Journal of Critical Care (2012) **27**, 533.e1–533.e10

ELSEVIER

Journal of Critical Care

The association between vital signs and major hemorrhagic injury is significantly improved after controlling for sources of measurement variability☆

Andrew T. Reisner MD[a,b,*], Liangyou Chen PhD[a], Jaques Reifman PhD[a,*]

[a]*Bioinformatics Cell, Telemedicine and Advanced Technology Research Center (TATRC), US Army Medical Research and Materiel Command (USAMRMC), ATTN: MCMR-TT, 504 Scott Street, Fort Detrick, MD 21702, USA*
[b]*Massachusetts General Hospital Department of Emergency Medicine, Boston, MA 02114, USA*

Keywords:
Vital signs;
Decision support systems;
Clinical;
Hemorrhage;
Data interpretation;
Statistical;
Emergency care;
Prehospital;
Trauma

Abstract

Purpose: Measurement error and transient variability affect vital signs. These issues are inconsistently considered in published reports and clinical practice. We investigated the association between major hemorrhagic injury and vital signs, successively applying analytic techniques that excluded unreliable measurements, reduced transient variation, and then controlled for ambiguity in individual vital signs through multivariate analysis.

Methods: Vital sign data from 671 adult prehospital trauma patients were analyzed retrospectively. Computer algorithms were used to identify and exclude unreliable data and to apply time averaging. An ensemble classifier was developed and tested by cross-validation. Primary outcome was hemorrhagic injury plus red cell transfusion. Areas under receiver operating characteristic curves (ROC AUCs) were compared by the test of DeLong et al.

Results: Of initial vital signs, systolic blood pressure (BP) had the highest ROC AUC of 0.71 (95% confidence interval, 0.64-0.78). The ROC AUCs improved after excluding unreliable data, significantly for heart rate and respiratory rate but not significantly for BP. Time averaging to reduce temporal variability further increased AUCs, significantly for BP and not significantly for heart rate and respiratory rate. The ensemble classifier yielded a final ROC AUC of 0.84 (95% confidence interval, 0.80-0.89) in cross-validation.

Conclusions: Techniques to reduce variability in vital sign data can lead to significantly improved diagnostic performance. Failure to consider such variability could significantly reduce clinical effectiveness or confound research investigations.

☆ Competing interests: Dr Andrew Reisner participated in a customer advisory meeting for General Electric Healthcare in 2008 and has received speaking fees from Masimo Corp. Dr Reisner is a coinvestigator on a US National Institutes of Health Bioengineering Research Partnership that includes Philips Healthcare.

* Corresponding author. Tel.: +1 301 619 7915; fax: +1 301 619 1983.
E-mail addresses: areisner@partners.org (A.T. Reisner), lchen@bioanalysis.org (L. Chen), jaques.reifman@us.army.mil (J. Reifman).

0883-9441/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jcrc.2012.01.006

1. Introduction

Vital signs measurement is a routine aspect of clinical practice and research protocols. Although it is known that transient variability and measurement error can, in principle, affect the accuracy of vital signs, what is unknown is the extent to which these factors affect diagnostic capabilities in actual clinical practice. Vital signs fluctuate through time because of transient perturbations (eg, medication boluses, bouts of pain, anxiety, coughing) as well as natural steady-state variability. In addition, the accuracy of vital sign data is affected by clinicians' technique [1]. For example, accurate blood pressure (BP) measurement using a cuff requires proper fit and positioning of the cuff, a relaxed and properly positioned extremity, and the absence of patient motion [2]. Significant discrepancies have been reported between different methods of measuring noninvasive BP [3]. Similarly, respiratory rate (RR) measurement is prone to technical error, whether measured by a clinician [4] or by a bedside monitor via impedance pneumography (IP) [5]. In one report, both triage nurses' measurements of RR and electronic measurement of RR revealed poor sensitivity for bradypnea and tachypnea, and the authors referred to RR as "the vexatious vital"[4]. Heart rate (HR) monitored by electrocardiography (ECG) can be unreliable, that is, if electrodes are improperly affixed, and false arrhythmia alarms are commonplace [6]. Multiple authors have called into question the value of HR in assessing the hemodynamic state of a patient because of its variable relationship with hypovolemia [7,8]. Finally, it is worth noting that the accuracy of vital sign data may vary considerably for different makes and models of measurement devices [9-11].

