Available online at www.sciencedirect.com
ScienceDirect
Journal of Biomedical Informatics 41 (2008) 469–478
Journal of Biomedical Informatics
www.elsevier.com/locate/yjbin

Decision tool for the early diagnosis of trauma patient hypovolemia

Liangyou Chen [a], Thomas M. McKenna [a], Andrew T. Reisner [a,b]
Andrei Gribok [a,c], Jaques Reifman [a,*]

[a] *Bioinformatics Cell, Telemedicine and Advanced Technology Research Center (TATRC), Building 363 Miller Drive, US Army Medical Research and Materiel Command (USAMRMC), Frederick, MD 21702-5012, USA*
[b] *Massachusetts General Hospital, Department of Emergency Medicine, Boston, MA 02114, USA*
[c] *Nuclear Engineering Department, University of Tennessee, Knoxville, TN 37919, USA*

Received 25 April 2007
Available online 18 January 2008

Abstract

We present a classifier for use as a decision assist tool to identify a hypovolemic state in trauma patients during helicopter transport to a hospital, when reliable acquisition of vital-sign data may be difficult. The decision tool uses basic vital-sign variables as input into linear classifiers, which are then combined into an ensemble classifier. The classifier identifies hypovolemic patients with an area under a receiver operating characteristic curve (AUC) of 0.76 (standard deviation 0.05, for 100 randomly-reselected patient subsets). The ensemble classifier is robust; classification performance degrades only slowly as variables are dropped, and the ensemble structure does not require identification of a set of variables for use as best-feature inputs into the classifier. The ensemble classifier consistently outperforms best-features-based linear classifiers (the classification AUC is greater, and the standard deviation is smaller, $p < 0.05$). The simple computational requirements of ensemble classifiers will permit them to function in small fieldable devices for continuous monitoring of trauma patients.
© 2007 Elsevier Inc. All rights reserved.

Keywords: Linear classifier; Ensemble classifier; Hemorrhage; Hypovolemia; Vital-signs; Decision assist; Monitoring; Physiology

1. Introduction

Two focuses of contemporary patient monitoring research include "smart" algorithms, which interpret multiparameter trends [1], and wearable sensors, which can have a wide range of form factors, some as innocuous as an article of clothing [2]. These new capabilities may improve decision-support in classic hospital environments and extend monitoring to novel arenas, such as the home or even a battlefield. In this investigation, our goal is to automate the detection of major hemorrhage (i.e., a physiologic state of hypovolemia) using a classification algorithm based on standard vital-signs measured before arrival at a receiving trauma center.

Consider a scenario in which a single field medic tends to four casualties. Knowing precisely which casualty is bleeding seriously would be invaluable for the caregiver, prompting field interventions (e.g., tourniquets and volume resuscitation), setting evacuation priorities, and activating the necessary resources. However, even in controlled clinical environments, such as emergency departments [3] and even intensive care units [4], conventional physiologic monitoring yields data that are noisy, incomplete, or erroneous due to artifact. The unreliability of pre- and in-hospital physiologic monitoring poses a major challenge for the development of advanced decision-support applications.

Practically speaking, any real-time decision-support algorithm should meet at least two important specifications. First, of course, the algorithm should be reasonably accurate. Second, the algorithm should provide consistent

* Corresponding author. Fax: +1 301 619 1983.
E-mail addresses: lchen@bioanalysis.org (L. Chen), Thomas.McKenna2@us.army.mil (T.M. McKenna), areisner@partners.org (A.T. Reisner), agribok@bioanalysis.org (A. Gribok), jaques.reifman@us. army.mil (J. Reifman).

1532-0464/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2007.12.002

performance despite inconsistent data availability and quality. We recently described methods to automatically distinguish reliable from artifactual physiologic measurements in support of this objective [5,6]. In this paper, we consider the related problem of missing measurements, since a consistent set of complete physiological measurements is difficult to guarantee. In hospital, especially for ambulatory patients, electrocardiography (ECG) leads may be disconnected or a pulse oximeter probe may come off a finger. Out of hospital, especially during helicopter transport of a patient, consistent data collection may be even more challenging [7].

The development of a classifier to identify major hemorrhage in trauma casualties is important for both military and civilian applications, because trauma is the leading cause of death for Americans ages 1 through 44 years [8], and major hemorrhage is the singular treatable cause of trauma mortality [9,10]. The earlier life-threatening hemorrhage is detected, the greater the opportunity exists for caregivers to provide life-saving therapy. In the future, a classifier that provides consistent performance despite inconsistent data availability may prove valuable for decision assistance if inserted into prehospital travel monitors for relatively short transports (e.g., from injury scene to civilian receiving hospital) or longer transports (e.g., military Critical Care Air Transport teams that use specially equipped aircraft to evacuate critically-injured patients from field hospitals to regional medical centers) [11]; into conventional hospital bedside monitoring systems; or in cutting-edge "wearable" systems worn during high-risk activities (e.g., military operations or firefighting).

2. Methods

2.1. Trauma data

This study is based on physiological time-series data collected from 898 trauma-injured patients during transport by medical helicopter from the scene of injury to the Level-I trauma center at the Memorial Hermann Hospital in Houston, Texas. Additional attribute data were collected retrospectively via chart review [12,13]. The time-series variables were collected by ProPaq 206EL vital-sign monitors [14], downloaded to an attached personal digital assistant, and ultimately stored in our database. The variables consist of ECG, photoplethysmogram, and respiratory waveform signals recorded at approximately 182, 91, and 23 Hz, respectively, and their corresponding monitor-calculated variables recorded at 1-s intervals [heart rate (HR), oxygen saturation of arterial hemoglobin (SaO_2), and respiratory rate (RR)]. In addition, systolic (SBP), mean (MAP), and diastolic (DBP) blood pressures were collected intermittently at multiminute intervals. The patient attribute data include items such as demographics, injury description, prehospital interventions, and hospital treatments. There are 100 variables of this type for each patient, and these data have already been subjected to a mining exercise [15].

