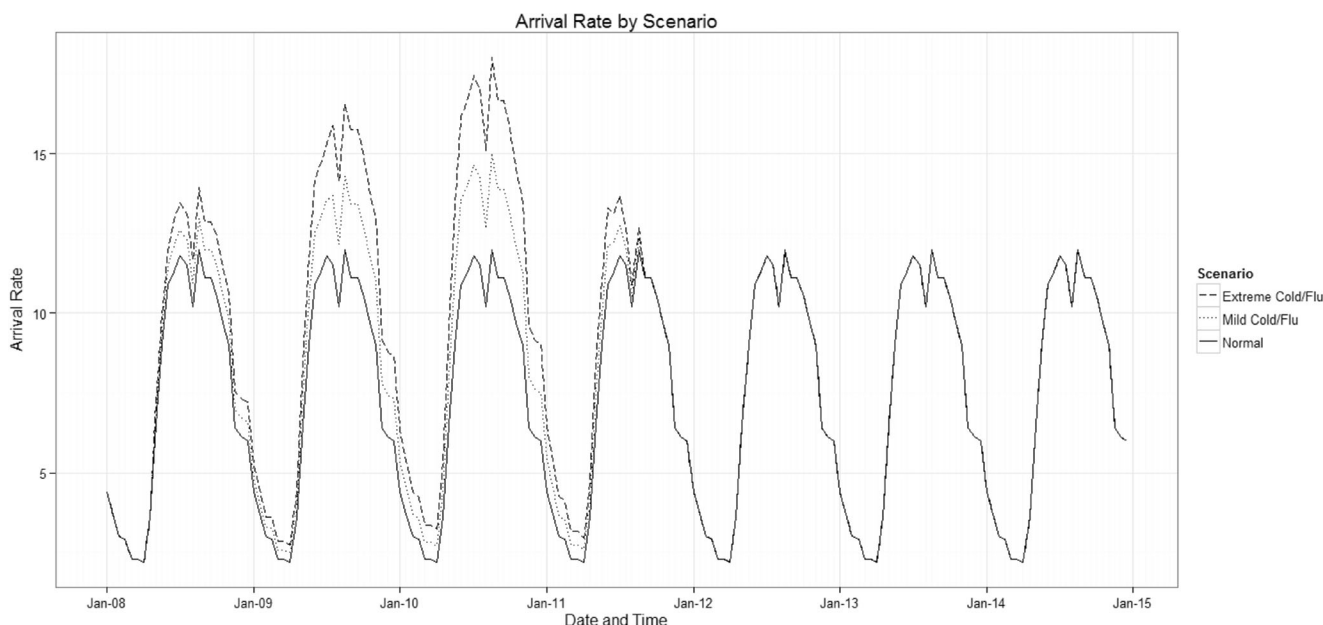| Day	Hour	Arrival rate	Day	Hour	Arrival rate	Day	Hour	Arrival Rate
1	0–8	100 %	2	13–16	119 %	4	4–6	119 %
1	8–11	105 %	2	16–20	121 %	4	6–7	117 %
1	11–15	107 %	2	20–23	122 %	4	7–9	115 %
1	15–18	108 %	3	0–1	122 %	4	9–11	111 %
1	18–22	109 %	3	1–9	123 %	4	11–13	108 %
1	22–23	110 %	3	9–15	124 %	4	13–14	105 %
2	0–3	110 %	3	15–23	125 %	4	14–16	103 %
2	3–8	112 %	4	0–2	123 %	4	16–23	100 %
2	8–13	115 %	4	2–4	121 %	5–7	0–23	100 %

Fig. 3 Hourly arrival rates of outbreak scenarios vs. Normal operating conditions

EDWIN threshold value of 2 for every hour of the day does not seem to be an effective threshold for this hospital. On the other hand, the published threshold values of 100 for NEDOCS and 7 for READI seem more reasonable as the score results under mildly loaded conditions and the associated 90th percentile thresholds fluctuate above and below these published thresholds.

