ORIGINAL ARTICLE

Development and validation of a machine learning algorithm and hybrid system to predict the need for life-saving interventions in trauma patients

Nehemiah T. Liu · John B. Holcomb · Charles E. Wade ·
Andriy I. Batchinsky · Leopoldo C. Cancio ·
Mark I. Darrah · José Salinas

Abstract Accurate and effective diagnosis of actual injury severity can be problematic in trauma patients. Inherent physiologic compensatory mechanisms may prevent accurate diagnosis and mask true severity in many circumstances. The objective of this project was the development and validation of a multiparameter machine learning algorithm and system capable of predicting the need for life-saving interventions (LSIs) in trauma patients. Statistics based on means, slopes, and maxima of various vital sign measurements corresponding to 79 trauma patient records generated over 110,000 feature sets, which were used to develop, train, and implement the system. Comparisons among several machine learning models proved that a multilayer perceptron would best implement the algorithm in a hybrid system consisting of a machine learning component and basic detection rules. Additionally, 295,994 feature sets from 82 h of trauma patient data showed that the system can obtain 89.8 % accuracy within 5 min of recorded LSIs. Use of machine learning technologies combined with basic detection rules provides a potential approach for accurately assessing the need for LSIs in trauma patients. The performance of this system demonstrates that machine learning technology can be implemented in a real-time fashion and potentially used in a critical care environment.

Keywords Machine learning · Artificial intelligence · Clinical decision support systems · Life-saving interventions · Trauma

N. T. Liu (✉) · A. I. Batchinsky · L. C. Cancio · J. Salinas
US Army Institute of Surgical Research, 3650 Chambers Pass,
Building 3610, Fort Sam Houston, TX 78234-6315, USA
e-mail: nehemiah.liu@us.army.mil

A. I. Batchinsky
e-mail: andriy.batchinsky1@us.army.mil

L. C. Cancio
e-mail: lee.cancio@us.army.mil

J. Salinas
e-mail: jose.salinas4@us.army.mil

J. B. Holcomb · C. E. Wade
Department of Surgery, Center for Translational Injury Research,
University of Texas Health Science Center at Houston,
6431 Fannin Street, Houston, TX 77030-1501, USA
e-mail: john.holcomb@uth.tmc.edu

C. E. Wade
e-mail: charles.e.wade@uth.tmc.edu

M. I. Darrah
Athena GTX, Inc., 3620 SW 61st Street, Suite 395,
Des Moines, IA 50321-2419, USA
e-mail: mdarrah@athenagtx.com

1 Introduction

Hemorrhage continues to be the leading cause of preventable mortality in trauma patients. For many of these patients, the key to effective assessment begins with the accurate identification and diagnosis of injury type and severity in order to determine the appropriate and timely treatment options. In this environment, potential delays in performing a life-saving intervention (LSI) may compromise a patient's stability, complicate injuries, and lead to increased morbidity or mortality. Although vital signs—such as a patient's body temperature, heart rate, blood pressure, and respiratory rate—play significant roles in monitoring trauma patients and are generally used as means to

Report Documentation Page

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
01 FEB 2014	**N/A**	**-**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Development and validation of a machine learning algorithm and hybrid system to predict the need for life-saving interventions in trauma patients	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)	5d. PROJECT NUMBER
Liu N. T., Holcomb J. B, Wade C. E, Batchinsky A. I, Cancio L. C, Darrah M. I, Salinas J.,	5e. TASK NUMBER
	5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
United States Army Institute of Surgical Research, JBSA Fort Sam Houston, Tx 78234	

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release, distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE	**UU**	**11**	
unclassified	**unclassified**	**unclassified**			

assess patient condition, they alone may not identify patient destabilization until late and often irreversible changes in state take place. The utility of available field vital signs depends heavily on concomitant interpretation by an expert provider, and vital signs are only measured intermittently during patient care at both prehospital and hospital phases. Previous studies have shown that standard vital signs available from monitors are no better in determining true injury status and severity in trauma patients than a simple physical examination [9]. Nonetheless, measurement and interpretation of electronic vital signs have become routine during prehospital and hospital care. In order to achieve more accurate diagnostic capabilities, new approaches based on the combinations of multiple vital signs, trends, and other information may be better suited for trauma diagnosis [5]. However, as approaches required for improving the sensitivity of diagnosis become more complex, use by providers may become increasingly difficult without the use of adjuncts capable of deriving solutions based on complicated mathematical formulas.

Because of the intricate relationships that exist between vital signs, time, and other factors, developing new approaches that take these items into account will also require the use of advanced information systems and computer algorithms coupled with technologies capable of processing and fusing multiple parameters, weight, and trends. These algorithms provide the capability to extract the maximum information content available from both single vital signs and combinations of multiple vital signs across time points. Machine learning (ML) technology is one approach that has been studied recently as a potential approach and solution for the multivariate processing of vital signs required to accurately diagnose the patient condition in the critical care environment [2, 10, 11, 14]. In addition, because such an environment requires rapid, accurate decisions within a short time frame, ML technology can potentially help describe possible interventions earlier. This approach uses information technology to mimic human decision making and provide an automated approach for processing vital signs and other patient data, with the aim of predicting needs rapidly for patient care. The use of ML technology may facilitate the process of triaging patients to appropriate trauma centers where experienced personnel can rapidly perform LSIs. By providing diagnostic support, computers and ML may fit well into a prehospital triage algorithm that focuses on LSIs as its optimal end point [3, 7]. Additionally, such technology may strengthen the ability to accurately triage trauma patients in the prehospital environment and hence improve the survival of patients that would have otherwise died.

