Automated Decision-Support Technologies for Prehospital Care of Trauma Casualties

Jaques Reifman,[1] Liangyou Chen,[1] Maxim Y. Khitrov,[1] Andrew T. Reisner[1,2]
[1]Bioinformatics Cell, Telemedicine and Advanced Technology Research Center, U.S. Army Medical Research and Materiel Command, MCMR-TT, 2405 Whittier Drive, Suite 200, Frederick, MD 21702 USA
[2]Massachusetts General Hospital Department of Emergency Medicine, Boston, MA 02114 USA

E-mails: jaques.reifman@us.army.mil; lchen@bioanalysis.org; mkhitrov@bioanalysis.org; areisner@partners.org

ABSTRACT

The military has long been interested in advanced decision-support capabilities for combat casualty care in which an automated computer algorithm processes available data and, through artificial intelligence, offers caregivers accurate information about the state of the casualty. However, two major obstacles have impeded these capabilities. First, routine vital signs have been speculated to be insensitive to prehospital major traumatic pathology. Second, there are numerous potential sources of decision-support failure, and it is not possible to investigate and address such potential limitations and demonstrate utility within the confines of a research laboratory. To address these obstacles, we retrospectively mined our trauma database consisting of vital signs and attribute data from 898 patients, and employed various signal-processing, artificial intelligence, and knowledge engineering technologies to develop an automated decision-support system. Our system for major hemorrhage diagnosis yielded an area under the receiver operating characteristic curve of 0.85 (95% confidence interval 0.80-0.90), with an 85% sensitivity and a 73% specificity, when retrospectively applied to the testing set of basic vital-sign data. In parallel, we developed a novel plug-and-play software/hardware system (termed APPRAISE) for automated, real-time data collection and prospective testing of decision-support algorithms in prehospital, clinical settings. Through simulations, we verified APPRAISE's real-time capability. Here, we summarize our technologies and findings in the development of an advanced medical system to reach the long-awaited goal of field deployment of automated decision-support tools for the triage and diagnosis of trauma casualties.

1.0 INTRODUCTION

Casualty care on the battlefield is challenging. In addition to major environmental distractions and dangers, caregivers may face suboptimal resources and incomplete diagnostic information. Proper assessment of the state of the casualty and determination of the emergent need (or not) for life-saving interventions can be problematic. The problem is exacerbated in mass casualty situations, where many casualties are tended by only a few caregivers who have received limited training on casualty care, while timely and proper diagnosis and treatment of casualties are critical for their ultimate survival. To overcome these challenges and optimize casualty outcome, it would be useful to develop and employ automated decision-support tools that could intelligently analyze early vital-sign data on the battlefield and autonomously generate reliable medical decisions. We have already witnessed decreases in morbidity and mortality in war wounds from the 19th century to the modern wars due to the adoption of better evacuation and treatment strategies, medical equipment, and care training systems [1]; however, about 15% of combat deaths in the recent Iraq war are still reported as being potentially survivable [2, 3]. The challenge for researchers is to develop decision-support

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
APR 2010		**00-00-2010 to 00-00-2010**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Automated Decision-Support Technologies for Prehospital Care of Trauma Casualties	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **U.S. Army Medical Research and Materiel Command,Telemedicine and Advanced Technology Research Center,504 Scott St,Fort Detrick,MD,21702**	8. PERFORMING ORGANIZATION REPORT NUMBER
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

Proceedings of the NATO/RTO HFM-182 Symposium on Human Use of Advanced Technologies and New Procedures in Medical Field Operations. Essen, Germany. 2010 April 19-21

14. ABSTRACT

The military has long been interested in advanced decision-support capabilities for combat casualty care in which an automated computer algorithm processes available data and, through artificial intelligence, offers caregivers accurate information about the state of the casualty. However, two major obstacles have impeded these capabilities. First, routine vital signs have been speculated to be insensitive to prehospital major traumatic pathology. Second, there are numerous potential sources of decision-support failure, and it is not possible to investigate and address such potential limitations and demonstrate utility within the confines of a research laboratory. To address these obstacles, we retrospectively mined our trauma database consisting of vital signs and attribute data from 898 patients, and employed various signal-processing, artificial intelligence, and knowledge engineering technologies to develop an automated decision-support system. Our system for major hemorrhage diagnosis yielded an area under the receiver operating characteristic curve of 0.85 (95% confidence interval 0.80-0.90), with an 85% sensitivity and a 73% specificity, when retrospectively applied to the testing set of basic vital-sign data. In parallel, we developed a novel plug-and-play software/hardware system (termed APPRAISE) for automated, real-time data collection and prospective testing of decision-support algorithms in prehospital, clinical settings. Through simulations, we verified APPRAISE?s real-time capability. Here, we summarize our technologies and findings in the development of an advanced medical system to reach the long-awaited goal of field deployment of automated decisionsupport tools for the triage and diagnosis of trauma casualties.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**14**	

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std Z39-18

tools that address the constraints of battlefield environments and assess their benefits.

The United States military has long been interested in developing advanced decision-support capabilities, which, based on physiological information collected by biosensors and an automated computer algorithm, can autonomously determine the injury state of casualties and formulate a prognosis and/or diagnosis. Such capabilities are especially important when triaging and monitoring multiple casualties during transport from the field to higher echelons of care. However, two major obstacles have impeded such capabilities. First, routine vital signs are notoriously unreliable in the prehospital environment, which is largely due to various confounding factors, such as motion artifacts, power supply interference, and changing psychological status of the soldiers, and the diagnostic value of prehospital vital signs for major traumatic pathologies has often been questioned [4-8]. Indeed, our automated *post hoc* analysis of vital-sign data of 898 trauma patients, collected during helicopter transport to a Level I trauma center, indicates that only 44% of heart rate (HR) records and 27% of respiratory rate (RR) records are of sufficient quality to be used for automated decision support [9, 10], and unreliable vital signs are significantly less informative for the diagnosis of major trauma pathologies than reliable vital signs [11, 12]. Second, there are numerous potential sources of decision-support failure, which can be due to the unique and unforeseeable battlefield conditions, confounding factors from medical interventions, and the high sensitivity of the algorithms to data outliers. Clearly, it is not possible to investigate and address such limitations and demonstrate utility within the confines of a research laboratory.