2.2. Datasets

2.2.1. Inclusion/exclusion criteria and outcomes

Two classes of patients are identified in the trauma database; those who received blood in the emergency room (hemorrhage, 169 cases) and those who did not (control, 729 cases). The patients who received blood are also required to have documented injuries that are consistent with hemorrhage, with at least one occurrence of the following: (a) laceration of solid organs, (b) internal bleeding as indicated by abdomino-pelvic hematoma or hemothorax, or (c) explicit vascular injury and operative repair, or limb amputation. Patients who received blood but do not meet the documented injury criteria (75 cases) are excluded from analysis.

Two datasets, referred to as Total and Illustrative datasets, are extracted from this patient population to develop and validate the classifier (Table 1).

2.2.2. Total dataset

This dataset includes many subjects with missing vital-sign data. It provides a test bed to evaluate the diagnostic capability of the classifier on the widest, most representative patient population. Patients in this dataset have at least one nonzero vital-sign (HR, RR, DBP, SBP, or SaO_2) available in every 2-min window during the patients' initial 16 min of transport. In our database, 23 hemorrhage and 173 control cases do not meet this minimal-data criterion.

2.2.3. Illustrative dataset

This is a subset of the Total dataset, comprised of patients with a complete set of *all five* of the vital-sign variables measured during the 5- to 7-min interval of their transport to the hospital. This dataset is used as a test bed for several computational exercises that require complete vital-signs for all included subjects, as described in Section 2.4.

Table 1
Population number and demographics of patients constituting the Total and Illustrative datasets

Dataset	Population	Gender[a]		Mean age	Type of injury		Control	Hemorrhage
		Male	Female		Blunt	Penetrating		
Total	627	473	153	38.8	555	65	556	71
Illustrative	492	373	119	38.2	435	51	437	55

[a] One patient had no assigned gender in the Total dataset.

2.3. Data specification

The values of the vital-sign variables used for classifier training and testing are calculated by three methods (''best-quality 5-s data,'' ''first 5-s data,'' and ''all data combined''). Subsequently, we compare these three methods. All calculations are based on 2-min time windows.

2.3.1. Best-quality 5-s data

All vital-sign data are rated in terms of their reliability by methods reported previously [5,6]. The best-quality (i.e., most reliable) data that are continuous for at least 5 s are identified in the time window, and the best-quality 5-s value is calculated as the mean of the first 5 s of the best-quality data. Note that, if all of the data in the time window are of poor quality, then the calculated value may be unreliable. The mean value is used because experiments with other estimators, such as the median, showed no performance differences. The data quality selection method intrinsically removes data outliers.

2.3.2. First 5-s data

Here, we use the mean of the *first* 5 s of vital-sign data in the time window, *without regard to their quality*. This method serves as a control for comparison with the other methods.

2.3.3. All data

With this method, vital-sign variables are calculated as the mean of all data in the 2-min time window, a longer time interval over which data are averaged, in contrast with the shorter 5-s time intervals used above.

2.4. Linear classifier and feature selection

In this paper, linear classifiers are used for discriminating between two patient outcome classes, control and hemorrhage, selecting the most-informative ''best'' vital-sign features, and constructing ensemble classifiers. Because a linear classifier can normally be applied only when all input variables are available (i.e., without missing variables), all computations involving standalone linear classifiers are performed exclusively to the Illustrative dataset, in which subjects have a complete set of vital-sign variables.

Linear classifiers employ a linear discriminant function $f = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + w_0$, where the vector of coefficients $\boldsymbol{w}^{\mathrm{T}}$ and the coefficient w_0 are learned from a training set, to evaluate a given input vector $\boldsymbol{x}$ against two classes. The linear classifier used here is trained using a least-squares method [16], which minimizes the squared difference between the classifier outputs, which generally fall in the $[0.0, 1.0]$ range, and the target outputs represented by binary 0 and 1 values. A decision threshold θ is used for classification, i.e., assigning a given input vector $\boldsymbol{x}'$ to the hemorrhage class if $f(\boldsymbol{x}') > \theta$.

2.4.1. Training and testing protocol

To obtain a ''representative'' classifier performance, each classifier (using either the Total or the Illustrative dataset) is trained/tested through 100 trials. Given a ratio of the dataset to be used for classifier training and testing, for each trial, the training data are randomly selected from the dataset without replacement, and the remaining data are used for testing. Because the two datasets have unbalanced control versus hemorrhage classes (almost 8:1), to reduce classifier bias, the classes are balanced by undersampling (i.e., randomly dropping) control patients until both classes have the same number of patients.

The ability of a classifier to accurately classify patients into the appropriate control and hemorrhage classes is quantified by the area under the receiver-operating-characteristic (ROC) curve (AUC) [17]. In Section 3, we report the average AUC for the 100 trials along with the associated standard deviation (SD). All AUCs refer to classifier performance on *test* data that are not used for training. The AUCs are calculated by trapezoidal integration.

2.4.2. Feature selection using a wrapper method

The ''best'' (most-informative) variables to be used in the input vector $\boldsymbol{x}$ can be selected by filter or wrapper methods [18]. Here, we use the wrapper method, which performs feature selection based on classifier performance. There are 31 possible combinations of the five vital-sign variables, and whichever combination yields the highest AUC is termed the ''best-features'' classifier. The wrapper procedure used here follows the approach described in Guyon and Elisseeff [18], and is summarized below.