The purpose of this study was to test and evaluate different types of artificial intelligence and ML methods for modeling injury severity (as defined by the need for an LSI

during prehospital and/or in the emergency department) of a set of retrospectively and prospectively collected trauma patients based on data collected over 5 years, most recently from the Wireless Vital Signs Monitor (WVSM, Athena GTX, Des Moines, Iowa) trial. Methods were implemented as a real software application module capable of real-time processing in a computer system with moderate performance specifications. A 90/10 cross-validation approach was used for designing the predictive capacity of the algorithm.

We hypothesized that because computers and ML can process large amounts of disparate data continuously, quickly, and accurately, they would not only benefit trauma diagnosis, especially in the context of prehospital triage, but also integrate well into an electronic system that may perform LSI predictions in real time.

A novelty of this study was the development and validation of an ML algorithm and hybrid system to predict the need for LSIs in trauma patients. While there have been numerous studies utilizing decision trees, conjunction rules, support vector machines, artificial neural networks, multilayer perceptrons, and logistic regression models in order to discriminate between different patient groups, to date, no study has investigated the possibilities of predicting in real time the needs for LSIs in trauma patients. The ultimate goal of this work was to address this shortcoming and provide physicians with a new tool for decision support.

2 Methods

This study was approved by the Institutional Review Boards of the US Army Institute of Surgical Research, Fort Sam Houston, TX, USA, and the University of Texas Health Science Center at Houston, Houston, TX, USA. We analyzed data from both the Trauma Vitals (TV) database and the WVSM to generate datasets for training and validating an artificial intelligence model, respectively.

2.1 TV database and protocol

Data in the TV database include severe trauma patients with blunt and penetrating injuries transported from the scene by helicopter service to a Level I trauma center in Houston, Texas, or San Antonio, Texas. Patients were monitored from the scene during transport using a Welch Allyn Propaq 206 (Welch Allyn, Skaneateles Falls, NY) monitor or Welch Allyn PIC 50 (Welch Allyn, Skaneateles Falls, NY) monitor. Propaq data were collected using a computerized personal digital assistant (PDA) attached to the monitor during transport. Data were stored in a nonvolatile memory card in the PDA for use during the study. All

numeric Propaq data were stored at a rate of 1 Hz. Waveform data were recorded at a rate of 182 Hz. PIC 50 data were stored in a built-in flash memory card attached to the monitor. PIC 50 numeric data were stored at a rate of one measurement every 3 min, coinciding with the patient's noninvasive blood pressure measurements. Waveform data were stored at a rate of 375 Hz. Data from the PDA and flash cards were extracted by research personnel and uploaded to the TV database for analysis. All nonelectronic data were manually recorded on the run sheet from the monitor's screen by Emergency Medical Services medics, then collected on a standardized form, and entered into the TV database. These included demographic data, physical examination results, Glasgow coma scores, and interventions performed on the patients in the field. LSIs consisted of endotracheal intubations, transfusions, tube thoracostomies, cardiopulmonary resuscitations, needle decompressions, angio-embolizations, cricothyrotomies, thoracotomies, and cardioversions.

Data for the ML model were selected based on a generic population sample that would prevent the algorithm from training to only a subset of the general population (which would result in a nongeneralizable model). We selected over 30 h worth of data corresponding to 79 prehospital patient records from the TV database based upon three criteria necessary to maximize the input parameter values required to provide an optimal learning set to the ML engine. These included (1) availability of vital signs and patient status summary scores (Murphy Factor, Athena GTX, Des Moines, Iowa) with values from 0 to 5 for the patient on a second-by-second basis, (2) blood pressure measurements over a minimum of 15 min that also changed from initial measurements (had a baseline shift to yield measurable slopes), and (3) heart rate measurements uncorrupted by electromechanical noise. Lengths of patient records varied from approximately 15–30 min. Records contained single episodes of data, sometimes missing one or more measurements from different vital signs over the episode's duration. Standard vital signs used during trauma care for patient assessment included heart rate (HR), systolic blood pressure (SBP), diastolic blood pressure (DBP), mean arterial pressure (MAP), respiratory rate (RR), and blood oxygenation (SpO_2). Combinations of these vital signs were also used to derive other measurements including shock index ($SI = HR/SBP$) and pulse pressure ($PP = SBP - DBP$). The Murphy factor is a patient status summary alarm that provides the medic with decision support capability by combining all vital signs, trends, and pulse characteristics recorded by the monitor, and applying a multivariate sensor fusion algorithm that generates a combined index of the patient condition in ranges from 0 to 5. Results are interpreted as 0–1 in green, 2–3 in yellow, and 4–5 in red, indicating a patient who is in a normal, low-priority, and high-priority condition, respectively. A baseline shift was defined as a dataset with at least one change from initial measurement over the data-recording time. This provided a learning dataset required for the ML algorithm to learn from changes in blood pressure during transport.