To explore the predictive power of each score more closely, we next calculated the percentage of trials (out of 100) for which each crowding score did not surpass the 90th percentile threshold before the onset of the outbreak, but did surpass the

threshold only after the patient load increased. In other words, we estimated the likelihood that each score provided an alert to clinicians that ED crowding conditions were changing. As a result of this experiment, we found that EDWIN provided alerts for the mild outbreak significantly more often than NEDOCS or READI. With alerts occurring in 73 % of the trials, EDWIN appears to have more predictive power than NEDOCS, which alerted in about 47 % of the trials, and READI pales in comparison to both other scores with only 6 % of trials resulting in alerts. Overall, these results and Fig. 3 suggest that EDWIN and NEDOCS have more predictive power than READI.
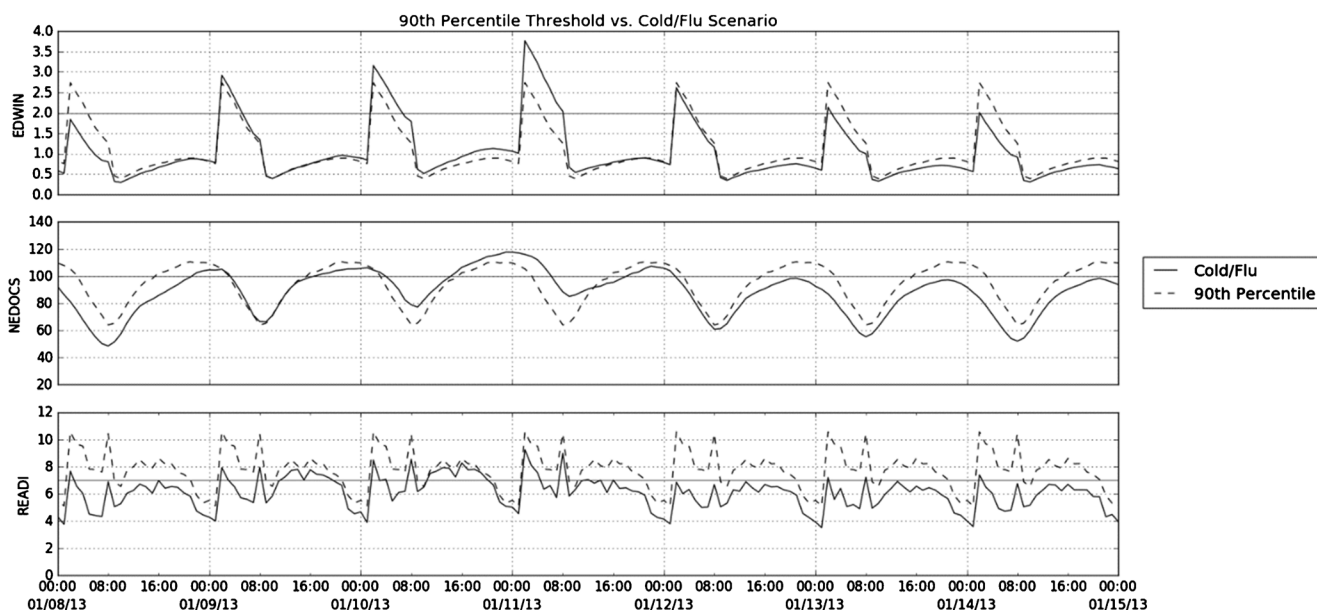

Fig. 4 Crowding score results: mild cold/flu outbreak vs. Threshold values

We also conducted a statistical test to compare how each score would perform with respect to LOS in terms of the predictive power. In particular, for each of the 100 replications, we determined whether each crowding score and the LOS successfully predicted the unusual load. We define a successful prediction as one where the score does not surpass the 90th percentile threshold before the onset of the mild outbreak but surpassed it only after the patient load increases due to the unusual event. We let the outcome of each experiment be one if the crowding score successfully detected the event and zero otherwise for each replication. We repeated the same procedure for the LOS, which resulted in four sequences of 0–1 observations, each sequence having 100 data points. Then, we divided these 100 data points into 10 groups for each sequence and took the group means. Finally, we applied a paired-t test using the resulting ten observations from each sequence to test the hypothesis that each score has the same mean fraction of successful predictions as the LOS. We obtained the following approximate 95 % confidence intervals on the difference between the average fraction of successful predictions for crowding scores and LOS: EDWIN: [0.002, 0.138], NEDOCS: [−0.282, −0.098], and READI: [−0.689,-0.511]. Although none of the intervals include zero, we see that EDWIN and NEDOCS do not perform too differently from the LOS but READI underperforms.