Over the past several years, our program has established a set of automated decision-support technologies for the prehospital care of trauma patients. These technologies were developed in a systematic fashion through retrospective mining of a civilian trauma database, and addressed several key issues relating to signal processing, artificial intelligence, and knowledge engineering. Specifically, our technologies are comprised of: *1*) signal-processing technologies that automatically determine the reliability of vital-sign data and extract vital-sign features; *2*) artificial intelligence classifiers that discriminate clinical outcomes, i.e., case and control groups, while considering real-world data problems, such as missing vital signs, noisy data, and unbalanced case-control groups; *3*) a plug-and-play software/hardware system for real-time field data collection and decision support; and *4*) prospective field validation. These technologies embody a fully automated decision-support system that can be taken into the field to monitor and diagnose trauma patients in real time. The system relies only on conventional vital-sign data, such as HR, RR, arterial blood oxygen saturation (SpO2), and systolic and diastolic blood pressures (BPs; SBP and DBP, respectively), and displays the maximal tolerance to noise and incomplete measurements that are emblematic of field-collected data. We have focused our system on the diagnosis of major hemorrhage, which is a major source of trauma mortality and is often treatable [13-15], and our system demonstrated good performance in hemorrhage diagnosis when retrospectively tested on the trauma database. Such capability is also useful for automated diagnosis, triage, and vigilant monitoring of combat casualties during long-distance air transport or in military medical treatment facilities.

In this paper, we review the above-mentioned technologies, focusing on their practical value in addressing two major obstacles: the unreliability and information content of field-collected vital signs and the ability to test decision-support tools in realistic environments. Our studies revealed that basic vital signs can offer surprisingly rich information for the automated diagnosis and decision support of trauma casualties. In addition, our preliminary assessment showed that our software/hardware system is capable of collecting and analyzing vital-sign data in real time. To this end, we propose a unique solution in which novel algorithms can be plugged into our real-time system in minutes, and they can be rapidly deployed in the field for prospective validation.

2.0 RETROSPECTIVE MINING OF TRAUMA DATA

2.1 Data Collection

Our decision-support system was developed based on the retrospective analysis of physiological vital-sign time-series data collected from 898 trauma-injured patients during transport by medical helicopter from the scene of injury to the Level I unit of the Memorial Hermann Hospital in Houston, TX [16-19]. The vital signs were measured by Propaq 206EL monitors (Welch Allyn; Skaneateles Falls, NY) [20] during transport, downloaded to an attached personal digital assistant, and ultimately stored in our database [21]. The time-series data consisted of high-frequency electrocardiogram (ECG), photoplethysmogram (PPG), and impedance pneumogram (IP) waveform signals recorded at 182, 91, and 23 Hz, respectively, and their corresponding monitor-calculated vital signs (HR, SpO2, and RR), recorded at 1-s intervals. In addition, SBP, DBP, and mean arterial pressure (MAP) were collected intermittently at multi-minute intervals. In addition, we collected 100 attribute data via retrospective chart review, which included demographics, injury descriptions, prehospital interventions, and hospital treatments [9, 10, 12, 15-19, 22, 23].

2.2 Outcome Definition

Our work focused primarily on the diagnosis of major hemorrhage induced by trauma, which is a significant source of mortality in the battle field and is often treatable [13-15]. We identified "case" and "control" outcomes for major hemorrhage using objective criteria. Because there are no perfect, indisputable retrospective definitions of major hemorrhage, we also tested the system using *alternative* hemorrhage definitions to verify that our system performed well given *any* reasonable definition of major hemorrhage. A good decision-support system should not be sensitive to the precise way in which the outcome of interest is defined; rather, it should yield similar performance for any reasonable definition of the clinical outcome. To test the performance of the system under alternative outcome definitions, we performed sensitivity analysis. The following definitions have been investigated:

- *Primary definition of major hemorrhage (cases)*: Major hemorrhage was defined as the receipt of blood transfusion within 24 h upon arrival at the hospital, along with documented anatomic injuries that were explicitly hemorrhagic. Such explicit injuries include one or more of the following: (*1*) laceration of solid organs, (*2*) thoracic or abdominal hematomas, (*3*) explicit vascular injury that required operative repair, or (*4*) limb amputation.

- *Alternative definitions of major hemorrhage*: We explored alternative definitions of major hemorrhage as receiving blood transfusion *or* a fluid transfusion of >3 liters, *and/or* documented anatomic injury (see above).

- *Controls*: In the primary outcome analysis, we excluded patients who received blood but did not meet the documented injury criteria, i.e., ambiguous hemorrhagic patients, and patients who died before arrival at the hospital (121 patients excluded); the remaining patients constituted the control group. In the primary and alternative definitions of major hemorrhage, we also explored control patients with/without exclusion of ambiguous hemorrhagic patients.

2.3 Demographics

Among the total 898 trauma patients, 110 (12%) patients did not have any vital-sign data collected during transport, 105 (12%) of the remaining patients had ambiguous hemorrhagic condition, and 9 (1%) additional

patients did not have any reliable vital-sign data based on our data reliability algorithms (see Section 3.1, "Vital-Sign Data Reliability"). The remaining 674 patients comprised our study population. Table 1 shows the demographics of the overall database as well as the study population.