2.2 WVSM database and protocol

Because a limitation of the initial TV cohort was the fact that times of actual LSIs were not recorded and stored in the TV database, an additional 82 h worth of data corresponding to 24 prehospital patient records were chosen based upon availability of times of recorded LSIs from the WVSM protocol. From June 27, 2011, to January 6, 2012, 305 consecutive patients transported from the injury scene via the Life Flight helicopter service to the Memorial Hermann Hospital, a Level I trauma center in Houston, Texas, were enrolled for this protocol. This included data captured from 104 patients wearing the WVSM system during transport to the Houston Level I trauma center.

Data in the WVSM database include severe trauma patients with blunt and penetrating injuries transported from the scene by helicopter service to a Level I trauma center in Houston, Texas. WVSM data were collected using a computerized server system that collected and stored all transport data from the WVSM device through a wireless connection once a patient arrived in the emergency department. Numeric data from the WVSM device were stored at a rate of 1 Hz. In addition, ECG waveform data from a single lead and pleth waveform data from a thumb-mounted pulse oximeter to the WVSM were recorded at the rates of 230 and 75 Hz, respectively. For trauma patients with concomitant lung injuries, respiration waveform data were also recorded at a rate of 10 Hz. Standard vital signs used during trauma care for patient assessment included the same vital signs recorded in the TV database (HR, SBP, DBP, MAP, RR, SpO_2, SI, and PP). All nonelectronic data were manually recorded on an electronic run sheet (Tablet PCR, Zoll Medical, Chelmsford, MA, USA) by Emergency Medical Services medics, then collected on a standardized form, and entered into the WVSM database (OpenClinica). These included demographic data, physical examination results, Glasgow coma scores, and interventions performed on the patients in the field. LSIs consisted of endotracheal intubations, transfusions, tube thoracostomies, cardiopulmonary resuscitations, needle decompressions, angio-embolizations, cricothyrotomies, thoracotomies, and cardioversions. Patients for analysis were selected based upon two criteria: (1) direct transport of the patient from the injury scene to the hospital and (2) an injury requiring hospital admission. Of these 104 patients, 32 received at least one LSI, while only
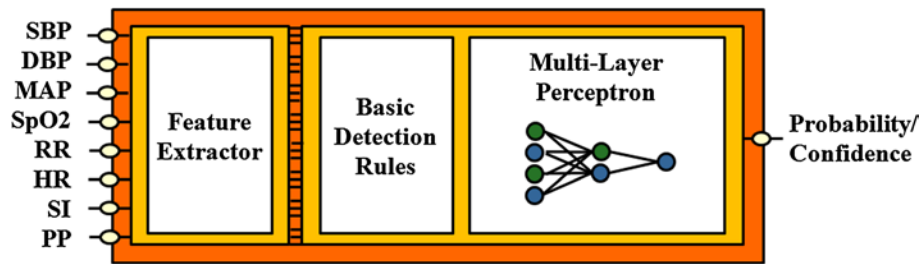
Fig. 1 A hybrid system for predicting the need for life-saving interventions in trauma patients. To predict the need for life-saving interventions, a hybrid system could employ the following components: (1) a component that extracts features from the measurements of various vital signs, (2) a simple rule-based algorithm that handles obtrusive cases (*features*) involving measurements that clearly indicated the need for some interventions, and (3) a machine learning algorithm (*multilayer perceptron*) that handles more obscure and complex cases (*features*) involving measurements unrecognized by the rule-based algorithm. If a patient's vital signs were clearly abnormal according to a set of basic rules, the patient would be classified as needing a life-saving intervention. Standard vital signs used during trauma care for patient assessment often include heart rate (HR), systolic blood pressure (SBP), diastolic blood pressure (DBP), mean arterial pressure (MAP), respiratory rate (RR), and blood oxygenation (SpO2). Combinations of these vital signs are also used to derive other measurements including shock index (SI = HR/SBP) and pulse pressure (PP = SBP − DBP)

24 patients had both recorded LSIs and corresponding LSI predictions. Actual LSIs were recorded only when the nurse/paramedic manually pressed a button on the WVSM data-capture-and-display interface. Only the start of each LSI was recorded. These 24 patients provided a validation set for this project. Lengths of these records varied from approximately 3–4 h. Records contained single episodes of data, sometimes missing one or more measurements from different vital signs over the episode's duration.

2.3 Design, validation, and analysis

Design of a hybrid system for LSI predictions employed two components: (1) a simple rule-based algorithm that would serve as a front end to handle obtrusive cases involving measurements that clearly indicated the need for some interventions and (2) an ML algorithm that would serve as an intelligent component to handle more obscure and complex cases involving measurements unrecognized by the front end (see Fig. 1). This configuration was a novelty proposed from our combined knowledge of ML and medicine. If a patient's vital signs were clearly abnormal according to a set of basic rules, the patient would be classified as needing an LSI (see rules below). Rules were based on the analysis of data from the TV database. "Normal" measurements included those within the 95 % confidence interval of the database. On the other hand, if vital signs were not obviously abnormal, the data would be passed to the ML algorithm. The basic detection rules were meant to filter out patients who required immediate attention. Order of the rules also reflected the relative importance of measurements for discriminating patient instability and their potential for affecting system performance.