6.2 Extremely loaded scenario

Although the mild outbreak scenario described in Table 4 is a reasonable approximation of what this ED typically sees during flu season, a more severe cold/flu scenario was also explored in the interest of further evaluating score behavior. This

scenario, referred to as "extreme outbreak," uses double the additional arrivals as in Table 4 (e.g., for Day 1 Hour 8–11, the arrival rate is 110 % of normal conditions), with the same hourly arrival breakdown. (See Fig. 3 for a comparison of hourly arrival rates under the normal operating conditions and the two outbreak scenarios.) Figure 5 shows the extreme outbreak results in the same manner as in the mild scenario. The crowding values in Fig. 4 represent the overall average of 100 trials of the extremely loaded model.

In the mild scenario in Fig. 4, we see the simulated crowding score results crossing the 90th percentile threshold by a reasonably small margin, but the results far surpass the threshold values under the extreme scenario in Fig. 5, particularly in EDWIN and NEDOCS. We also see READI briefly crossing the 90th percentile threshold value under the extreme outbreak, unlike the mild scenario, although the READI score is still very jagged and generally uninformative. EDWIN first exceeds the 90th percentile threshold and remains above it starting around 9:00 AM on Day 1, around the true onset of the extreme outbreak. NEDOCS is shortly behind EDWIN, exceeding the threshold around 8:00 PM on Day 1. READI first surpasses the 90th percentile threshold around 10:00 PM on Day 2, but fluctuates above and below the threshold throughout the duration of the extreme scenario. The start of the longest period of time for which READI exceeds the threshold does not occur until 8:00 AM on Day 3.

In Fig. 5, we see that the published threshold value of 2 for EDWIN again does not seem to be useful for this hospital. The extreme outbreak crowding score generally meetsor surpasses 2 during the 0–9 h of each day, which is known to be due to the staffing levels at this ED. Similarly, the published threshold value of 7 for READI also does not seem to be useful for
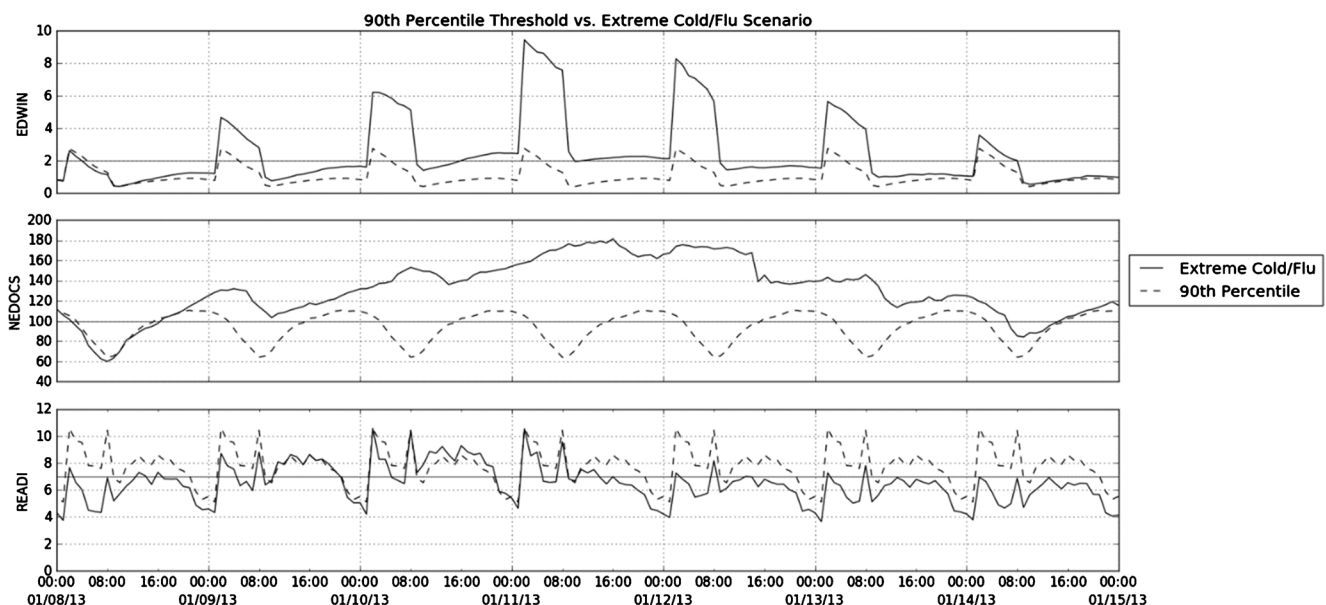