Table 1: Demographics of the overall database and the study population

Characteristics	Overall database	Study population
Population size	898	674
Male	660* (73%)	501 (74%)
Female	234* (26%)	173 (26%)
Mean age, yr	37 (SD† 16)	37 (SD 15)
Blunt injury	778 (87%)	599 (89%)
Mortality	94 (10%)	41 (6%)
Prehospital intubated	201 (22%)	115 (17%)
Major hemorrhage‡	94 (10%)	78 (12%)

*4 patients had no assigned gender in the overall database.
†Standard deviation.
‡Received blood transfusion in the hospital and also had documented injuries that were consistent with major hemorrhage.

2.4 Statistical Methods

The diagnostic performance of our decision-support system was evaluated by constructing receiver operating characteristic (ROC) curves, which provide a tradeoff relationship between sensitivity and specificity of the system's decision outputs under a varying threshold, and by calculating the corresponding area under the curve (AUC) for each ROC curve, which provides a metric of the system's overall performance. An AUC of 1.00 represents a perfect system and 0.50 represents a random one. We used ROCKIT freeware (University of Chicago) [24] for these analyses, which automatically partitions the system's decision outputs into at most 20 intervals for the ROC-curve construction. ROCKIT assumes a binormal ROC model, that is, data for each of the outcomes (i.e., major hemorrhage cases and controls) are considered to be normally distributed. Under this assumption, each ROC curve is transformed into a straight line on the normal-deviate axes, whose ordinate intercept "*a*" and slope "*b*" are estimated by the maximum likelihood method [24]. The AUC is computed based on its mathematical relationship with *a* and *b*. The ROC curves estimated from this method are smoother than empirically evaluated ROC curves and can better represent the relationship between the outputs of the decision-support system and the clinical outcomes.

3.0 SINGAL-PROCESSING TECHNOLOGIES

Physiological time-series data collected in the field are noisy, and the diagnostic value of prehospital vital signs has been questioned [4, 5]. Even in-hospital vital signs are prone to erroneous measurement [6, 25]. The determination of vital-sign data reliability is a primary challenge in developing automated decision-support systems and, therefore, we employed various signal-processing technologies to rigorously and systematically analyze prehospital vital-sign data, and created a complete set of data-processing algorithms to evaluate the reliability of each vital-sign time series and extract useful information, or features, from the vital signs.

3.1 Vital-Sign Data Reliability

We have developed a set of algorithms to automatically evaluate the reliability of vital-sign time-series data collected in the field, utilizing redundant information present in high-frequency waveforms (ECG, PPG, and IP), based on which we algorithmically derived HR and RR, and the corresponding monitor-calculated vital signs (HR, SpO2, and RR), and the physiological relationships between SBP, DBP, and MAP [9, 10, 12]. Our decision-support system then used these data-reliability algorithms to automatically *identify* and then *exclude* unreliable vital-sign data. We have shown that reliable data are superior to and significantly more predictive of clinical outcomes than unreliable data (as determined by our automated algorithms) [12, 16]. Regarding the importance of data reliability, we have found the following:

- Our data-reliability algorithms rated each vital-sign datum on an integer-scale quality index (QI) ranging from "0" to "3," reflecting vital-sign reliability from the least to the most reliable [9, 10, 12]. We compared the algorithm-based rating of the vital-sign reliability, accepting QI≥2 as reliable, with human expert ratings and found that they concurred 90% of the time. However, in our database, we found that only 27% of RR, 44% of HR, 40% of SpO2, and 87% of BP records were reliable based on our algorithms.

- Our RR-reliability algorithm evaluated IP waveform (the source of the monitor-computed RR) and identified rhythmic and clean segments. We found that RR that was computed exclusively from these clean, rhythmic waveform segments is statistically superior to standard measures of RR as a predictor of hospital intubation and major hemorrhage [9, 16]. Indeed, computed in this fashion, elevated RR is as diagnostic of major hemorrhage as hypotension (ROC AUC of 0.77 for RR, of 0.71 for SBP, and of 0.60 for MAP) [16]. However, ~50% of the patients in our database did not have "reliable" RR.

- Another algorithm evaluated the reliability of the PPG waveform, a key component of pulse oximetry. In our database, we found that standard prehospital hypoxia (SpO2 ≤91%) has a positive predictive value (PPV) of ≤75% as a predictor of thoracic or intracerebral injury. Prehospital hypoxia concurrent with a clean PPG waveform has a significantly higher PPV (≥95%) for the same outcomes [17]. However, this measurement has low sensitivity because the majority (~70%) of the PPG-waveform data in our database were found to be unreliable.

- Our HR reliability algorithm evaluated the ECG waveform and considered if there is agreement between several different methods of computing HR. The algorithm was previously compared versus blinded human experts for several hundred ECG waveform excerpts [10]. When the HR algorithm identified reliable data, in 97% of the cases, blinded human experts concurred that the waveform was clean and, in 100% of those cases, concurred with the monitor's reported HR. When the algorithm identified unreliable data, humans agreed 85% of the time, suggesting that the algorithm was more selective than the human experts [10].

- The BP-reliability algorithm compared the HR measured by an oscillometric noninvasive BP cuff against the ECG HR and also checked that the relationships among SBP, MAP, and DBP were physiological [12]. Reliable SBP, as determined by this algorithm, was found to be statistically superior to unreliable SBP as a predictor of major hemorrhage [17].