The extent to which these factors affect diagnostic capabilities in actual practice is relevant to the design and interpretation of clinical investigations. If vital sign data were often polluted by inaccuracies, then there would be a bias toward the null hypothesis, where positive study effects might be masked (ie, type II study errors). Alternatively, failure to describe key methodology that improved vital sign accuracy (eg, superior equipment, training, or study protocols) would make it harder for others to replicate a successful study. Consider that some reports support the usefulness of prehospital severity scores for trauma patients [12-14], whereas other studies found those scores ineffective [15,16]. In these examples, the reports lacked any explicit consideration of the measurement apparatus, clinical protocols, and quality assurance processes related to vital sign measurements; and inconsistency in how vital signs were measured could have contributed to the heterogeneous findings. More broadly, there are diverse sets of conflicting reports with a shared failure to detail vital sign measurement methodology, for example, the risk of volume resuscitation of trauma patients with uncontrolled hemorrhage [17,18], the benefit of rapid response teams for inpatients with physiologic deterioration [19-22], and the benefit of early goal-directed resuscitation for septic shock [23]. It is possible that different approaches to vital sign measurements contributed to the inconsistencies of the reports' findings.

We investigated the association between standard vital signs and major hemorrhagic injury in a population of prehospital trauma patients using computational techniques that excluded unreliable measurements, reduced transient perturbations, and reduced ambiguity of individual vital signs. We compared these results with conventional analyses. The findings are applicable to the clinical evaluation of hemorrhage, which is the single most treatable cause of mortality in trauma patients [24,25]. Moreover, the findings may relate to a range of applications because the extent to which different analytic methods yield significantly different results indicates the importance of considering these factors in clinical practice and research studies.

2. Materials and methods

2.1. Clinical data collection

This was a retrospective analysis of a database, originally collected and analyzed by Cooke et al [26] with institutional review board approval, of trauma patients during transport by air ambulance from the scene of injury to a level I trauma center [26]. Between August 2001 and April 2004, the following physiologic data were measured in a convenience sample by Propaq 206EL monitors (Protocol Systems, Beaverton, Ore) and archived using a networked personal digital assistant: ECG and IP recorded at 182 and 23 Hz, respectively; the corresponding HR and RR output at 1-second intervals; and systolic BP (SBP) and diastolic BP (DBP) measured intermittently at multiminute intervals. Clinical data were collected during retrospective chart review, including demographics, prehospital interventions, hospital treatments, and injury descriptions. Subsequently, vital sign data from 788 patients were uploaded to our data warehousing system [27]. Protected health information was not included.

All data analysis was performed using MATLAB v7 (MathWorks, Natick, Mass).

2.2. Vital sign reliability

For each vital sign value, reliable data were identified by automated algorithms that rated each datum on an integer scale of 0 to 3 from least reliable to most reliable. Vital sign data rated 2 or 3 were considered reliable; otherwise, they were unreliable. Detailed descriptions of these algorithms have been previously reported [28-30]. Here, we provide an overview of the methodology. The algorithms analyze moving windows of physiologic data. The algorithms rate the reliability of vital signs computed

from the data windows based on (1) a computerized assessment of the ECG or IP waveforms' reliability and (2) a comparison between the rates output by the Propaq 206EL vs an independent calculation of the HR or RR performed by the algorithm. In practice, when waveforms demonstrate clear, rhythmic beats or breaths and the rates output by the Propaq 206EL match the algorithms' own calculations, then the corresponding HR or RR is rated as reliable. Conversely, when the waveforms are noisy with irregular, heterogeneous beats or breaths and/or there were major discrepancies between the rates output by the Propaq 206EL vs the algorithms' own calculations, then the HR or RR is rated as unreliable. The underlying rationale is the assumption that clean ECG or IP waveforms lead to reliable HR or RR measurements and that HR or RR tends to be reliable when 2 independent calculation methods yield similar results.

In prior validation, the reliability rating of RR using the automated algorithms typically concurred with clinicians who independently applied the reliability criteria to a set of test cases [28,30]. In 99% of the test cases, the automated algorithm agreed with the clinician RR rating (±1 level), where high RR reliability ratings were found to be associated with smaller differences between computer-calculated and human-calculated RR (average differences of 1.7 and 8.1 breaths per minute for the best and worst RR reliability ratings, respectively). Likewise, there was close agreement (within ±5 beats per minute) between computer-calculated and human-calculated HR in 97% of the test cases rated 2 or 3 by the automated HR reliability algorithm [30].

The BP reliability algorithm determined if the ratio between SBP, DBP, and mean pressure is physiologic and if the HR measured by the inflatable oscillometric cuff matches the ECG HR [29]. The algorithm does not attempt to distinguish between unequal HRs because of motion artifact vs unequal HRs because of nonperfusing electrical beats, for example, premature contractions; in the latter case, it would be possible for reliable BP data to be misclassified as unreliable.

2.3. Subject selection

The primary study population consisted of patients with any reliable vital sign datum within the initial 15 minutes of prehospital monitoring. We also studied 3 subgroups: patients with pairs of at least 1 reliable and 1 unreliable (*a*) HR, (*b*) RR, and (*c*) BP. In the primary analysis, we excluded the "ambiguous outcome" patients who received red blood cell (RBC) transfusions but lacked documented injuries that were indisputably hemorrhagic (see below). These cases were reincluded in sensitivity analyses (see below). Also excluded were the few patients who died before any diagnostic imaging or surgical exploration, when it could not be determined whether the patient died to major hemorrhage vs other critical pathology.