In addition, distributions of initial nonzero BP measurements and mean non-BP measurements across all patient records in the TV database were used to formulate detection rules. First, identification of tail regions of BP-related distributions was combined with knowledge of expected BP ranges (normal SBP 90–120 mm Hg; normal DBP 60–80 mm Hg; normal MAP: DBP + (SBP − DBP)/3 mm Hg) to derive lower-bound and upper-bound thresholds. Similarly, identification of tail regions of non-BP-related distributions was combined with knowledge of expected ranges (normal HR 60–100 bpm; normal RR 12–20 breaths per minute; normal SpO$_2$ 94–100 % at sea level; normal SI 0.5–0.7 bpm/mm Hg) to derive thresholds. Decision tables, decision trees, and/or conjunction rules were then employed to tailor rules.

Rates of change (slopes) and mean and maximum measurements were used to train an ML algorithm so that it could respond quickly to measured trends and disparities in a patient's vital signs. We used linear regression to calculate slope values, ignoring those values equal to zero to derive an estimate of the rate of change for numeric values across time. A sliding window of 180 s was used to calculate slopes for all non-BP-related vital signs. Because BP-related measurements were recorded every 3 min, a sliding window of 540 s was used to calculate slopes for these measurements. Data were configured for input into an ML modeler (WEKA, University of Waikato, New Zealand) to generate the ML model [8].

The main criteria to train the ML algorithm using the TV database were a strong correlation between inputs and outputs, preferably, with correlation coefficient greater than 70 %, and a low mean absolute error, with values less than 30 %. From a system perspective, the ML algorithm needed to produce smooth continuous outputs (probabilities) between 0 and 1, indicating the need for an LSI. Because these outputs could not be binomial nor jump sporadically up and down with discontinuities, data corresponding

to patients with LSIs were not separated from data corresponding to patients without LSIs. Visual assessment of the outputs was required to evaluate system performance.

To design our ML algorithm, we used a tenfold cross-validation approach [1, 6, 13]. Because standardizing the inputs improves the numerical condition of the data for training, we preprocessed the data before training the classifier by replacing all unknown or missing features for each given patient record in the dataset with zeros and normalizing all other features so that features fall within the range −1–1. In particular, we used a maximum–minimum normalization rule as follows:

$$\bar{x}_{i,j} = \frac{x_{i,j} - \frac{1}{2}\left(x_{i,j_{\max}} + x_{i,j_{\min}}\right)}{\frac{1}{2}\left(x_{i,j_{\max}} - x_{i,j_{\min}}\right)}$$

$\forall\, i = 1, …, N,\, \forall\, j = 1, …, M$ and $j \subset M_i$, where $x_{i,j}$ denotes the jth feature value of feature set i, N denotes the number of instances (feature sets) in the training data, M denotes the number of features in an instance, and M_i denotes the set of known features in the feature set i. In addition, we replicated each feature set, replacing unknown or missing feature values for each set with averaged values over all values in the training data and then normalizing all feature values using the rule above. This second set was only used for providing a confidence interval for outputs from the first dataset, not for model training.

The following features formed a feature set of our training data for designing the MLP: slope of SBP, current SBP, slope of DBP, current DBP, slope of MAP, current MAP, slope of SpO$_2$, mean SpO$_2$, slope of RR, mean RR, slope of HR, mean HR, slope of inverted SI, mean inverted SI, slope of PP, current PP, maximum SBP, maximum DBP, maximum MAP, maximum SpO$_2$, maximum RR, maximum HR, maximum inverted SI, and maximum PP. Moreover, classifications were obtained by remapping Murphy scores to a scale between 0 and 1, i.e., a nominal probability. The final training data consisted of over 110,000 feature sets. Thus, these training data covered more than 30 h of data, ranging across different physiologic, temporal, and spatial conditions.

Validation involved the WVSM protocol and its patient records and was accomplished by determining the output of our hybrid system at the time of the recorded LSI and the maximum output of our system 60 s, 3 min, and 5 min prior to the recorded LSI. In other words, the observation window ended at the time of the recorded LSI. Further validation was done by sampling outputs during the first 5 min of each patient record and 5 min prior to the start of each LSI. An initial analysis classified prediction outputs (probabilities) >30 % as true positives (TPs) and otherwise as false negatives (FNs). Similarly, a second analysis classified outputs >50 % as TPs and otherwise as FNs, respective of the analysis.

3 Results

3.1 Model development

The demographics of the 79 patients included in this study are depicted in Table 1; likewise, the demographics of the WVSM patients are shown in Table 2. Quartiles were established for age. Race and age were not different between those patients who received at least one LSI and those who received none, nor did male gender predispose to an LSI. Likewise, increasing patient age did not increase the frequency of an LSI in this sample/study. Of the 79 patients, 24 (30 %) did not require an LSI. The other 55 patients received a total of 124 LSIs. Thirty-nine percent (48) of the LSIs were performed prehospital, 60 % (74) in the emergency room, and 1 % (2) elsewhere. Interventions consisted of the following: 42 endotracheal intubations, 42 transfusions, 18 tube thoracostomies, eight cardiopulmonary resuscitations, five needle decompressions, five angio-embolizations, two cricothyrotomies, two thoracotomies, and one cardioversion. Table 3 describes the hybrid system's front-end component, that is, the basic detection rules that were used to identify patients who required immediate interventions. As a note, the value of 0.9 (90 %) in the table was arbitrary and only used to indicate the fact that abnormal measurements should alert the provider's attention to a strong need for an LSI.