The wrapper method is applied within the context of the training and testing protocol for the linear classifier discussed above. However, it only applies to the Illustrative dataset, in which each subject has all five vital-sign variables. For a given ratio of classifier training/testing data, and for each of the 100 training/testing (''outer'') trials, the wrapper procedure involves a set of 100 additional trials (''inner'' trials) performed with the training data. For each inner trial, the training data are randomly sampled (without replacement) so that 50% are used to train 31 different linear classifiers (each employing one of the 31 possible feature combinations), and the remaining 50% are used for testing. Next, we compute the testing AUC for each one of the 31 classifiers. This process is repeated 100 times (corresponding to the 100 inner trials), and the average AUC for each classifier over these 100 (inner) trials is used to identify the ''best'' features, which are then used to train the associated ''best-features'' classifier with the entire training data set. Finally, we compute the AUC for the ''best-features'' classifier using the testing data. This process is repeated 100 times, once for each one of the 100 outer trials of the training and testing protocol.

2.5. Ensemble classifier

We employ ensemble classifiers to address the problem of missing vital-sign data. In addition, ensemble classifiers have been reported to provide improved classification accuracy [19], because the integration of multiple separate classifiers, reporting an "ensemble" behavior, is less susceptible to idiosyncrasies in the data. An ensemble classifier consists of multiple linear "base" classifiers and an "aggregator" that combines the decisions of the base classifiers. Aggregation can be achieved by different methods, such as majority vote, median, or average. In general, the performance of ensemble classifiers is weakly dependent on the selected aggregation method [20,21]. Our preliminary results confirm this observation; therefore, for convenience, we aggregate the results of the base classifiers by averaging.

Using all combinations of the five vital-sign variables (HR, RR, DBP, SBP, and SaO_2), the largest-possible ensemble classifier consists of 31 base classifiers, including five classifiers with one input, 10 classifiers with two inputs, 10 classifiers with three inputs, five classifiers with four inputs, and one classifier with five inputs. However, our initial tests show that there is no performance improvement, in terms of AUC, by using more than three inputs to the base classifier. Hence, our ensemble consists of the 25 linear base classifiers corresponding to the 25 possible combinations of one, two, and three input variables.

Ensemble classifiers are employed in two groups of computations: (a) using the Illustrative dataset for a direct comparison with the "best-features" linear classifier (see Section 2.4 above), and (b) using the Total dataset for benchmarking the performance against real-world applications with frequent missing data. In the latter case, we need to make adjustments to the training and testing protocol discussed in Section 2.4. During both training and testing, the number of base classifiers used (from 1 to 25) varied, depending on the availability of vital-sign data for each patient.

2.6. Classifier performance and statistical analysis

Differences between AUCs of two classifiers, e.g., ensemble versus best-features classifier, are tested for statistical significance by Wilcoxon matched-pairs signed-ranks tests [22]. The test verifies whether the observed difference between two sets of observations is statistically not different from zero, which represents the null hypothesis. The Wilcoxon test is a nonparametric analogue to the paired Student's *t*-test, but it allows the differences to be non-normally distributed.

3. Results

3.1. Individual variable versus composite-variable features

We test the hypothesis that the interaction of variables with each other (i.e., a composite relationship, such as HR/SBP or SBP–DBP) offers more information than separately inputting each of the same basic variables (HR, SBP, and DBP) into a classifier. HR/SBP is known as the shock index and SBP–DBP as the pulse pressure, and both are effective at signaling cardiovascular hypovolemia in certain applications [23,24]. We test the classification performance of linear classifiers using these composite-variables as input features, along with three additional composite features (Table 2) that appeared promising based on attempts to generate useful features by our group. For comparison, we test the performance of linear classifiers using the same variables as individual input features (Table 2). No significant differences in classification performance are apparent, suggesting that either variable format is adequate for input into the classifier. We observe the same outcome when applying nonlinear classifiers (feedforward artificial neural networks and support vector machines, results not shown). Consequently, for all subsequent analyses, we only use the basic vital-signs as input features into all classifiers.

3.2. Best-features-based classifiers versus ensemble classifiers

We evaluate whether a single classifier using the best set of vital-sign inputs ("best-features" classifier) provides better discrimination than an aggregation of classifiers using all possible sets of the variables ("ensemble" classifier). Using the linear wrapper method (Section 2.4), we resample the Illustrative dataset 100 times, and observe that there is no consistent "best" set of vital-sign variables to use as input features to a linear classifier. The most common best-features, composed of the HR and SBP variables, are selected only 14% of the time (Fig. 1). It is notable that, within the 100 resamples, a majority of the combinations that can be obtained using five vital-sign variables are selected as the best-feature set (21 out of a maximum

Table 2
Comparison of the classification performance of linear classifiers using independent variables versus composite-variable features

Variables	Test AUC
Shock index = HR/SBP	0.76 (SD 0.06)
HR, SBP	0.75 (SD 0.06)
Pulse pressure (PP) = SBP − DBP	0.73 (SD 0.06)
SBP, DBP	0.71 (SD 0.07)
Hemorrhage Index = (HR × RR)/(MAP[a] × PP)	0.73 (SD 0.06)
HR, RR, SBP, DBP	0.74 (SD 0.06)
RR/PP	0.67 (SD 0.08)
RR, SBP, DBP	0.72 (SD 0.06)
HR/PP	0.75 (SD 0.10)
HR, SBP, DBP	0.75 (SD 0.07)

AUCs (mean and SD) show test results from 100 trials, where for each trial 50% of the Illustrative dataset are used for training and 50% for testing.

[a] Mean arterial pressure (MAP) = (1/3) × SBP + (2/3) × DBP.