3.2 Feature Extraction

"Feature" refers to some function of the continual physiological data, which seeks to provide diagnostically

useful information to the decision-support system. For example, a feature may be as simple as the moving average or the maximum value of a continually monitored vital sign. Other features can be computationally more sophisticated, such as the rate of change of a signal or its signal power at a specific frequency. We have developed algorithms that automatically compute features from standard vital signs, taking into consideration their variation through time. Accordingly, we explored parameters related to temporal patterns that have diagnostic value, such as peaks, troughs, means, standard deviations, differences, and rates of change as well as much more complex "shapes" that appear as the parameter is plotted through time. Regarding prehospital vital-sign features, we have several key findings:

- During 21 min of transport time to a trauma center, vital-sign trends do not offer clinically useful discriminatory power (ROC AUCs) to distinguish major hemorrhage cases versus controls (for all vital signs, ROC AUCs were not significantly better than 0.50) [17].

- For each of the five basic vital signs, HR, RR, SpO2, SBP, and DBP, their 21-min averages for major hemorrhage cases are statistically significantly different than controls [17].

- For each of the five vital signs, the magnitude of the 95% data range (averaged over all subjects) is considerably larger than the magnitude of any temporal trend [17].

- Using the 21-min averages of RR, HR, and SpO2 versus their initial values yielded improvements in ROC AUCs anywhere from +0.02 to +0.10 [17].

- Using wavelet transformation to represent the "shape" of HR plotted through time with a small number of parameters (i.e., wavelet coefficients) allows us to identify associations between temporal HR patterns and major hemorrhage. A combination of ten automatically selected temporal HR patterns yielded a sensitivity of 0.68 and a specificity of 0.79 [18]; however, only 47 hemorrhage cases out of the 94 total cases in our database had continuous reliable HR for >2 min as required for this analysis, and the most frequent HR temporal pattern was present in only 23% of these cases.

- We have combined multiple vital signs into a single "composite" variable via certain mathematical relationships. For example, SBP and DBP were combined through subtraction to define the pulse pressure (PP), and RR was divided by PP to define the "breath index" (BI) [11], which is a novel predictor of major hemorrhage. Composite variables show increased AUCs compared with individual vital signs [11, 15], but they are not statistically significantly different than a linear combination of the basic constituent vital signs [15].

- Respiratory-induced waveform variation (RIWV) in the PPG is a feature that has been associated with hypovolemia in mechanically ventilated patients and in controlled laboratory environments [26, 27]. We found that RIWV is a statistically independent predictor of major hemorrhage in our patient database, above and beyond a full set of standard vital signs, and that RIWV moderately improves the overall AUC of a multivariate statistical model [23]. However, as noted above, the majority of PPG-waveform data in our database were found to be "unreliable."

4.0 ARTIFICIAL INTELLIGENCE CLASSIFIERS

The diagnostic capability of our decision-support system is attained through artificial-intelligence-based classifiers, which are computer algorithms that take as inputs one or more vital-sign features and generate a

decision output to discriminate between two clinical outcomes, or classes, e.g., major hemorrhage cases and controls, in our application. These artificial intelligence classifiers are typically "trained" on a proportion of a retrospectively collected dataset, and subsequently applied to the remaining "testing" dataset (or prospectively collected data) to discriminate between the possible clinical outcomes.

4.1 Classifier Development

Our classifier algorithms were developed through a systematic retrospective analysis of our database. First, we thoroughly analyzed vital-sign features extracted from our patient database, such as the mean, variation, trend, and shape of the vital signs, and examined their discriminatory value under various time and data-reliability conditions. For example, as mentioned above, we found that using the 21-min averages of RR, HR, and SpO2 versus their initial values as input features to classifiers yielded improved performance, and that reliable RR, SBP, and SpO2 are more predictive of major hemorrhage than unreliable measurements. Next, we compared univariate classifiers (i.e., classifiers with only one input vital-sign feature) against multivariate classifiers (i.e., classifiers with multiple input vital-sign features), and established that multivariate classifiers provided improved discrimination. In addition, we compared the results obtained with linear against nonlinear classifiers, such as artificial neural networks and support vector machines, and found that nonlinear classifiers were not necessarily superior to simpler techniques [15, 22]. Finally, we used ensemble classifiers to address the problem of missing vital-sign data. An ensemble classifier consists of multiple linear "base" classifiers and an "aggregator" that combines the outputs of the base classifiers. Ensemble classifiers have been reported in the literature to provide improved classification accuracy [28], because the integration of multiple separate classifiers, reporting an "ensemble" behavior, is less susceptible to idiosyncrasies in the data. Regarding ensemble classifiers, we have the following key findings.

- Our initial tests indicated that all of the vital-sign variables contain information useful for classification and that there is no consistent best-feature set for input into a classifier [22]. We also verified that composite variables, including PP and BI, do not provide additional information than their constituent vital signs [11, 15, 22]. Therefore, we only included the five monitor-calculated basic vital signs, HR, RR, SpO2, SBP, and DBP, as input features into the ensemble classifier.

- There was no performance improvement, in terms of AUC, by using more than three inputs to the base classifier. Hence, our ensemble consisted of 25 linear base classifiers corresponding to the 25 possible combinations of one, two, and three input vital-sign features [22].

- In general, the performance of ensemble classifiers is only weakly dependent on the selected aggregation method, such as majority vote, median, or average [29, 30]. Our preliminary results confirmed this observation and, therefore, for convenience, we aggregated the results of the base classifiers by averaging.

- The performance of the ensemble classifier increases with the increasing length of reliable vital-sign time-series data [31, 32]. Accordingly, for classification at a given time t, we found that we should use all reliable data available up to time t.

4.2 Classifier Evaluation

We evaluated the performance of the ensemble classifier through separate training and testing datasets that were randomly selected from our database, where data from the training dataset were used to train the ensemble classifier, and data from the testing dataset were used to evaluate the classifier's diagnostic

performance. The use of separate training and testing datasets prevented us from overestimating the classifier's performance. To obtain a "representative" classifier performance, we trained/tested each classifier through 100 trials, each using 50% of the data for training and the remaining 50% for testing. Because the two datasets had unbalanced hemorrhage versus control classes (almost 1:8), to reduce classifier bias, the training classes were balanced by upsampling (i.e., randomly repeating) hemorrhage patients until both classes had the same number of patients. For each simulation, we computed the mean AUC and the corresponding 95% confidence interval (CI) based on the 100 testing trials.