We trained and compared several ML models, including decision trees, conjunction rules, support vector machines, artificial neural networks, multilayer perceptrons, and logistic regression models. Models were generated for the 110,000+ feature sets using WEKA and binary and continuous classes. In order to develop a real-time hybrid system to predict the need for LSIs, i.e., output a continuous probability, we limited comparisons to artificial neural networks, multilayer perceptrons, and logistic regression models and sought models that yielded highest correlation and lowest errors. Comparisons of top cross-validation results are shown in Table 4.

Comparisons among several proposed models proved that a multilayer perceptron (MLP) would best implement the ML algorithm in the novel hybrid LSI prediction system. This ML model consisted of 24 inputs, 12 hidden nodes that each contained a set of 24 optimized weights, and one output that contained a set of 12 optimized weights. The back-propagation algorithm (learning rate 0.05, momentum 0.2) was used to train the MLP (as well as all other algorithms in Table 4). The activation function employed by the MLP was the sigmoid function. Given 111,028 feature sets, the WEKA tool spent exactly 10.3 h (36,861.2 s) to generate weights for the MLP nodes. In Table 4, the high correlation coefficient of 0.8072 indicates that the predicted probabilities of the MLP matched the desired probabilities reasonably well. In addition, the

Table 1 Demographics of selected patients from the Trauma Vitals database

Variable	All patients		Patients with life-saving interventions (LSIs)		Number of LSIs
	Number N	Percentage N/79	Number n	Percentage n/N	
All patients	79	100	55	72	124
GCS total* Mean 10 ± 5					
3	22	28	22	100	53
4	3	4	3	100	8
5	1	1	1	100	1
6	3	4	3	100	8
7	2	3	2	100	4
8	1	1	1	100	1
10	1	1	0	0	0
11	1	1	1	100	2
12	2	3	2	100	4
13	8	10	4	50	7
14	13	16	3	23	10
15	22	28	13	59	26
Gender					
Female	29	37	19	66	44
Male	50	63	36	72	80
Race					
White/Caucasian	41	52	29	71	69
Black	7	9	6	86	12
Hispanic	19	24	12	63	24
Asian/Pacific	1	1	1	100	3
Not recorded	11	14	6	55	16
Age Mean 37 ± 15 Quartiles					
18–23	21	27	16	76	36
25–35	19	24	15	79	34
36–47	20	25	11	55	26
49–86	19	24	13	79	28
Systolic BP* Mean 79 ± 24					
0–66	20	25	18	90	47
67–77	20	25	14	70	35
80–94	20	25	16	80	36
96–128	19	25	7	37	6
Diastolic BP* Mean 43 ± 21					
0–32	20	25	17	85	43
33–45	21	27	16	76	41
46–58	18	23	14	78	30
59–83	20	25	8	40	10
Heart rate* Mean 112 ± 26					
74–100	13	16	6	46	12
108–128	14	18	12	86	31

	Variable	All patients		Patients with life-saving interventions (LSIs)		Number of LSIs
Table 1 Continued		Number N	Percentage N/79	Number n	Percentage n/N	
	135–185	5	6	5	100	13
	Unknown	47	60	32	68	68
	Respiratory rate* Mean 20 ± 7					
LSI life-saving intervention, *GCS* Glasgow Coma Scale/ Score, *BP* blood pressure (mm Hg), heart rate (beats per minute), respiratory rate (breaths per minute), *SpO₂* saturation of peripheral oxygenation (%)	6–10	8	10	8	100	18
	12–20	26	33	13	50	27
	21–38	25	32	14	48	29
	Unknown	20	25	20	100	50
	SpO₂*					
	97	1	1	1	100	3
	100	1	1	1	100	3
* Entry values taken from the run sheet	Unknown	77	98	53	69	118

MLP model achieved a relatively low mean absolute error of 0.1612, which equals the sum total of the absolute differences between each desired probability and its predicted probability divided by the total number of instances during cross-validation.

Our ML algorithm system was able to generate outputs commensurate with baseline changes in the patients' vital signs in real time. Through a graphical interface, we analyzed these results, and for selected records, we plotted predicted probabilities against patient features in order to assess the influence of every feature on the prediction model. We illustrate our analyses by showing an example in Fig. 2.

When all vital signs were available, RR and SpO₂ were most discriminative in detecting patient instability and affecting system performance. These results agreed with the fact that measurements outside of the 95 % confidence interval of measurement distributions (from the TV database) would immediately trigger the basic detection rules of the hybrid system. When RR and SpO2 measurements were missing from the input set, BP-related vital signs (SBP, DBP, MAP, and PP) were most discriminative in detecting patient instability (see Fig. 2).

3.2 Validation

To validate the model, we employed an additional set of data derived from 305 patients of which 37.7 % required an LSI (Table 3). Of the 199 LSIs, 90 (45 %) were performed prehospital and 109 (55 %) in the emergency department.

There were 295,994 feature sets from 82 h of real-world patient data to validate the hybrid classification system. Table 5 shows confusion matrices for the initial analysis, as described in the "Design, validation, and analysis" section. Importantly, the system was able to obtain a sensitivity of 89.8 % within 5 min of recorded LSIs when a probability >30 % was denoted as a TP. Moreover, the system achieved a positive predictive accuracy of 96.4 % for observation windows described in the previous section.