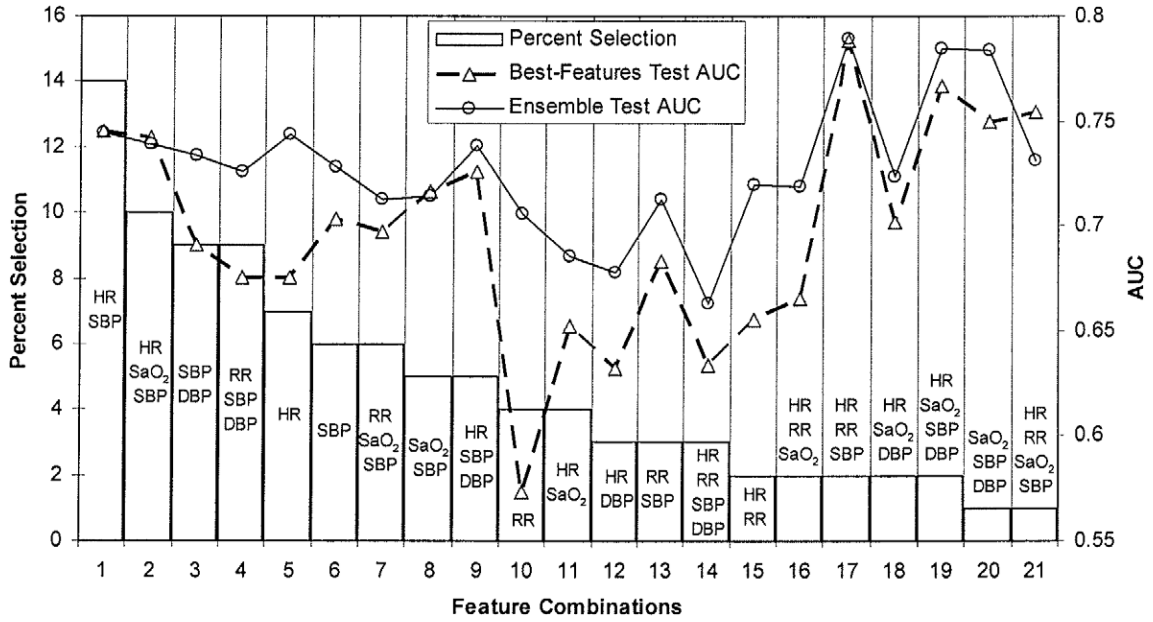

Fig. 1. Percent selection (left ordinate) denotes the number of times a vital-sign combination (indicated on the histograms) is selected as the best-feature by the wrapper method. The AUC (right ordinate) denotes the mean classification performance by the best-features (triangles) and ensemble (circles) classifiers. The best-features and ensemble classifiers are trained and tested on the same 100 sets (30% training and 70% testing) resampled from the Illustrative dataset. The AUCs indicate the mean classification performance by the best-features and the ensemble classifiers for the sets characterized by each of the feature combinations. Variability of the AUCs is not shown for clarity but averaged 0.04 AUC units over all of the feature combinations for both the best-features and ensemble classifiers.

of 31 possible combinations). These results indicate 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, since the best-features are attained only within the context of a selected population. This is a key finding and suggests that classifiers based on presumptive best-features will not be stable across larger populations.

The best-features classifiers are compared with the ensemble classifier. The mean AUCs for each of the 21 combinations are compared, and the ensemble classifier consistently performs better than each of the best-features classifiers (Fig. 1). The difference in performance is statistically significant ($p < 0.001$).

A further comparison of best-features versus ensemble classifiers is performed to determine the sensitivity of classifier performance to the ratio of data used for training and testing. Both classifiers show diminished performance, expressed as mean AUCs, at small training-to-testing ratios (upper panel, Fig. 2). The AUCs became more erratic at either extreme of the training-to-testing ratio and yield an increase in the standard deviation, SD (lower panel, Fig. 2). The ensemble classifier performs better than the best-features classifiers in terms of mean AUCs and their associated SDs ($p < 0.001$ and $p < 0.05$, respectively). Based on these and the previous results, it is clear that linear classifiers in an ensemble structure perform better than single linear classifiers applied to input features identified as "best" from a subsample of a population.

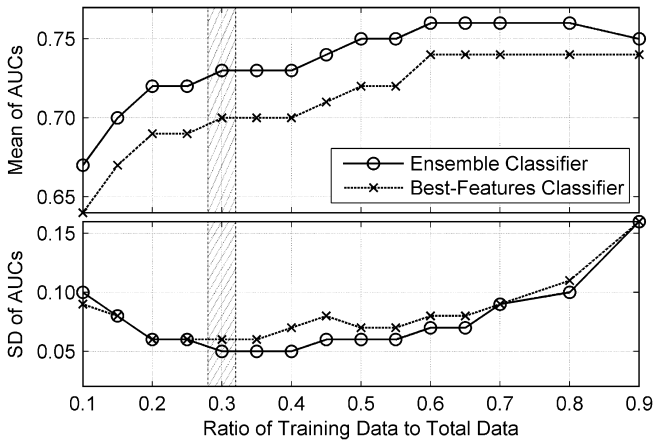

Fig. 2. Comparison of means and SDs of test AUCs calculated using ensemble and best-features classifiers. Given a set ratio of training data, 100 random sets of training and testing data are taken from the Illustrative dataset. The best vital-sign features are selected by the wrapper method (Section 2.4) from the training data, and a linear classifier is trained using the best-features. AUCs over the 100 test sets are averaged, and the mean and SD are shown in the upper and lower panels, respectively (crosses). The same 100 random sets are used to train and test ensemble classifiers, and their associated mean AUCs and SDs are also shown in the upper and lower panels (circles). For comparison, the shaded area denotes the mean of all of the best-features and ensemble classifier AUCs and SDs derived from the 30% training and 70% testing sets in Fig. 1.

3.3. Ensemble classifier performance

The ensemble classifier, as the average of linear base classifiers using all combinations of one, two, and three

vital-sign variables, has a good classification performance, generating a test AUC of 0.76 (SD 0.05) for distinguishing hemorrhage from control patients in the Total (i.e., maximum patient population) dataset. The associated ROC curve optimal operating point, as calculated by both the Youden index and the closest to (0, 1) criteria [25,26], yields a sensitivity of 0.69 (SD 0.08) and specificity of 0.68 (SD 0.09); alternatively, the specificity is 0.40 (SD 0.10) at a clinically-relevant sensitivity of 0.90. The classification performance is not due to fortuitous events because random scrambling of control and hemorrhage classes in the dataset decreases the classifier performance to an AUC of 0.59. These results indicate that the classifier learns information that is present in the vital-sign data.