The ensemble classifier, as the average of linear base classifiers using all combinations of one, two, and three vital-sign features, had good diagnostic performance, generating a test AUC of 0.85 (95% CI 0.80-0.90) for distinguishing hemorrhage from control patients in the study population. The classifier offered a sensitivity of 85% at a specificity of 73% when retrospectively applied to the testing dataset (i.e., the set of data that is different from the set used to train the algorithm). In contrast, we found that early field hypotension (i.e., SBP ≤110 mmHg) was only 47% sensitive and 87% specific, as shown in Figure 1. Note that these results are dependent on how features are computed and on data selection criteria, and may vary and degrade if applied to casualties with low-quality or missing data [22].

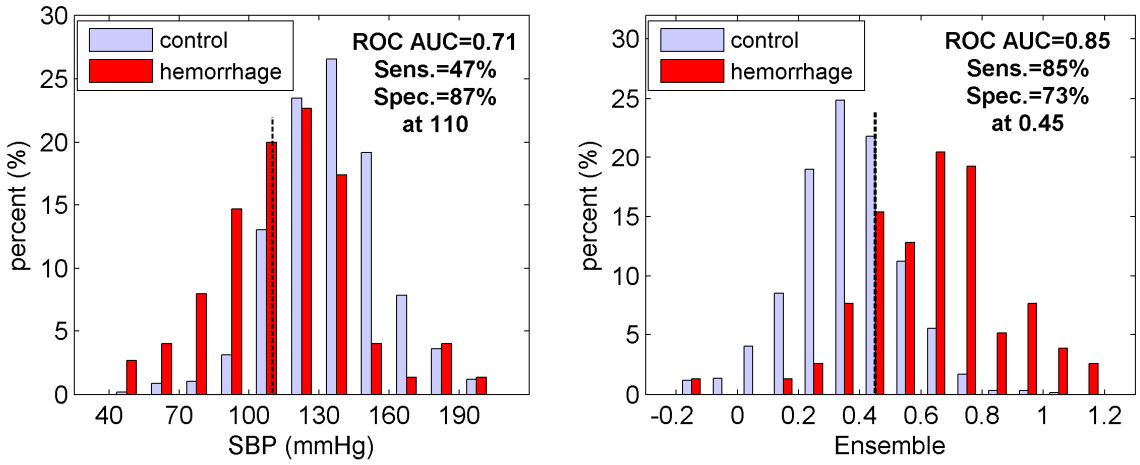

Figure 1. Histogram showing how effectively different physiology measurements are able to discriminate between casualties with major hemorrhage (dark red bars) and control patients (light blue bars). We summarize the receiver operating characteristic curve with the area under the curve (ROC AUC), where 1.00 represents a perfect classifier and 0.50 represents a random classifier. We also report the optimal performance point [33] (dotted line) for each method. Left: Using the initial field measurement of systolic blood pressure (SBP), with a cut-off of 110 mmHg, fewer than 50% of major hemorrhage patients are identified (sensitivity = 47%). Right: Using the output of our ensemble classifier [22], we see a statistically significant improvement in classification, with sensitivity as high as 85%.

The ensemble classifier was notably tolerant of missing data. Among the 674 patients in the study population, only 399 (59%) patients had reliable data for each vital sign at some time during the transport. This means that if we had used a conventional multivariate classifier for decision making that required the availability of each and every vital sign, we would have missed 275 (41%) patients. We did require, however, that each patient had at least one reliable measurement from any of the five basic vital signs. This rather mild requirement allowed us to diagnose all but 9 (1%) out of the total 898 patients. We are developing signal-processing technology that can recover vital signs from noisy waveform signals to further increase data availability [34].

5.0 DEVELOPMENT OF FIELD-DEPLOYABLE SYSTEM

Our decision-support algorithms for data processing and hemorrhage diagnosis were developed retrospectively using our trauma database. Such retrospectively developed algorithms may be subject to idiosyncrasies of our specific database, and may not perform as well in real-time, prospective clinical settings because of unanticipated operational problems and multiple confounding factors, such as motion artifacts, caregiver interventions, and sensor failures, which are ubiquitous in unrestricted clinical environments. To evaluate the prospective performance of any in-house-developed decision-support algorithm and assess its sensitivity and potential limitation in actual clinical environments, we developed a plug-and-play software/hardware system, termed APPRAISE (Automated Processing of the Physiologic Registry for the Assessment of Injury Severity), that can collect physiological data, and run and test decision-support algorithms in real time [35, 36].

5.1 The APPRAISE System for Real-Time Data Collection & Testing of Decision-Support Algorithms

Our APPRAISE system is an integrated hardware and software system developed for the real-time collection of physiological data and for the plug-and-play testing of novel decision-support algorithms during actual clinical operations. The APPRAISE hardware system consists of a standard-of-care vital-sign monitor, and a ruggedized ultramobile personal computer (PC) [see Figure 2]. The vital-sign monitor is a Welch Allyn's (Skaneateles Falls, NY) Propaq Encore 206EL monitor [20] with the Acuity Port option, and the ruggedized PC is Roper Mobile Technology's (Tempe, AZ) Switchback [37] running the Microsoft Windows XP operating system with 2 GB of memory and a 32-GB solid-state drive. The Propaq is mounted on top of a small cage, which houses the PC and acts like a pedestal (Figure 2, left panel). The PC is connected to the Propaq through an RS-232 to USB adapter.