4 Discussion

Although the application of ML algorithms to datasets began over 50 years ago and now has roots in multiple disciplines [4, 12], only recently has this technology been introduced to trauma research. Furthermore, ML technology has rarely been applied to trauma diagnosis, decision support, or clinical practice for the trauma patient. This study was designed to advance trauma patient care through the development and validation of an ML algorithm and hybrid system to predict the need for LSIs in trauma patients. In previous work, only ML and new vital signs were explored for their utility to discriminate between LSI and non-LSI patients [2]. Neither standard vital signs nor trends were used for identifying LSI patients. Likewise, numerous studies utilizing various ML techniques in order to discriminate between different patient groups have been conducted. However, to date, no study has investigated the possibilities of predicting in real time the needs for LSIs in trauma patients using ML and other information.

By producing over 110,000 feature sets from various vital sign measurements of a select cohort of trauma patients, we intended to capture the synergistic complexities among vital signs, derived statistics, time, and spatial/environmental factors—complexities that may not be understood by the health practitioner in an emergency situation. Since feature sets retrospectively scored patients on a per-second basis, time became an integral part of real-time system design. Furthermore, unlike previous work [2, 10, 11, 14], system design involved not only the development

Table 2 Demographics of selected patients from the Wireless Vital Signs Monitor protocol

Variable	All patients		Patients with LSIs		Prehospital LSIs		Hospital LSIs	
	# N	% N/305	# n	% n/N	# i	% i/90	# j	% j/109
All patients	305	100	115	38	90	100	109	100
Gender								
Female	104	34	32	31	31	34	44	40
Male	201	66	83	41	59	66	65	60
Race								
White/Caucasian	191	63	71	37	58	64	62	57
Black	30	10	12	40	10	11	13	12
Hispanic	64	21	28	44	19	21	27	25
Asian/Pacific	3	1	2	67	1	1	5	5
Not recorded	17	5	2	12	2	3	2	1
Age Mean 39 ± 16 Quartiles								
18–26	76	25	30	40	30	33	37	34
27–36	76	25	28	37	22	25	19	17
37–50	76	25	26	34	17	19	27	25
51–85	77	25	31	40	21	23	26	24
Heart rate* Mean 95 ± 19 Quartiles								
53–80	76	25	21	28	16	18	18	17
81–92	68	22	22	32	13	14	13	11
93–105	72	24	20	28	17	19	19	17
106–170	74	24	45	61	40	44	54	50
Unknown	15	5	7	47	4	5	5	5
Systolic BP* Mean 133 ± 26 Quartiles								
61–118	74	24	45	61	31	34	50	46
120–133	73	24	24	33	19	21	11	10
134–105	72	24	16	22	13	14	17	16
106–170	74	24	25	34	22	25	27	25
Unknown	12	4	5	42	5	6	4	3

LSI life-saving intervention, *BP* blood pressure (mm Hg), heart rate (beats per minute)

* Entry values taken from the run sheet

of an ML model but also formulation of basic detection rules.

In addition, we chose to develop a real-time ML algorithm system that incorporates an MLP based upon the ability to handle complex datasets and perform well on nonlinear data, especially missing data. Moreover, MLPs yield numerical outputs equivalent to probabilities, provide easy real-time implementation in software, and learn through conventional techniques (such as the back-propagation algorithm). An MLP has a major strength over a traditional artificial neural network in that it uses a hidden layer or layers of nodes and transforms every weighted sum using a nonlinear function before making any threshold comparisons. Hence, MLPs not only distinguish, if possible, the

instances of classes in some feature space, but also join isolated convex regions into a single class [13].

For this study, the term "prediction" denoted the probability that a patient needs an LSI at a particular time. While ML may help predict whether a patient should receive an LSI, the accuracy of that prediction and its confidence interval depend upon the availability of measurements and their buffered histories. In other words, we expected that the longer our system could buffer measurements and calculate features, the more reliable our system would perform. As such, initial predictions would only make sense with respect to their place in time, and confidence intervals would only improve as time goes on. As a part of the design process, we explored the types and numbers of features that

would best assist ML. To add robustness to system design and validation, datasets included feature sets that contained missing vital sign measurements.

Interpretation of outputs during this study influenced use and performance of the hybrid prediction system; likewise, selection of outputs in a given time frame. When a probability >30 % was denoted as a TP, the system was able to obtain an accuracy of 89.8 % within 5 min of recorded

Table 3 Basic detection rules

if ((current SBP < 90 mm Hg) or (current SBP > 200 mm Hg)) then   output = .9;
else if ((current DBP < 40 mm Hg) or (current DBP > 140 mm Hg)) then   output = .9;
else if ((current PP < 20 mm Hg) or (current PP > 100 mm Hg)) then   output = .9;
else if ((current MAP < 60 mm Hg) or (current MAP > 180 mm Hg)) then   output = .9;
else if (mean SpO₂ < 85) then   output = .9;
else if (mean HR > 130 bpm) then   output = .9;
else if ((mean SI < .2) or (mean SI > 1.6)) then   output = .9;
else if (max SBP > 120 mm Hg) and (max DBP > 80 mm Hg) and (max MAP > 100 mm Hg) and (mean HR > 115 bpm) and (max SpO₂ ≤ 95)) then   output = .9;
else if ((current MAP > 131 mm Hg) and (max RR > 40 breaths per minute)) then   output = .9;
else if ((max SBP > 160 mm Hg) and (max DBP > 120 mm Hg) and (mean RR > 40 breaths per minute)) then   output = .9;
else   use machine learning algorithm

SBP systolic blood pressure (mm Hg), *DBP* diastolic blood pressure (mm Hg), *PP* pulse pressure (mm Hg), *MAP* mean arterial pressure (mm Hg), *SpO₂* saturation of peripheral oxygenation (%), *HR* heart rate (beats per minute), *SI* shock index (beats per minute/mm Hg), *RR* respiratory rate (breaths per minute)

LSIs. As this selection time frame was narrowed, the hybrid system achieved smaller accuracies. On the other hand, as the selection time frame increased, the simple rule-based algorithm played a greater role in indicating patient destabilization. When a probability >50 % was denoted as a TP, the system was able to obtain an accuracy of 69.5 % within 5 min of recorded LSIs.