2.4. Primary outcome

Major hemorrhagic injury was defined as a documented injury that unequivocally causes some loss of blood volume (laceration or fracture of a solid organ, thoracic or abdominal hematoma, vascular injury that required operative repair, or limb amputation) and RBC transfusion within 24 hours.

2.5. Comparison of reliable vs unreliable vital signs

We computed the patients' proportions of reliable vital signs (median and interquartile range). For the 3 subgroups with at least 1 reliable and 1 unreliable vital sign—HR, RR, and BP—we computed each patient's mean of the reliable and of the unreliable data and compared the population mean of the subjects' means with Student *t* test for paired data (note that the *t* test is valid for normal and nonnormal distributions as long as there are enough subjects per distribution, eg, 30 or more [31]).

To compare diagnostic performance, we repeated the following statistical computation 100 times for each vital sign: from each patient, we randomly selected 1 reliable and 1 unreliable measurement, then computed receiver operating characteristic (ROC) curves for the selected reliable and the unreliable data using the method of DeLong et al [32]. We computed the difference between the areas under those curves (ROC AUCs) and averaged the results from the 100 cycles. This methodology avoided biases due to those patients with a surplus of measurements and unequal ratios of reliable vs unreliable measurements between patients.

2.6. Association between vital signs and major hemorrhagic injury within the initial 15 minutes

For each vital sign, we computed the univariate ROC AUC for (*a*) the first nonzero value, (*b*) the first reliable value, (*c*) the last reliable value, and (*d*) the average of all reliable values within 15 minutes.

We performed multivariate analysis using ensemble classification, a collection of multivariate regression models. Each of the models within the ensemble is a standard linear regression model, and their outputs are simply averaged to yield the ensemble classifier output [33]. Ensemble classification is able to classify subjects with incomplete data, as is explained below. This property was important because many patients lacked reliable data for every vital sign.

Each regression model within the ensemble used 1, 2, or 3 of the following parameters: HR, RR, SBP, and SBP − DBP. The final ensemble was composed of all possible combinations (14 total regression models). We applied cross-validation, randomly partitioning 50% of the study population for classifier training. Each model was trained using the subset of patients who possessed at least 1 reliable measurement of each model parameter within the initial 15 minutes, using the average of all reliable values from the initial 15 minutes. Next,

Table 1 Population description

Characteristic	Study population
Population size, n	671
Male/female, n [a]	498/172
Age (y), mean (SD)	38 (15)
Blunt injury, n (%)	596 (89)
Mortality, n (%)	41 (6)
Prehospital intubation, n (%)	115 (17)
Major hemorrhagic injury, n (%)	78 (12)
% Reliable HR for patient, median (IQR)	62 (4-84)
% Reliable RR for patient, median (IQR)	16 (0-45)
% Reliable SBP for patient, median (IQR)	100 (83-100)

Patients with at least 1 reliable vital sign datum within 15 minutes after exclusion of cases who received RBC transfusions but lacked documented injuries that were indisputably hemorrhagic (see text for details). IQR indicates interquartile range.

[a] Sex unknown for 1 patient in the database.

we tested the ensemble classifier in the remaining 50% of the patients. For each patient, we only used those regression models for which the patient had the necessary reliable data during the initial 15 minutes and used the models' average output as the final output. This process was repeated for 100 cycles, each time randomly repartitioning the patients into training/testing sets. We computed the mean ROC AUC of those 100 testing cycles.

2.7. Sensitivity analyses

We repeated the ensemble classification using 4 alternative methodologies: (*a*) reinclusion of the "ambiguous outcome" patients, treating them as nonhemorrhage control cases; (*b*) redefining "major hemorrhagic injury" as a documented hemorrhagic injury, as above, plus RBC transfusion *or* at least 3 L of crystalloid infusion; (*c*) redefining "major hemorrhagic injury" as the receipt of at least 5 U of RBC regardless of the documented injuries; and (*d*) using reliable vital sign data from only the initial 10 minutes.

3. Results

The database had 788 records with at least 1 nonzero vital sign datum. One hundred seventeen cases were excluded (105 were "ambiguous outcome" cases subsequently reintroduced in the sensitivity analysis described below, whereas 12 lacked any reliable vital sign data). Table 1 shows the population characteristics, with 12% having major hemorrhagic injury, 17% with prehospital intubation, and 6% overall mortality. Respiratory rate data had the lowest rate of reliability, whereas BP data had the highest.

Table 2 shows reliable data compared with unreliable data. Unreliable measurements of HR, RR, and SBP all had significantly elevated values vs their reliable counterparts and tended to have reduced ROC AUCs.