Fig. 5 Crowding score results: extreme cold/flu outbreak vs. Threshold values

this hospital, as the extreme crowding score fluctuates up and down around 7. However, the published threshold value of 100 for NEDOCS seems to be reasonable. Throughout the duration of the extreme outbreak, the NEDOCS score stays consistently above 100, reflecting that the ED is experiencing an extreme patient load.

As in the case for the mild scenario, we calculated the percentage of trials (out of 100) which resulted in an alert (the score surpassing the 90th percentile threshold for the first time after the onset of the extreme outbreak) to clinicians. We found that all three scores performed similarly under the more extreme conditions. In particular, EDWIN correctly alerted in 73 % of the trials, NEDOCS detected the extreme conditions in about 47 % of the cases, and finally, READI alerted correctly only in 6 % of the trials. We also repeated the statistical test to compare and contrast the performance of each crowding score with LOS described in the last paragraph of Section 6.1 and found almost the same 95 % confidence intervals. We have observed that although the extreme outbreak case yielded identical results with the mild case in terms of the number of alerts, the alerts seems to occur earlier in most replications.

7 Conclusions and discussion

With the goal of comparison of three alternative ED crowding scores in terms of their predictive power, we have utilized patient data to build a discrete-event simulation model of a North Carolina academic hospital's ED. We first validated our model by comparing its outcomes (average length of stay and the three crowding scores) to the actual data. We later conducted several experiments using this simulation model to compare the prediction and detection capabilities of the EDWIN, NEDOCS, and READI crowding scores under a hypothetical scenario with raised arrival rates over a period of four days due to an outbreak.

Our simulation model and resulting analysis led to the following conclusions:

1. EDWIN and NEDOCS appear to be helpful measures of current ED crowdedness. NEDOCS best depicts the crowdedness at this ED when we compare the crowding scores to the average length of stay in the ED. In particular, under normal operating conditions, NEDOCS peaks slightly before midnight and decreases until 9 a.m. the following morning, which provides an accurate picture of typical crowding at this particular ED as perceived by clinicians.
2. EDWIN and NEDOCS demonstrated predictive power for anticipating impending crowdedness as a result of a hypothetical disease outbreak. EDWIN captures the simulated outbreak (both at the mild and extreme levels) the earliest on average. Furthermore, EDWIN demonstrated the most predictive power by providing an alert to

changing ED conditions for around 73 % of all the simulated outbreak trials compared to NEDOCS alert rate of around 47 % across all trials.
3. READI does not appear to be a good fit for this ED. The overall daily pattern of the READI score in all scenarios considered (normal operating conditions and extreme cases) generally does not seem to show the true ebb and flow of crowdedness at this ED. Furthermore, the READI score results in a curve that is too jagged and abrupt. Additionally, the READI score provided alerts in only 6 % of the simulated outbreak trials, which suggests that READI does not have good predictive power.

Another major outcome of our simulation study was that the recommended threshold values in the literature for the three crowding scores did not appear to be ideal for this ED. The EDWIN and NEDOCS scores approach the threshold values of 2 and 100, respectively, nearly every night under normal operating conditions. Even in the mild outbreak scenario, EDWIN and NEDOCS fluctuate above and below their threshold values throughout the duration of the increased patient flow, and READI remains below its threshold value. For this particular hospital, we believe that a more realistic interpretation of these scores would come from comparing them against an upper percentile (such as the 90th percentile) of each score (as a function of time) based on historical data. Our simulation results for the outbreak scenario demonstrated that such a percentile-based threshold would be an effective predictive tool for detecting impending crowding.