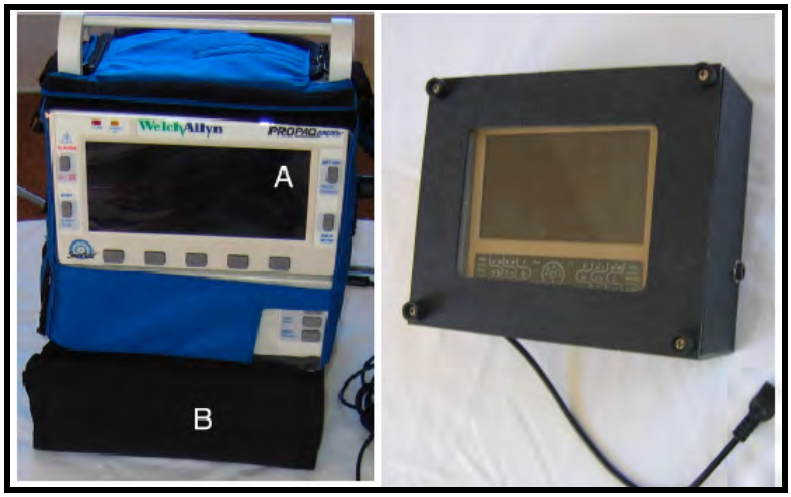

Figure 2. The APPRAISE software/hardware system for clinical field trials of decision-support algorithms [35, 36]. Left: A standard Propaq travel monitor (A; Welch Allyn; Skaneateles Falls, NY) is mounted atop a protective cage (B), which contains a ruggedized Switchback personal computer (PC) [Roper Mobile Technology; Tempe, AZ] [37]. These two devices are connected together; see text for details. Right: Bottom view of the cage, showing the exposed touch-sensitive screen of the ruggedized Switchback PC.

The APPRAISE software system consists of in-house-developed software for collecting physiological time-series waveform and vital-sign data, which runs on the ruggedized PC. During patient monitoring, the software system timestamps and stores physiological data received from the monitor. It then runs, in real time, decision-support algorithms implemented in MATLAB [38], which is a high-level computer language and the *de facto* standard for digital signal processing, data analysis, and rapid software prototyping throughout engineering and physical science communities, to perform analysis on the collected data. Any decision-support algorithm that is implemented in MATLAB can be run in the APPRAISE system and will require only a matter of several minutes to install. As a result of this plug-and-play functionality, it will be relatively feasible to deploy our system in additional arenas in which Propaq monitors are in use.

To test the APPRAISE system, as well as any decision-support algorithm it carries prior to their field use, we developed a Propaq emulator, which simulates the transmission and processing of physiological time-series data, such as the ones available in our trauma database. The emulator transmits data to the PC using the Propaq communication protocol, allowing us to test the functionalities of both the APPRAISE system and any accompanying decision-support algorithm. One exemplary use of the emulator is to introduce known types of errors in the simulated data, e.g., to mimic situations of short communication break, sudden unexpected hardware interruption, and signal corruption by certain noise, and test whether the entire system can function continuously regardless of these errors. Most importantly, after actual data have been collected, we can use the emulator to playback the time-series data in a deterministic and simulated environment, so that field failure scenario can be retrospectively examined. Such functionality is especially important for the development of medical decision-support tools that must work robustly and reliably under chaotic combat field environments.

5.2 Evaluation of the APPRAISE System

We assessed the performance of the APPRAISE system by simulating the real-time data environment for patients in our database, while running the decision-support algorithms we have created, including the data-reliability and hemorrhage-diagnosis algorithms. Specifically, we randomly selected 20 patients from the study population with different degrees of data reliability and data length, and ran the Propaq emulator to simulate real-time data transmission and processing. We compared the input data submitted to the APPRAISE system with what were made available to the decision-support algorithms to verify that data collection was functioning correctly. We also compared the results of the decision-support algorithms with those obtained off-line from a desktop PC. We were able to perfectly duplicate the algorithms' results, thus proving that the entire system was performing as intended.

Our decision-support algorithms were able to perform in real time on the APPRAISE system. The average algorithm execution time, including estimating vital-sign data reliability, vital-sign feature extraction, and ensemble classification, was 16 s, and the maximal time was 34 s for the randomly selected 20 patients. This performance test supports the prospective use of diagnostic decision support as frequently as every minute. In case any individual algorithm's execution time exceeded one minute, the algorithm would execute at the next available time slot, or immediately if two time slots had already been skipped. These results ensure that our decision-support algorithms can be readily deployed for real-time applications.

6.0 PROSPECTIVE FIELD VALIDATION

We have just started the initial phase of the prospective validation of the APPRAISE system, which integrates field data collection and decision support in a single small unit, for hemorrhage diagnosis. We have obtained

institutional review board approval from the Massachusetts General Hospital (MGH) and the U.S. Army (Fort Detrick, MD) to perform a prospective field trial of our APPRAISE system and related decision-support algorithms in acute trauma patients en route to MGH via Boston MedFlight helicopters.

7.0 CONCLUSION

We have developed a system in which data from standard vital-sign monitors are used as diagnostic indicators of major hemorrhage, incorporating cutting-edge, real-time artificial intelligence classifier algorithms that provide decision support for trauma casualty care. Our algorithms' retrospective performance, expressed as the area under an ROC curve was 0.85, which falls within a "good" classifier range. An AUC of 0.85 can be interpreted to mean that, applying our algorithm to two subjects, one from each outcome class, the subjects will be accurately classified 85% of the time [39]. The algorithms used only conventional apparatus, avoiding the need to train thousands of clinicians to use novel instrumentation, and they are applicable to 99% of the field patients who present any available vital sign. Such capabilities are not fanciful: we are currently field-testing these algorithms during the transport of civilian trauma casualties using the APPRAISE system.