Table 3 reports the cumulative diagnostic yields of the investigative techniques. The ROC AUCs were significantly improved for initial HR and initial RR when reliability was considered. The ROC AUCs were significantly improved for SBP when the average of all its reliable values was used, whereas these were nonsignificantly increased for the average of reliable HR or RR. (In regard to the effects of mechanical ventilation on RR, the average of all reliable RR yielded an ROC AUC of 0.72 [95% confidence interval {CI}, 0.62-0.80] for spontaneously breathing patients and 0.64 [95% CI, 0.46-0.78] for mechanically ventilated patients.)

Applied to all 671 patients in the study population, the ensemble classifier yielded an ROC AUC of 0.84 (95% CI, 0.80-0.89) in cross-validation. This AUC was significantly greater than any univariate vital sign. The classifier could identify 36% of major hemorrhagic injury cases with greater

Table 2 Reliable compared with unreliable vital signs

Vital sign	Population with at least 1 reliable and 1 unreliable vital sign, n	Patients with major hemorrhagic injury, n (%)	Patients' proportion of reliable data (%), median (IQR)	Reliable data, mean (SD)	Unreliable data, mean (SD)	Reliable vs unreliable data, *P* value (Student *t* test)	Unreliable vital signs: ΔROC AUC for Dx of major hemorrhagic injury, mean (upper/lower range)
HR	632	72 (11)	65 (7-85)	95 (20)	99 (20)	<.001 [a]	−0.02 (−0.05/+ 0.01)
RR	388	52 (13)	39 (20-61)	27 (7)	37 (17)	<.001 [a]	−0.11 (−0.18/−0.03)
SBP	217	34 (16)	75 (67-86)	127 (22)	138 (37)	<.001 [b]	−0.12 (−0.21/−0.03)
DBP	221	34 (15)	75 (67-86)	72 (15)	76 (75)	NS	−0.02 (−0.09/+ 0.04)

Populations included only those patients determined to have at least 1 reliable and 1 unreliable vital sign measurement, according to the reliability algorithms, at any time during their transport. Shown are the patients' means of reliable vs unreliable data for all patients (computing first the mean of each patient and then computing the mean of the patients' means). Student *t* test for paired data was used to test for significant differences between patients' means. Finally, the change in ROC AUC in the diagnosis of major hemorrhagic injury is shown, when one random unreliable measurement was used in place of a random reliable measurement (see text for details of this calculation). NS indicates not significant. Dx indicates diagnosis.

[a] Reliable vs unreliable data are also significant ($P < .001$) in hemorrhage cases alone and in control cases alone.

[b] Reliable vs unreliable data are also significant in hemorrhage cases alone ($P = .01$) and in control cases alone ($P < .001$).

Table 3 Areas under receiver operating characteristic curves for the diagnosis of major hemorrhagic injury with application of vital sign reliability criteria, time averaging, and multivariate (ensemble) classification

Vital sign	Population	ROC AUC (95% CI)			
		First nonzero	First reliable	Last reliable	All reliable
HR	At least 1 reliable HR (n = 625)	0.60 (0.53-0.68)	0.71 (0.63-0.77)[a]	0.72 (0.65-0.78)[a]	0.73 (0.65-0.79)[a]
RR	At least 1 reliable RR and intubated (n = 85)	0.52 (0.46-0.58)	0.64 (0.55-0.72)[a]	0.63 (0.53-0.71)	0.67 (0.58-0.75)[a]
RR	At least 1 reliable RR and spontaneous breathing (n = 313)	0.55 (0.48-0.61)	0.64 (0.53-0.74)	0.68 (0.56-0.77)[a]	0.72 (0.62-0.80)[a]
SBP	At least 1 reliable SBP (n = 648)	0.71 (0.64-0.78)	0.74 (0.67-0.80)	0.77 (0.70-0.83)	0.79 (0.73-0.85)[a,b]
DBP	At least 1 reliable DBP (n = 648)	0.62 (0.54-0.69)	0.64 (0.56-0.71)	0.64 (0.56-0.71)	0.63 (0.55-0.71)
Ensemble classifier	At least 1 reliable HR or reliable RR or reliable SBP (n = 671)	NA	NA	NA	0.84 (0.80-0.89)[c]

Ensemble classification was applied to the overall study population. For RR, results are also provided separately for intubated patients and for spontaneously breathing patients. The method of DeLong [32] for paired data was used to test for statistically significant differences of ROC AUCs. NA indicates not applicable.

[a] ROC AUC significantly ($P < .05$) increased vs ROC AUC for "first nonzero" value.

[b] ROC AUC significantly ($P < .05$) increased vs ROC AUC for "first reliable" data.

[c] Ensemble ROC AUC significantly increased vs ROC AUC for "all reliable" HR data ($P < .001$), "all reliable" RR data ($P < .001$), "all reliable" SBP data ($P < .05$), and "all reliable" DBP data ($P < .001$).

than 60% positive predictive value (PPV) and greater than 85% of hemorrhage cases with 24% PPV (Fig. 1).