A major advantage of an ensemble classifier is that it can deal effectively with missing data. This is important because it is likely that a full set of vital-sign data will occasionally be unavailable for a patient due to sensor malfunction, misplacement, motion artifacts, or other circumstances. This situation occurred while collecting data for patients comprising the Total dataset. The HR vital-sign records are present in 99% of the cases, but the other vital-sign records are only present 91% of the time. Requiring combinations of the vital-sign data will further decrease the population until, in the most restrictive case in which the patients must have all five vital-sign records (i.e., the Illustrative dataset), only 78% of the original Total patient population remains. The ensemble classifier is relatively resistant to such missing data; randomly dropping vital-sign variables from the Illustrative dataset only slowly degrades the classifier's performance. For instance, a loss of 40% of the vital-sign data results in only a 9% decrement in performance (Table 3).

Table 3
The effect of randomly removing variables on ensemble classifier performance

Ratio (%)	AUC
0	0.75 (SD 0.06)
5	0.74 (SD 0.06)
10	0.73 (SD 0.06)
15	0.72 (SD 0.06)
20	0.71 (SD 0.07)
30	0.69 (SD 0.07)
40	0.68 (SD 0.08)
50	0.65 (SD 0.07)
60	0.62 (SD 0.09)

Vital-sign variables are randomly dropped in set ratios from the Illustrative dataset, and the average AUC, based on 100 trials at each ratio, is calculated. The ensemble classifiers are trained on a random 50% sample of the changed dataset and tested on the remaining 50%.

3.4. Ensemble classifier performance over patient transport time and data quality

The influence of time and data properties on classifier performance is compared by training three ensemble classifiers at a single time point on variables calculated from vital-sign data by different methods (Section 2.3) and then by testing the classifiers using equivalently calculated variables as input in sequential 2-min time windows over a total of 16 min of patient transport time. All three classifiers significantly improve their classification performance over time (Fig. 3, $p < 0.05$ by t-tests of the linear regression coefficients), suggesting that the ensemble classifier responds to time-dependent changes in the vital-sign data. The amount of data used to calculate the variables is important if unqualified data are used. As Fig. 3 shows, the average AUCs of classifiers using variables calculated as the mean over the 2-min window of data are significantly better than those of classifiers using variables calculated as

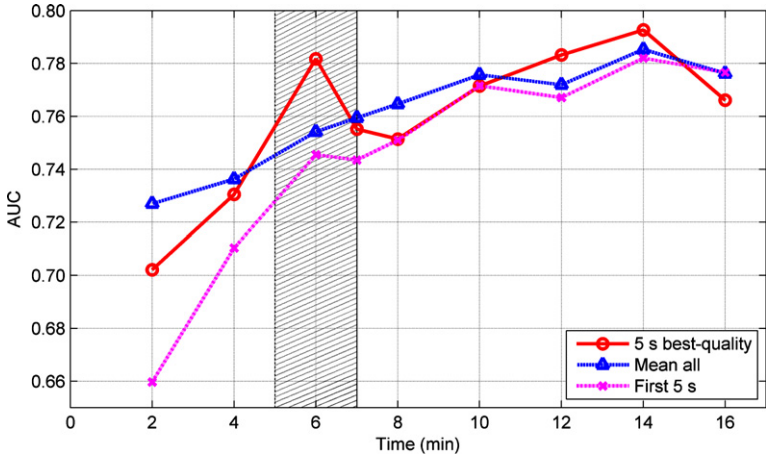

Fig. 3. Ensemble classifiers are trained and tested on 100 random sets (50% training and 50% testing) drawn from the Total dataset using vital-sign variables calculated by three different methods, as described below. The classifiers are initially trained and tested using data collected during the 5- to 7-min interval of patient transport (shaded area) and then tested on the remainder of the data in sequential 2-min intervals throughout 16 min of study. The values of the vital-sign variables are calculated in 2-min time windows as the means of: (a) 5 s of the best-quality data (5 s best-quality, circles), (b) the entire 2-min window (mean all, triangles), and (c) 5 s of unqualified data (first 5 s, crosses). The mean AUC results from classification of the 100 test sets are shown. The standard deviations are not shown for clarity but averaged 0.05 AUC units over 16 min for each of the data instances.

the mean over the first 5 s of data from each 2-min window ($p = 0.008$). However, taking the reliability of the data into account alters this relationship. Variables calculated using only 5 s of the best-quality data yield classification performance that is not significantly different from those attained from variables calculated using the 2-min mean of unqualified data ($p = 0.734$). Finally, classification performance is better for classifiers tested on variables calculated from 5 s of the best-quality data compared with variables calculated from 5 s of data chosen, without consideration of its quality, from the beginning of each 2-min window ($p = 0.027$).

4. Discussion

We have formulated an algorithm that uses basic vital-signs to identify major hemorrhage in trauma patients during helicopter transport to a hospital, i.e., identifies which casualties are bleeding so severely that they will require a life-saving blood transfusion. Based on our retrospective analysis of 627 subjects with 71 cases of major hemorrhage, this classifier's performance, expressed as the area under a receiver operating characteristic curve, is near 0.75, which falls within a "good" classifier range. Over the course of transport, the typical patient has progressively less missing data (26% missing variables in the first 2 min, dropping to 5% missing variables beyond 8 min), and our classifier grows even more accurate, peaking near 0.80 (Fig. 3). This classifier functions without regard to gender, age, type of injury, or other confounding factors, such as medics' intervention during transport. An AUC of 0.80 can be interpreted to mean that, applying this classifier to two subjects, one from each outcome class, the subjects will be accurately classified 80% of the time [27].