The APPRAISE system we developed offers the ability to rapidly insert a novel decision-support algorithm into actual clinical operations. This provides major benefits. First, algorithm developers can get rapid feedback on any new algorithm. Because algorithm development can be challenging, it is essential to run real-world tests and make multiple iterative improvements. The system promotes iterative design cycles with minimal expense and time. Second, the system can simultaneously compare competing analytic strategies and algorithms at the same time. Such head-to-head comparisons will promote good technologic approaches and identify inferior ones, so that future development resources are invested appropriately. Finally, in the future, the system can be modified to accept and process novel investigational sensors and provide output of the decision-support algorithms to medics in real time. Ultimately, the system should illuminate the capabilities and limitations of standard vital signs for decision support and offer a benchmark against which novel sensors and approaches should be compared.

Given the initial successful development of the APPRAISE system for hemorrhage diagnosis, we plan to accelerate and expand the scope of our program's activities in the near future. Our objective is to deliver a complete, integrated suite of automated decision-support tools for assessing the need for life-saving interventions of trauma casualties due to: (1) major hemorrhage, (2) respiratory compromise, (3) traumatic brain injury, and (4) overall mortality. We will apply novel signal-processing techniques that *correct* unreliable physiological signals (reliance on noisy data significantly impairs algorithm performance, as discussed above), build statistical models that can be used for prospective diagnostic applications, and integrate the technologies we have developed, i.e., integration of data-reliability, data-correction, and multiple diagnostic algorithms, into a single, integrated software application. The system will be a fully functional, clinically validated prototype that can be provided to an industry partner for full productization. In doing so, we will also possess the infrastructure necessary for further testing of commercial implementations of these technologies. Ultimately, our work will establish the full extent of what can be done with the monitoring capabilities in widespread use today, so the U.S. Army can better evaluate other novel technologies and promote promising new sensors while disregarding those that are no more useful than standard vital signs.

ACKNOWLEDGEMENTS

This work was partially supported by the Combat Casualty Care Research Area Directorate of the U.S. Army

Medical Research and Materiel Command, Fort Detrick, Maryland, USA.

DISCLAIMER

The opinions or assertions contained herein are the private views of the authors and are not to be construed as official or as reflecting the views of the U.S. Army or the U.S. DoD. This paper has been approved for public release with unlimited distribution.

REFERENCES

[1] M. M. Manring, A. Hawk, J. H. Calhoun, and R. C. Andersen, "Treatment of war wounds: a historical review," *Clin Orthop Relat Res,* vol. 467, pp. 2168-91, Aug 2009.

[2] J. Holcomb, J. Caruso, N. McMullin, C. E. Wade, L. Pearse, L. Oetjen-Gerdes, H. R. Champion, M. Lawnick, W. Farr, S. Rodriguez, and F. Butler, "Causes of death in US Special Operations Forces in the global war on terrorism: 2001-2004," *US Army Med Dep J,* pp. 24-37, Jan-Mar 2007.

[3] J. F. Kelly, A. E. Ritenour, D. F. McLaughlin, K. A. Bagg, A. N. Apodaca, C. T. Mallak, L. Pearse, M. M. Lawnick, H. R. Champion, C. E. Wade, and J. B. Holcomb, "Injury severity and causes of death from Operation Iraqi Freedom and Operation Enduring Freedom: 2003-2004 versus 2006," *J Trauma,* vol. 64, pp. S21-6; discussion S26-7, Feb 2008.

[4] D. C. Garner, "Noise in medical helicopters," *JAMA,* vol. 266, pp. 515-6, Jul 1991.

[5] R. B. Low and D. Martin, "Accuracy of blood pressure measurements made aboard helicopters," *Ann Emerg Med,* vol. 17, pp. 604-12, Jun 1988.

[6] P. B. Lovett, J. M. Buchwald, K. Sturmann, and P. Bijur, "The vexatious vital: neither clinical measurements by nurses nor an electronic monitor provides accurate measurements of respiratory rate in triage," *Ann Emerg Med,* vol. 45, pp. 68-76, Jan 2005.

[7] Z. V. Edmonds, W. R. Mower, L. M. Lovato, and R. Lomeli, "The reliability of vital sign measurements," *Ann Emerg Med,* vol. 39, pp. 233-7, Mar 2002.

[8] C. L. Tsien and J. C. Fackler, "Poor prognosis for existing monitors in the intensive care unit," *Crit Care Med,* vol. 25, pp. 614-9, Apr 1997.

[9] L. Chen, T. M. McKenna, A. T. Reisner, and J. Reifman, "Algorithms to qualify respiratory data collected during the transport of trauma patients," *Physiol Meas,* vol. 27, pp. 797-816, Sep 2006.

[10] C. Yu, Z. Liu, T. McKenna, A. T. Reisner, and J. Reifman, "A method for automatic identification of reliable heart rates calculated from ECG and PPG waveforms," *J Am Med Inform Assoc,* vol. 13, pp. 309-20, May-Jun 2006.

[11] L. Chen, A. T. Reisner, A. Gribok, T. M. McKenna, and J. Reifman, "Can we improve the clinical utility of respiratory rate as a monitored vital sign?" *Shock,* vol. 31, pp. 574–80, Jun 2009.

[12] A. T. Reisner, L. Chen, T. M. McKenna, and J. Reifman, "Automatically-computed prehospital severity scores are equivalent to scores based on medic documentation," *J Trauma,* vol. 65, pp. 915-23, Oct 2008.

[13] A. Sauaia, F. A. Moore, E. E. Moore, K. S. Moser, R. Brennan, R. A. Read, and P. T. Pons, "Epidemiology of trauma deaths: a reassessment," *J Trauma,* vol. 38, pp. 185-93, Feb 1995.