The sensitivity analyses yielded the following ROC AUCs for major hemorrhagic injury, which were similar to the primary analysis: (*a*) inclusion of the ambiguous outcome patients, 0.82 (95% CI, 0.77-0.87); (*b*) use of RBC transfusion or at least 3 L of crystalloid infusion as the outcome, 0.83 (95% CI, 0.79-0.87); (*c*) inclusion of

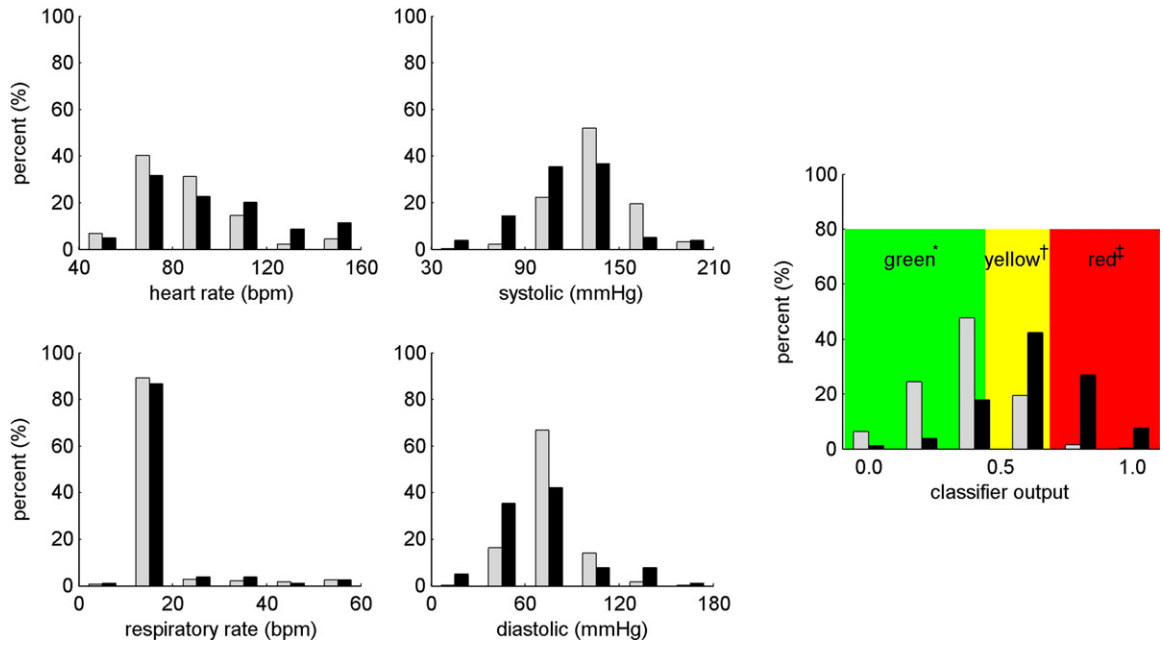

Fig. 1 Histograms of basic vital signs and of the multivariate ensemble classifier for major hemorrhagic injury cases vs control cases. Histograms for each basic vital sign (HR, RR, SBP, and DBP) using the first nonzero value and the output of the multivariate ensemble classifier (using cross-validation with distinct training/testing data; see text for details). Patient populations for each histogram correspond to the populations in Table 3, whereas multivariate ensemble classification was applied to the entire study population. Right: Ensemble output (testing data) averaged from 100 iterations of cross-validation. Using one cutoff, ensemble classification yielded a sensitivity of greater than 85% and a PPV of 24%; patients below this threshold lie in the green background field. Using an alternative cutoff, ensemble classification offered a sensitivity of 36% and a PPV of greater than 60%; patients above this threshold lie in the red background field. *Green zone: 383 control cases and 11 major hemorrhagic injury cases; †yellow zone: 192 control cases and 39 major hemorrhagic injury cases; ‡red zone: 18 control cases and 28 major hemorrhagic injury cases.

ambiguous outcome patients and changing the outcome to the receipt of at least 5 U of RBC, 0.81 (95% CI, 0.76-0.86); and (*d*) use of only the initial 10 minutes, 0.81 (95% CI, 0.76-0.86).

4. Discussion

We found that accounting for measurement error and physiologic variability can significantly improve the association between vital signs and major hemorrhagic injury. Vital signs may be more informative about a trauma patient's circulatory state than previously appreciated in reports that did not explicitly consider these factors [26,34-36]. Moreover, these findings may inform the design and interpretation of a range of clinical trials that involve vital signs and how vital signs are used in clinical practice. The implications are cautionary, suggesting that such factors are important to consider. At the same time, these findings also suggest potential solutions.

The computational techniques used in this analysis have been previously described [28-30,33,37,38]. Here, the techniques were integrated to determine their cumulative effects in a population of trauma patients. These techniques significantly improved the association of vital signs and major hemorrhagic injury without the need for consideration of the patients' baseline vital signs, administration of medications, anatomical location of the injury, age, or mechanism of injury. Applied cumulatively, diagnostic performance exceeded prior reports on the individual techniques [33]. The vital sign patterns correctly classified by these techniques were not always self-evident by eye (eg, Fig. 2).