There are three major issues to discuss. First, if there were *no* missing vital-sign data, how well would our ensemble classifier perform, compared with other options? Second, is our ensemble classifier an appropriate solution to missing data? Third, what implications do our results have for real-world application?

We explore the first issue with the "Illustrative dataset," which is the set of subjects with complete vital-sign data. One important finding is that there is no "best" vital-sign(s) for identifying hemorrhage: when we examine 100 subsets of this dataset, there is *no one consistent set of vital-signs* that are the most discriminatory (Fig. 1). For instance, in 2% of the subsets, HR, SaO_2, and DBP prove to be the best multivariate discriminator between major hemorrhagic and control cases. In 4%, RR and SBP are the best. The most frequent "best set," SBP and HR, occurs in merely 14% of the subsets. Overall, there are 21 different "best sets" of vital-signs identified during our exploration of 100 randomly selected subsets. The fact that there is no consistent "best set" of vital-signs, and that all the vital-signs contain at least some discriminatory information, means that a conventional linear classifier runs the risk of overfitting to any subpopulation. Also, because of this information heterogeneity, a conventional linear classifier may have inconsistent accuracy in different subsets. When we compute these "best set" regression models against the ensemble classifier using all five vital-signs, we find (unsurprisingly) that the "best set" underperforms.

As seen in Fig. 1, the ensemble classifier tends to perform as well as, or better than, a single linear classifier when applied to 100 training/testing trials. The advantage of ensemble classifiers appears to persist through different ratios of training/testing cases, as shown in Fig. 2. Lastly, we show that the ensemble classifiers are indeed learning the "true" underlying vital-sign profiles of major hemorrhage. When we perform a simulation that randomly changes the subjects' outcome classes while preserving their vital-signs, classifier performance degrades, as expected. This provides additional confirmation that our classifier can identify vital-sign patterns consistent with hemorrhage.

Ensemble classifiers offer statistical, computational, and representational advantages compared with a single classifier [19,28]. Statistically speaking, a single classifier, generated from a particular training set, may not apply to a broader population, whereas the *average* of an *ensemble* of separate classifiers should have a more consistent performance throughout different subgroups, yielding a more stable, and perhaps more accurate, performance. Computationally, the ensemble classifier may better approximate the "true," but unknown, discriminatory function when multiple base classifiers, each trained with different training data, are combined. Lastly, the ensemble structure may be a fundamentally superior representation of the relationship between the independent and dependent variables, encoding relationships that cannot be represented by any single base classifier [19,28,29].

When we compared linear with nonlinear classifiers for this application, including feedforward artificial neural networks trained with a conjugate gradient algorithm [30], support vector machines trained with linear and radial basis kernels [31], and classification trees [32], we found that none of these alternatives offered better performance than a basic linear classifier [33]. This might be explained by our limited volume of training data, which hindered classifier optimization through cross validation. We also attempted to improve the ensemble classifier performance via bagging [34] and boosting [35]. However, they yielded the same overall performance, AUC of 0.76 (SD 0.05). We interpret this lack of improvement to mean that the ensemble classifiers are reasonably stable, so their performance is not improved by further resampling and aggregation [19].

Our basic justification for using ensemble classifiers is their ability to tolerate missing variables [36]. Although prior investigators have reported promising results using machine learning algorithms in predicting major hemorrhage [37], mortality in patients with head trauma [38], and trauma outcomes [39], such methods require the availability of all independent predictor variables, all the time. The ensemble classifier presented in this paper is free from such limitation, since it uses whichever variables are

available at any given time, thus significantly increasing the overall system availability.

Because we are using just the basic vital-signs (HR, RR, DBP, SBP, and SaO_2), we are able to use all combinations of one, two, or three vital-signs as inputs into the base classifiers that comprise the ensemble (there was no improvement in performance when we used more than three vital-signs per base classifier). In practice, when a subject is missing a vital-sign, we drop whichever base classifiers require the missing vital-sign as input, and the remaining base classifiers make up the "new" ensemble. The system is able to classify the patient as long as a single vital-sign variable is available. Using our Illustrative dataset, we quantify the effects of missing data (Table 3)—the slow degradation of performance as a function of increasing loss of data.

Missing data is a real problem in our Total dataset, collected in the real-world during prehospital patient transport, and we speculate that missing data will be a major problem during any physiologic monitoring in unstructured environments (home, battlefield, disaster scene, etc.). Applying to the Total dataset, we learn in practice how the ensemble classifier might perform: during the first 2 min of transport, where 26% of the variables are missing, the classifier yields a 0.70 AUC performance. The AUC rises through time, when there is less and less missing data (see Table 4). This validates that our ensemble classifier functions as intended: it classifies subjects with whatever information (or lack thereof) is available. This property should be an asset to any real-world application, when complete data availability cannot be taken for granted.

There are alternative techniques to handle missing data, such as imputation or expectation-maximization (EM) algorithms [40], which we decided are not ideal for our application. Imputation techniques require either availability of similar data (which, because of the limited size of the prehospital dataset, we do not have), or the availability of a probabilistic model for the data-generating mechanism, which may not be effective given the heterogeneities in our data (this heterogeneity is exemplified in Fig. 1). More importantly, imputation implicitly requires knowledge of the outcome class (to establish the extent of data similarity), which we do not know during real-time application. Ultimately, these factors motivated our selection of an ensemble of linear base classifiers. In the future, a direct comparison with alternative solutions for missing data may be warranted.

Table 4
The relationship between the ensemble classifier AUCs and the percentages of missing and reliable variables in the Total dataset as a function of time

Time interval (min)	Ensemble classifier AUC	% of missing variables in the Total dataset	% of reliable variables in the Total dataset
0–2	0.70	26	39
2–4	0.73	14	50
4–6	0.78	8	56
6–8	0.75	6	58
8–10	0.77	5	59
10–12	0.78	5	59
12–14	0.79	4	57
14–16	0.77	4	53
Correlation with AUC	1.00	-0.90 ($p = 0.002$)	0.87 ($p = 0.005$)

The percentages of missing and reliable variables are significantly correlated with AUC ($p < 0.05$).