[14] R. Peng, C. Chang, D. Gilmore, and F. Bongard, "Epidemiology of immediate and early trauma deaths at an urban Level I trauma center," *Am Surg,* vol. 64, pp. 950-4, Oct 1998.

[15] L. Chen, A. T. Reisner, T. M. McKenna, A. Gribok, and J. Reifman, "Diagnosis of hemorrhage in a prehospital trauma population using linear and nonlinear multiparameter analysis of vital signs," *Conf Proc IEEE Eng Med Biol Soc,* vol. 2007, pp. 3748-51, 2007.

[16] L. Chen, A. T. Reisner, A. Gribok, T. M. McKenna, and J. Reifman, "Can we improve the clinical

utility of respiratory rate as a monitored vital sign?," *Shock,* vol. 31, pp. 574–80, Jun 2009.

[17] L. Chen, A. T. Reisner, A. Gribok, and J. Reifman, "Exploration of prehospital vital sign trends for the prediction of trauma outcomes," *Prehosp Emerg Care,* vol. 13, pp. 286-94, Jul-Sep 2009.

[18] L. Chen, A. Gribok, A. T. Reisner, and J. Reifman, "Exploiting the existence of temporal heart-rate patterns for the detection of trauma-induced hemorrhage," *Conf Proc IEEE Eng Med Biol Soc,* vol. 2008, pp. 2865-8, 2008.

[19] L. Chen, A. T. Reisner, and J. Reifman, "Automated beat onset and peak detection algorithm for field-collected photoplethysmograms," *Conf Proc IEEE Eng Med Biol Soc,* vol. 2009, pp. 5689-92, Sep 2009.

[20] "Propaq Encore Monitors," *Internet: www.welchallyn.com/products/en-us/x-11-ac-100-000000000 1101.htm*, [accessed Feb 2010].

[21] T. M. McKenna, G. Bawa, K. Kumar, and J. Reifman, "The physiology analysis system: an integrated approach for warehousing, management and analysis of time-series physiology data," *Comput Methods Programs Biomed,* vol. 86, pp. 62-72, Apr 2007.

[22] L. Chen, T. M. McKenna, A. T. Reisner, A. Gribok, and J. Reifman, "Decision tool for the early diagnosis of trauma patient hypovolemia," *J Biomed Inform,* vol. 41, pp. 469-78, Jan 2008.

[23] L. Chen, A. T. Reisner, A. Gribok, and J. Reifman, "Is respiration-induced variation in the photoplethysmogram associated with hypovolemia in patients with acute traumatic injuries?" *In press, Shock* (Feb 2010).

[24] "ROCKIT," *Internet: http://www-radiology.uchicago.edu/krl/KRL_ROC/software_index6.htm,* [accessed Feb 2010], Chicago: Kurt Rossmann Laboratories, University of Chicago.

[25] D. W. Jones, L. J. Appel, S. G. Sheps, E. J. Roccella, and C. Lenfant, "Measuring blood pressure accurately: new and persistent challenges," *JAMA,* vol. 289, pp. 1027-30, Feb 2003.

[26] G. Natalini, A. Rosano, M. Taranto, B. Faggian, E. Vittorielli, and A. Bernardini, "Arterial versus plethysmographic dynamic indices to test responsiveness for testing fluid administration in hypotensive patients: a clinical trial," *Anesth Analg,* vol. 103, pp. 1478-84, Dec 2006.

[27] M. Shamir, L. A. Eidelman, Y. Floman, L. Kaplan, and R. Pizov, "Pulse oximetry plethysmographic waveform during changes in blood volume," *Br J Anaesth,* vol. 82, pp. 178-81, Feb 1999.

[28] T. G. Dietterich, "Ensemble Methods in Machine Learning," *First International Workshop on Multiple Classifier Systems,* vol. 1857, pp. 1-15, 2000.

[29] J. Kittler and F. M. Alkoot, "Sum Versus Vote Fusion in Multiple Classifier Systems," *IEEE Transactions on Pattern Analysis and Machine Intelligence* vol. 25, pp. 110-5, 2003.

[30] L. I. Kuncheva, "A theoretical study on six classifier fusion strategies," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 24, pp. 281-186, 2002.

[31] L. Chen, A. Gribok, A. T. Reisner, and J. Reifman, "Can We Improve Classifier Consistency and Reduce False Alarms for the Continual Monitoring of Patients?" *In preparation,* Feb 2010.

[32] A. T. Reisner, L. Chen, and J. Reifman, "Are prehospital vital signs indicative of hemorrhage in trauma casualties? A new look at an old problem," *In preparation,* Feb 2010.

[33] W. J. Youden, "Index for rating diagnostic tests," *Cancer,* vol. 3, pp. 32-5, Jan 1950.

[34] A. Gribok and J. Reifman, "A robust method to estimate instantaneous heart rate from noisy electrocardiogram waveforms," *Under review,* Feb 2010.

[35] M. Y. Khitrov, M. Rutishauser, K. Montgomery, A. T. Reisner, and J. Reifman, "A platform for testing and comparing of real-time decision-support algorithms in mobile environments," *Conf Proc IEEE Eng Med Biol Soc,* vol. 2009, pp. 3417-20, Sep 2009.

[36] J. Reifman, M. Y. Khitrov, A. T. Reisner, L. Chen, and T. M. McKenna, "A system for real-time collection and analysis of vital signs and prediction of clinical outcomes, Invention Disclosure," *U.S. Army Medical Research and Materiel Command, Ft. Detrick,* Oct 2009.

[37] "Switchback," *Internet: www.ropermobile.com/products/switchback*, [accessed Feb 2010].

[38] "The MathWorks," *Internet: www.mathworks.com*, [accessed Feb 2010].

[39] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology,* vol. 143, pp. 29-36, Apr 1982.