4.1. Clinical implications

We have shown that reliable vital sign data have a significantly higher association with a life-threatening pathophysiology, even as unreliable measurements were commonplace (Table 1). These findings support the adherence to proper vital sign measurement techniques; even better than excluding unreliable data, as was done in this retrospective study, would be reducing unreliable measurements in the first place. When procuring monitoring apparatus, it would be desirable to prioritize makes and models that possess maximum accuracy [11-13]. In addition, the study implies a potential benefit to continuing training of clinical staff to enhance the diagnostic value of vital sign

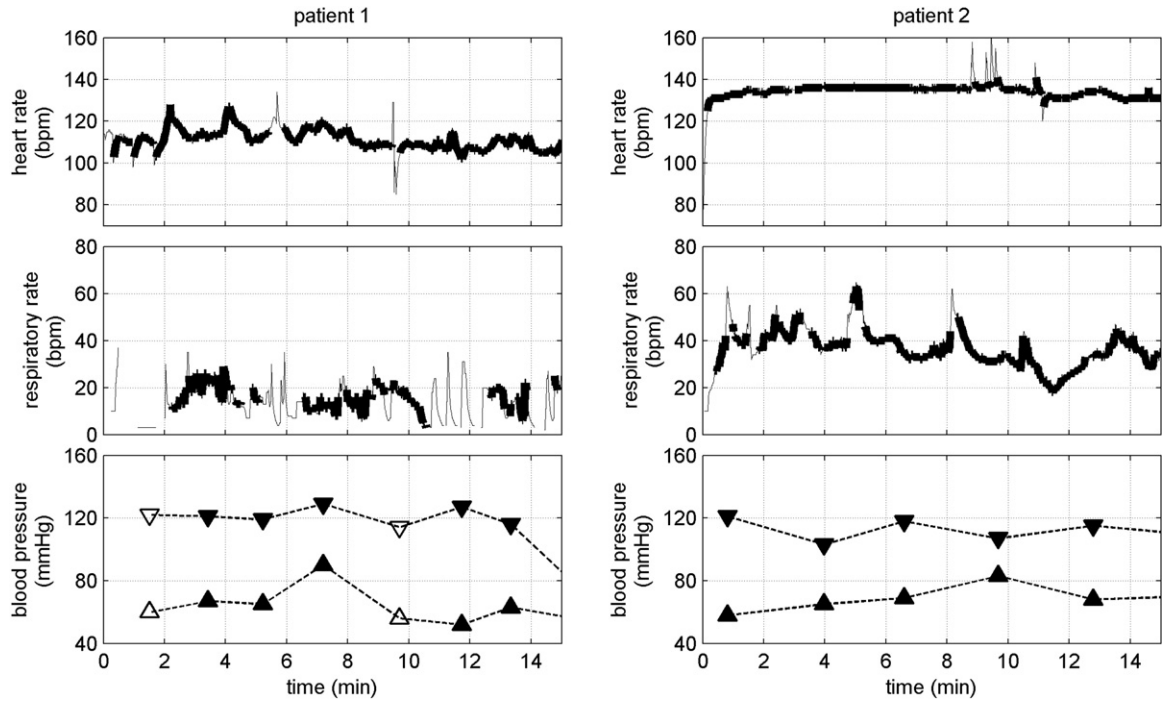

Fig. 2 Case examples for which the presence or absence of major hemorrhagic injuries can be identified by patterns in the vital signs. Both cases had HRs of more than 120 beats per minute and normotension. In patient 1, the ensemble multivariate classifier—which weighs the HR, RR, and systolic and pulse pressures—indicated that major hemorrhagic injury was unlikely (ie, the classifier output lay in the low-risk green zone, with a 97% negative predictive value; see Fig. 1). Patient 1 did not require RBC transfusion and was diagnosed with a cerebral contusion and a neck injury without major vascular involvement. In patient 2, the ensemble multivariate classifier indicated that major hemorrhagic injury was probable (ie, the classifier output lay in the high-risk red zone, with a 60% PPV; see Fig. 1). Patient 2 had a grade III liver laceration; a fractured, disrupted pelvis; and a femur fracture and received 3 U of RBC. Thin lines and open triangles indicate unreliable data according to the automated algorithms; thick lines and solid triangles indicate reliable data.

measurement. Sources of unreliable vital sign data include poor electrode placement (eg, chest hair causing poor skin adhesion), excessive patient movement, and poor placement of BP cuffs. It would be truly revealing to study, prospectively, which sources of error are the most problematic and whether the association between vital signs and pathology can be enhanced through focused training.