Our classifier offers good, but not excellent, accuracy. We believe that this reflects a limitation to the information in the basic vitals signs. We have shown that alternatives to the ensemble, other linear and nonlinear classifiers, are not more effective. Also, as shown in Table 2, there is no apparent advantage when the variables are arranged into a composite structure (e.g., the shock index HR/SBP); classifiers containing the basic vital-signs seem to yield the same information. Beyond these basic vital-signs, there are other physiological measurements that have been shown to be diagnostically useful in trauma patients. These include cardiac index and transcutaneous oxygen tension indexed to the fractional inspired oxygen concentration [41], blood base excess [42], and heart rate variability [12]. Information may also exist in the temporal changes of the basic vital-sign data. It is possible that, in the future, our ensemble classifier could be improved by incorporation of additional physiologic variables along with the basic vital-sign variables.

The ensemble classifier performs better over time (Fig. 3). We have already speculated that this is largely a function of better data availability (Table 4). We do not think this is due to progressive changes in physiology. In a subset of 296 patients with two SBP measurements, one from the first 10 min and one from the subsequent 10 min, the ROC AUC is identical for both time periods (AUC = 0.73). In general, we do not find any evidence for progressive physiologic evolution to explain the rising AUC through time. It is also possible that the medics' therapies may be affecting the physiology through time. However, we do not believe this is a major factor either. When, in addition to the basic vital-signs, we input the volume of fluid resuscitation given to each subject into our classifiers, the AUC typically rises just +0.02 units. Based on this, we conclude that the volume of fluid resuscitation is a minor factor in the underlying relationship between vital-signs and major hemorrhage (moreover, volume replacement therapy would tend to *mask* the physiology of hemorrhage, rather than make it more evident through time).

This study focuses on the problem of missing data, but our results also address the matter of *unreliable* data. Use of 5 s of the "best-quality data" available (determined by our own automated algorithms) yields better classification performance than unqualified data. Using a simple 2-min average of all the data gives classification performance that is equivalent to using 5 s of qualified data. These findings suggest that, in a fieldable implementation of the classifier,

the collected data might be filtered by either averaging values over a certain time window, or, if time is of the essence, by automatically qualifying the data before input to the classifier. In addition, Table 4 shows that there is a positive correlation between percentage of reliable data and classifier AUC. This further indicates the strong dependency of improved classification on data quality.

5. Conclusions

It is possible to classify trauma patients into those that show physiological responses consistent with a cardiovascular hypovolemic state and those that are normovolemic. The classifier uses five basic vital-sign variables, HR, RR, DBP, SBP, and SaO_2, collected at 1-s intervals, since no advantage is apparent if the variables are arranged into a composite structure. The classifier is constructed from simple linear classifiers in an ensemble configuration, which is able to tolerate missing vital-sign data, and is more reliable than classifiers based on the identification of the "best" variable features. The classifier is robust enough to work with simple, unfiltered vital-sign data, but its performance can be marginally improved by using the best-quality data available or by accumulating mean vital-sign values over a longer time period. The ensemble classifier can be an important element in ongoing efforts to develop reliable, fast, and small devices to monitor a patient's physiologic state in real-time to provide caregivers with additional information to assist in the care of their charges.

6. 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 US Army or the US Department of Defense. This paper has been approved for public release with unlimited distribution.

Acknowledgments

This work was partially supported by the Combat Casualty Care Directorate of the US Army Medical Research and Materiel Command, Fort Detrick, Maryland. We are grateful to the University of Texas Health Science Center and to COL John Holcomb and Dr. Jose Salinas of the US Army Institute of Surgical Research for providing access to the Vital-Signs (Trauma) database.