4.1 Limitations

This study had a number of limitations. The sizes of the training and validation datasets were small, i.e., they contained less than 120 h of data from less than 110 patients in total. Moreover, the results were preliminary due to the dataset sizes, criteria for selecting the data, the training dataset used to design our ML algorithm, and the fact that nonpresence of an LSI does not equate to a need for an LSI. Therefore, we tended to err on the side that certain measurements may indicate the possible need of an intervention rather than indicate an LSI is not required. This is the basic concept of overtriage that is a central tenet of trauma care. In order to trade off the requirement of low outputs (<10 %) that indicated stable measurements with the requirement of high outputs (>90 %) that indicated patient needs, we accepted the middle ground that a system would gravitate extremes toward the center, and we chose, instead, to add an offset to ML outputs according to a power-law adjustment (see Table 6). In other words, outputs closer to 100 % were subtracted with a smaller power of 2, whereas outputs closer to 0 % were subtracted with a larger power of 2. This adjustment compensated for the bias in the training dataset so that the real-time ML system could yield a wide range of values, including very small (e.g., 0 %) and very large (e.g., 100 %) predictions.

Lastly, this study did not investigate the impact of noise and artifacts in the measurements on the real-time performance of our system. Although the training dataset contained missing data and erroneous measurements and system design employed safeguards against abnormal measurements, future studies using larger datasets and noisy measurements will be required to test system performance thoroughly and improve system robustness.

Table 4 Comparisons of cross-validation results for various machine learning models

| Detail	Multilayer perceptron	Single logit	Multilayer perceptron	Single logit
Correlation coefficient	0.8072	0.5669	0.7790	0.5660
Mean absolute error	0.1612	0.3004	0.1395	0.3061
Root mean squared error	0.2251	0.3867	0.1784	0.3897
Relative absolute error, %	44.4068	60.3197	58.7536	61.4533
Root relative sq. error, %	59.2866	77.4999	62.7230	78.0853
Number of features	24	24	16	16
Total number of instances	111,028	111,028	111,028	111,028
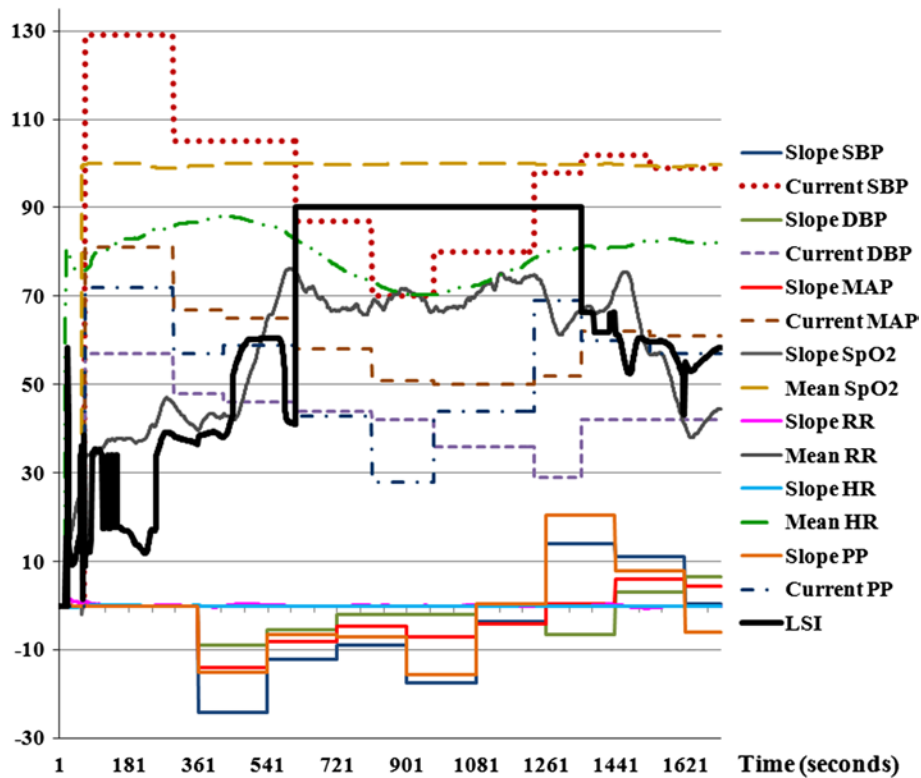
Fig. 2 Plots of prenormalized features and predictions versus time for a trauma patient record. Standard vital signs used during trauma care for patient assessment included heart rate (HR), systolic blood pressure (SBP), diastolic blood pressure (DBP), mean arterial pressure (MAP), respiratory rate (RR), and blood oxygenation (SpO2). Combinations of these vital signs were also used to derive other measurements including shock index (SI = HR/SBP) and pulse pressure (PP = SBP − DBP). The following features were extracted for a hybrid system in order to predict the need for life-saving intervention (LSI as a probability): slope of SBP, current SBP, slope of DBP, current DBP, slope of MAP, current MAP, slope of SpO2, mean SpO2, slope of RR, mean RR, slope of HR, mean HR, slope of inverted SI, mean inverted SI, slope of PP, current PP, maximum SBP, maximum DBP, maximum MAP, maximum SpO2, maximum RR, maximum HR, maximum inverted SI, and maximum PP. Linear regression was used to calculate slope values, ignoring those values equal to zero to derive an estimate of the rate of change for numeric values across time. A sliding window of 180 s was used to calculate slopes for all non-BP-related vital signs. Because BP-related measurements were recorded every 3 min, a sliding window of 540 s was used to calculate slopes for these measurements. For this particular patient, the hybrid system described in this paper yielded appropriate outputs corresponding to the input feature set. The region where the solid black line remained at 90 was a result of the rule-based algorithm of the hybrid system detecting BP-related measurements outside of "normal" range values