Certain techniques suggested by this report might be applied at the bedside to assess the state of the casualty. For example, when patients arrive at the hospital, clinicians expecting obvious vital sign trends might be misled because we have found that transient perturbations may mask the underlying trends and that measurements made at the end of transport are not necessarily more useful than the preceding prehospital measurements. Shapiro et al [39] and Lipsky et al [40] reported that, among patients who arrived normotensive in the emergency department, one or more episodes of preceding hypotension were associated with higher acuity. Our findings suggest that, in addition to the most recent measurements, clinicians should consider the time average of recent data, which we have shown can be significantly more diagnostic.

This raises the question of what duration of time window is optimal for computation of average values of recent vital sign data, for example, 5, 15, or 60 minutes. The goal of the time averaging is to filter out transient perturbations; but if the time window gets too large, then time averaging can actually obscure trends developing in later data. Therefore, it is important that the time window should not be too large. We speculate that averaging over more than 15 minutes may not be diagnostically optimal, but this is difficult to answer definitively with the current data set because the records are of such heterogeneous duration.

Simultaneous consideration of multiple vital signs can also improve the value of the data. For instance, low BP could represent significant blood loss, the patient's normal baseline, or reduced adrenergic tone. Tachycardia and tachypnea suggest the former, normal rate and respiration suggest baseline physiology, and bradycardia and bradypnea suggest sympatholysis. Clinicians may be unable to mentally compute a multivariate statistical model; but a simple multivariate metric, such as the shock index (the ratio of HR and SBP [41,42]), can be applied at the bedside.

4.2. Research implications

We demonstrated that accounting for sources of measurement variability can yield significantly different results when analyzing vital sign data. Accordingly, we recommend the following steps for clinical research involving vital sign data: (*a*) report the make and model of any monitoring equipment used and, when available, provide accuracy citations [12,13,43]; (*b*) report relevant in-service training, or its absence, of the clinical staff; (*c*) keep the measurement environment as consistent as possible to reduce transient variability, or else use the average of several measurements; and (*d*) consider the use of validated clinical scores or propensity scores to supplement or replace individual vital signs.

In addition, we note that there has been academic interest in novel types of physiologic sensors intended to improve patient monitoring. The cost and effort necessary to adopt new sensor modalities might be weighed against the findings in this report, which are that standard vital signs can be significantly improved through application of some simple techniques. Academically, we suggest that new monitoring modalities should be directly compared against conventional monitoring, with consideration given to the sources of variability highlighted here.

4.3. Specific findings

Systolic BP was the best univariate predictor. We [37] and others [44,45] have previously found that prehospital trauma patients demonstrate substantial temporal variability. We reduced the effects of transient perturbations by using the time average of serial vital sign measurements, which yielded significantly higher ROC AUCs for SBP, higher than either the initial or final prehospital SBP. Diastolic BP alone was a weak predictor; but we found that it provides additional information independent of SBP because it is useful to compute pulse pressure, the difference between SBP and DBP [33]. In spontaneously breathing patients, reliable RR was a useful predictor of hemorrhage. This finding was anticipated by classic physiologic reports that demonstrated that blood flow to the carotid body chemoreceptors is reduced in early hemorrhage because of compensatory vasoconstriction. "Stagnant" hypoxia then develops in the chemoreceptors, triggering an increased respiratory drive and tachypnea [46-49]. Interestingly, this RR reliability algorithm was not originally developed to diagnose major hemorrhagic injury per se, but to identify intervals in the IP that matched clinicians' opinions that the respiratory waveform was rhythmic and consistent [28]. Used as a diagnostic tool, we found that reliable RR data were significantly more diagnostic than unreliable RR. We observed that unreliable RR was often falsely elevated (ie, biased) because of motion artifacts in the pneumogram that were incorrectly counted as additional breaths.

Only a subset of patients (59%) had a complete set of reliable vital signs within 15 minutes. This was consistent with prior reports that unreliable vital sign data are all too typical in clinical practice [1,2,4-6]. To deal with missing data, we used an ensemble classifier for multivariate classification, which was significantly better than univariate classification. In a prior report, the ensemble classifier was applied to a moving 2-minute window of vital sign data [33]. That approach was not as successful because, in any given 2-minute window, there was an exaggerated proportion of missing data and there was major minute-to-minute variability that, here, we successfully filtered out by time averaging over 15 minutes (see above). In addition, the

current ensemble uses pulse pressure instead of DBP and does not incorporate oxygen saturation, thus excluding weak univariate predictors.

4.4. Automated diagnostic algorithms

It is technically feasible to run this investigation's analysis algorithms in real time, automatically distinguishing between normovolemic vs hemorrhagic vital sign patterns. We speculate that such automated, continuous analysis could improve the quality and safety of any monitored patient, especially when the clinical staff is distracted or inexperienced. In addition, protocols for triage or resuscitation could be considered using the algorithm's output as a starting point that may be more clinically valid than any sole vital sign. Lastly, in some cases, the algorithm could enhance the judgment of the clinician (eg, cases such as in Fig. 2). Similar types of automated analysis of vital sign data may likewise prove useful for other clinical applications, such as early detection of acute deterioration of hospital ward patients [50].