References

[1] Schoenberg R, Sands DZ, Safran C. Making ICU alarms meaningful: a comparison of traditional vs. trend-based algorithms. Proc AMIA Symp 1999:379–83.
[2] Grossman P. The LifeShirt: a multi-function ambulatory system monitoring health, disease, and medical intervention in the real world. Stud Health Technol Inform 2004;108:133–41.
[3] Lovett PB, Buchwald JM, Sturmann K, Bijur P. The vexatious vital: neither clinical measurements by nurses nor an electronic monitor provides accurate measurements of respiratory rate in triage. Ann Emerg Med 2005;45:68–76.
[4] Friesdorf W, Konichezky S, Gross-Alltag F, Fattroth A, Schwilk B. Data quality of bedside monitoring in an intensive care unit. Int J Clin Monit Comput 1994;11:123–8.
[5] Yu C, Liu Z, McKenna T, Reisner AT, Reifman J. A method for automatic identification of reliable heart rates calculated from ECG and PPG waveforms. J Am Med Inform Assoc 2006;13(3):309–20.
[6] Chen L, McKenna T, Reisner A, Reifman J. Algorithms to qualify respiratory data collected during the transport of trauma patients. Physiol Meas 2006;27(9):797–816.
[7] Hunt RC, Bryan DM, Brinkley VS, Whitley TW, Benson NH. Inability to assess breath sounds during air medical transport by helicopter. JAMA 1991;265(15):1982–4.
[8] Hoyert DL, Mathews TJ, Menacker F, Strobino DM, Guyer B. Annual summary of vital statistics: 2004. Pediatrics 2006;117(1):168–83.
[9] Sauaia A, Moore FA, Moore EE, Moser KS, Brennan R, Read RA, Pons PT. Epidemiology of trauma deaths: a reassessment. J Trauma 1995;38(2):185–93.
[10] Peng R, Chang C, Gilmore D, Bongard F. Epidemiology of immediate and early trauma deaths at an urban Level I trauma center. Am Surg 1998;64(10):950–4.
[11] Montgomery SP, Swiecki CW, Shriver CD. The evaluation of casualties from Operation Iraqi Freedom on return to the continental United States from March to June 2003. J Am Coll Surg 2005;201(1):7–12. discussion 12-3.
[12] Cooke WH, Salinas J, Convertino VA, Ludwig DA, Hinds D, Duke JH, et al. Heart rate variability and its association with mortality in prehospital trauma patients. J Trauma 2006;60(2):363–70. [discussion, p. 370].
[13] Holcomb JB, Salinas J, McManus JM, Miller CC, Cooke WH, Convertino VA. Manual vital signs reliably predict need for lifesaving interventions in trauma patients. J Trauma 2005;59(4):821–8. [discussion, p. 828–829].
[14] Propaq Encore Reference Guide. Welch Allyn Inc. Beaverton, OR, 1998. Available from: http://www.monitoring.welchallyn.com/products/portable/propaqencore.asp.
[15] Holcomb JB, Niles SE, Miller CC, Hinds D, Duke JH, Moore FA. Prehospital physiologic data and lifesaving interventions in trauma patients. Mil Med 2005;170(1):7–13.
[16] Webb A. Statistical pattern recognition. New York: Oxford University Press; 1999.
[17] Obuchowski NA. Receiver operating characteristic curves and their use in radiology. Radiology 2003;229(1):3–8.
[18] Guyon I, Elisseeff A. An introduction to variable and feature selection. In: Guyon I, Elisseeff A, editors. Special issue on variable and feature selection. J Mach Learn Res 2003;3:1157–82.
[19] Dietterich TG. Ensemble methods in machine learning. MCS '00: Proceedings of the First International Workshop on Multiple Classifier Systems. Lect Notes Comput Sci 2000;1857:1–15.
[20] Kuncheva LI. Theoretical study on six classifier fusion strategies. IEEE Trans Pattern Anal Mach Intell 2002;24(2):281–6.
[21] Kittler J, Alkoot FM. Sum versus vote fusion in multiple classifier systems. IEEE Trans Pattern Anal Mach Intell 2003;25(1):110–5.
[22] Wilcoxon F. Individual comparisons by ranking methods. Biometrics 1945;1:80–3.
[23] Birkhahn RH, Gaeta TJ, Van Deusen SK, Tloczkowski J. The ability of traditional vital signs and shock index to identify ruptured ectopic pregnancy. Am J Obstet Gynecol 2003;189(5):1293–6.
[24] Pestel GJ, Hiltebrand LB, Fukui K, Cohen D, Hager H, Kurz AM. Assessing intravascular volume by difference in pulse pressure in pigs submitted to graded hemorrhage. Shock 2006;26(4):391–5.
[25] Youden WJ. Index for rating diagnostic tests. Cancer 1950;3(1):32–5.
[26] Perkins NJ, Schisterman EF. The inconsistency of "optimal" cutpoints obtained using two criteria based on the receiver operating

characteristic curve. Am J Epidemiol 2006;163(7):670–5. [Epub 2006 Jan 12].
[27] Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143(1):29–36.
[28] Kittler J, Hatef M, Duin RP, Matas J. On combining classifiers. IEEE Trans Pattern Anal Mach Intell 1998;20(3):226–39.
[29] Xu L, Krzyzak A, Suen CY. Methods of combining multiple classifiers and their applications to handwriting recognition. IEEE Trans Syst Man Cybern 1992;22(3):418–35.
[30] Haykin S. Neural networks a comprehensive foundation. 2nd ed. New Jersey: Prentice-Hall, Inc; 1999.
[31] Chang CC, Lin CJ. LIBSVM: a library for support vector machines 2001. Available from: http://www.csie.ntu.edu.tw/~cjlin/libsvm.
[32] Matlab® 7.1.0. Tree-based model for classification. The MathWorks, Inc. 1984–2005.
[33] Chen L, Reisner A, McKenna T, Gribok A, Reifman J. Diagnosis of hemorrhage in a prehospital trauma population using linear and nonlinear multiparameter analysis of vital signs. 29th IEEE EMBS Ann Int Conf, Lyon, France 2007, p. 2748–51.
[34] Breiman L. Bagging predictors. Mach Learn 1996;24(2):123–40.
[35] Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. J Comput Syst Sci 1997;55(1):119–39.
[36] Mohammed HS, Stepenosky N, Polikar R. An ensemble technique to handle missing data from sensors. SAS 2006—IEEE Sensors Appl Symp (SAS). Houston, Texas USA, 2006, p. 101–5. 10.1109/SAS.2006.1634246.
[37] Blackmore CC, Cummings P, Jurkovich GJ, et al. Predicting major hemorrhage in patients with pelvic fracture. J Trauma 2006;61(2):346–52.
[38] Eftekhar B, Mohammad K, Ardebili HE, Ghodsi M, Ketabchi E. Comparison of artificial neural network and logistic regression models for prediction of mortality in head trauma based on initial clinical data. BMC Med Inform Decis Mak 2005;5(1):3.
[39] DiRusso SM, Sullivan T, Holly C, Cuff S, Savino J. An artificial neural network as a model for prediction of survival in trauma patients: validation for a regional trauma area. J Trauma 2000;49(2):212–23.
[40] Molenberghs G, Verbeke G. Multiple imputation and the expectation-maximization algorithm. In: Models for discrete longitudinal data. Springer series in statistics. New York: Springer Verlag; 2005. p. 511–29.
[41] Lu KJ, Chien LC, Wo CC, Demetriades D, Shoemaker WC. Hemodynamic patterns of blunt and penetrating injuries. J Am Coll Surg 2006;203(6):899–907. [Epub 2006 Oct 2].
[42] Siegel JH, Rivkind AI, Dalal S, Goodarzi S. Early physiologic predictors of injury severity and death in blunt multiple trauma. Arch Surg 1990;125(4):498–508.