To summarize, our recommendation for this hospital is to use EDWIN and NEDOCS for assisting health care professionals at detecting unusual crowding situations. In particular, tracking one or both of these scores throughout the day in conjunction with a historical-data-based threshold alert system (such as the 90th percentile threshold proposed in this paper), would alert the ED management to an unusual increase in the crowdedness, which could lead to better preparation and eventually better patient care outcomes.

Limitations of this study The discrete-event simulation model and resulting analysis were based on patient data at one particular hospital. Therefore, the results presented here may not extend to hospitals of different sizes or with different characteristics. Additionally, all parameter estimations and distributional fits are based on one month of patient data for a winter month. Although the historical evidence shows that this ED does not experience drastic seasonal changes in terms of patient loads and service times, additional work may be needed to extend our results to other seasons. While we acknowledge that these assumptions created an imperfect simulation model, the reactions of the crowding scores to these levels of ED crowdedness yielded meaningful conclusions about their strengths and weaknesses.

References

1. Asplin B, Magid D, Rhodes K, Solberg L, Lurie N, Carmago CA (2003) Conceptual model of emergency department crowding. Annals of Emergency Medicine 42(2):173–180

2. Bernstein S, Verghese V, Leung W, Lunney A, Perex I (2003) Development and validation of a new index to measure emergency department crowding. Academic Emergency Medicine 10:938–942

3. Connely L, Bair A (2004) Discrete event simulation of emergency department activity: a platform for systems-level operations research. Academic Emergency Medicine 11:1177–1185

4. Derlet R, Richards J, Kravitz R (2001) Frequent overcrowding in U.S. emergency departments. Academic Emergency Medicine 8: 151–155

5. Epstein S, Tian L (2008) Development of an emergency department work score to predict ambulance diversion. Academic Emergency Medicine 13:421–426

6. Hoot N, Aronsky D (2006) An Early Warning System for Overcrowding in the Emergency Department. AMIA 2006 Symposium Proceedings 3:39–43

7. Hoot N, LeBlanc L, Jones I et al (2008) Forecasting emergency department crowding: a discrete event simulation. Annals of Emergency Medicine 52(2):116–125

8. Hwang U, McCarthy M, Aronsky D et al (2011) Measures of crowding in the emergency department: a systematic review. Academic Emergency Medicine 18:527–538

9. Jones S, Allen T, Flottemesch T, Welch S (2006) An independent evaluation of four quantitative emergency department crowding scales. Academic Emergency Medicine 13:1204–1211

10. Pines J, Iyer S, Disbot M, Hollander J, Schofer F, Datner E (2008) The effect of emergency department crowding on patient satisfaction for admitted patients. Academic Emergency Medicine 15:825–831

11. Reeder T, Burleson D, Garrison H (2003) The overcrowded emergency department: a comparison of staff perceptions. Academic Emergency Medicine 10:1059–1064

12. Reeder T, Garrison H (2008) When the safety net is unsafe: real-time assessment of the overcrowded emergency department. Academic Emergency Medicine 8:1070–1074

13. Sun B, Hsia R, Weiss R et al (2013) Effect of emergency department crowding on outcomes of admitted patients. Annals of Emergency Medicine 61:605–611

14. Tanabe P, Gimbel R, Yarnold P, Kyriacou D, Adams J (2004) Reliability and validity of scores on the emergency severity index version 3. Academic Emergency Medicine 11:59–65

15. Weiss S, Derlet R, Arndahl J et al (2004) Estimating the degree of emergency department overcrowding in academic medical centers: results of the national ED overcrowding study (NEDOCS. Academic Emergency Medicine 11:38–50

16. Weiss S, Ernst A, Nick T (2006) Comparison of the National Emergency Department Over- crowding scale and the emergency department work index for quantifying emergency department crowding. Society for. Academic Emergency Medicine 13:513–518

17. Weng S, Wang L (2011) Simulation Optimization for Emergency Department Resources Allocation. Proceedings of the 2011 Winter Simulation Conference (IEEE):1231–1238

18. Wiler J, Griffey R, Olsen T (2011) Review of modeling approaches for emergency department patient flow and crowding research. Academic Emergency Medicine 18:1371–1379