Table 5 Confusion matrices for the performance of the hybrid system

Recorded LSI	True positive (TP ≥30 %)							
	LSI algorithm (at recorded time)		LSI algorithm (within 60 s)		LSI algorithm (within 3 min)		LSI algorithm (within 5 min)	
	Y	N	Y	N	Y	N	Y	N
Y	41	18	45	14	52	7	53	6
N	10	108	5	113	3	115	2	116
	24 TPs had max ≥90 %		34 TPs had max ≥90 %		36 TPs had max ≥90 %		37 TPs had max ≥90 %	

LSI life-saving intervention, *TP* true positive, *max* maximum value

In summary, we developed and validated an algorithm and system to predict the probability of a trauma patient requiring an LSI. The system is composed of an MLP and rules for predicting the need for LSIs in both prehospital and emergency department trauma patients. The performance of our system demonstrates that ML technology combined with basic detection rules may provide valuable support in assessing trauma patients within the

Table 6 Power-law adjustment for system outputs as probabilities

if ((.80 ≤ output) and (output < .90))
then　　output = output – .02;
else if ((.70 ≤ output) and (output < .80))
then　　output = output –.04;
else if ((.60 ≤ output) and (output < .70))
then　　output = output –.08;
else if ((.50 ≤ output) and (output < .60))
then　　output = output –.16;
else if (output < .50) {
then　　output = output –.32;

critical care environment. Future studies will expand on the described approach utilizing assigned prediction probabilities derived from this initial effort and include system validation in a clinical trial with both recorded LSIs and times of performance.

References

1. Alpaydin E (2004) Introduction to machine learning, 1st edn. MIT Press, Cambridge, pp 1–16
2. Batchinsky AI, Salinas J, Jones JA, Necsoiu C, Cancio LC (2009) Predicting the need to perform life-saving interventions in trauma patients using new vital signs and artificial neural networks. Lect Notes Comput Sc 5651:390–394
3. Baxt WG, Jones G, Fortlage D (1990) The trauma triage rule: a new, resource-based approach to the prehospital identification of major trauma victims. Ann Emerg Med 19:1401–1406
4. Buchanan BG (2005) A (very) brief history of artificial intelligence. AI Magazine 26:53–60
5. Chen L, Reisner AT, Gribok A, Reifman J (2009) Exploration of prehospital vital sign trends for the prediction of trauma outcomes. Prehosp Emerg Care 13:286–294
6. Duda RO, Hart PE, Stork DG (2001) Pattern classification, 2nd edn. John Wiley and Sons, New York, pp 215–349
7. Garner A, Lee A, Harrison K, Schultz CH (2001) Comparative analysis of multiple-casualty incident triage algorithms. Ann Emerg Med 38:541–548
8. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH (2009) The WEKA data mining software: an update. SIGKDD Explor 11:10–18
9. Holcomb JB, Salinas J, McManus JM, Miller CC, Cooke WH, Convertino VA (2005) Manual vital signs reliably predict need for life-saving interventions in trauma patients. J Trauma 59:821–829
10. Hravnak M, Devita MA, Clontz A, Edwards L, Valenta C, Pinsky MR (2008) Cardiorespiratory instability before and after implementing an integrated monitoring system. Crit Care Med 177:A842
11. Hravnak M, Edwards L, Clontz A, Valenta C, DeVita MA, Pinsky MR (2008) Defining the incidence of cardio-respiratory instability in step-down unit patients using an electronic integrated monitoring system. Arch Intern Med 168:1300–1308
12. Patela VL, Shortliffea EH, Stefanellic M, Szolovitsd P, Bertholde MR, Bellazzic R, Abu-Hanna A (2009) The coming of age of artificial intelligence in medicine. Artif Intell Med 46:5–17
13. Rohwer R, Wynne-Jones M, Wysotzki F (1994) Neural Networks. In: Michie D, Spiegelhalter DJ, Taylor CC (eds) Machine learning, neural and statistical classification. Series in artificial intelligence. Ellis Horwood Publishing, New York, pp 84–105
14. Tarassenko L, Hann A, Young D (2006) Integrated monitoring and analysis for early warning of patient deterioration. Br J Anaesth 97:64–68