4.5. Limitations

There are several factors to consider in terms of the internal validity of this study. First, there is no gold standard definition to retrospectively distinguish true hemorrhagic injury vs minor (or non) hemorrhagic injures. We therefore analyzed several alternative outcome definitions. The similar results, regardless of the specific definition, suggest that the findings were not an artifact of the outcome definition but will be similar given any reasonable definition of hemorrhagic injury (note that our database did not contain parameters such as base deficit and pH). Our findings would be further strengthened if future investigations demonstrate comparable findings given additional end points and pathologic processes.

As a second limitation, the present findings depended on our algorithms to identify reliable vital signs; and the results might be different with different algorithms. However, in developing these algorithms, we found that most analytic methodologies that we explored yielded similar results because, in practice, the different algorithms only differed about borderline cases, a minority of the data set [51]. In most of the cases, which were clearly reliable (eg, HR based on very clean ECG) or clearly unreliable (eg, HR based on very noisy ECG), different versions of the algorithms that we explored yielded consistent ratings of vital sign reliability. (Note that these reliability algorithms were not a priori developed to diagnose major hemorrhage but to match clinicians' opinions regarding whether waveform segments were clean with well-defined heartbeats [27] or breaths [28].)

Third, the data set was notable in that many patients were missing a full set of reliable data. However, we contend that this is a salient finding of the study, rather than a limitation, because it emphasizes the prevalence of unreliable vital sign data. At the same time, it did not hamper the univariate analyses because there were suitably large populations for each analysis. Finally, for the multivariate analysis, we were able to report a valid ROC AUC for the broadest study population (any patient with at least 1 reliable vital sign within the first 15 minutes) by using an ensemble classifier, which can tolerate missing data. The performance of the ensemble classifier was assessed through cross-validation, that is, with distinct training and testing patient populations.

In terms of the external validity of the study, the issues that we studied have been previously recognized [1,2,4-6]. This report offers a novel, quantitative analysis of their magnitude of effect in actual prehospital practice. It is not certain to what extent the quantitative results of this analysis will apply to different clinical settings, for example, emergency department vs hospital ward vs ground EMS, and different make and model of patient monitors. Likewise, there may be salient differences given alternative populations, for example, patients older in age with a higher rate of β-blocker medication. However, the study population of this report was reasonably large (>600 subjects); and such considerations were outside its scope. This analysis provides a prima facie demonstration that each of the factors is important and that specific strategies can significantly alter diagnostic test characteristics of routine clinical data. Further work is warranted to explore these factors in a diversity of clinical arenas and populations.

5. Conclusion

The study is notable for quantifying the magnitude of the effect of physiologic variability and measurement error on a diagnostic application of vital signs. These sources of variability were commonplace in this clinical data analysis. Techniques that accounted for the variability yielded significantly improved diagnostic test characteristics. Vital sign data are often treated uncritically in published reports. The findings here suggest that these factors should be carefully considered when using vital signs in clinical practice or research protocols.

Acknowledgments

Funding/support:

This work was funded by the Combat Casualty Care Research Area Directorate of the US Army Medical Research and Materiel Command, Fort Detrick, MD.

Role of sponsor:

The sponsor did not participate in the design and conduct of the study; collection, management, analysis, and interpretation of the data; or preparation, review, or approval of the manuscript.

Disclaimer:

The opinions and 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 of the US Department of Defense. This article has been approved for public release with unlimited distribution.

References

[1] Edmonds ZV, Mower WR, Lovato LM, et al. The reliability of vital sign measurements. Ann Emerg Med 2002;39:233-7.

[2] Jones DW, Appel LJ, Sheps SG, et al. Measuring blood pressure accurately: new and persistent challenges. JAMA 2003;289:1027-30.

[3] Davis JW, Davis IC, Bennink LD, et al. Are automated blood pressure measurements accurate in trauma patients? J Trauma 2003; 55:860-3.

[4] Lovett PB, Buchwald JM, Sturmann K, et al. 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.

[5] Friesdorf W, Konichezky S, Gross-Alltag F, et al. Data quality of bedside monitoring in an intensive care unit. Int J Clin Monit Comput 1994;11:123-8.

[6] Chambrin MC. Alarms in the intensive care unit: how can the number of false alarms be reduced? Crit Care 2001;5:184-8.

[7] Brasel KJ, Guse C, Gentilello LM, et al. Heart rate: is it truly a vital sign? J Trauma 2007;62:812-7.

[8] Victorino GP, Battistella FD, Wisner DH. Does tachycardia correlate with hypotension after trauma? J Am Coll Surg 2003;196:679-84.

[9] Staessen JA, Fagard R, Thijs L, et al. A consensus view on the technique of ambulatory blood pressure monitoring. The Fourth International Consensus Conference on 24-Hour Ambulatory Blood Pressure Monitoring. Hypertension 1995;26:912-8.

[10] O'Brien E, Waeber B, Parati G, et al. Blood pressure measuring devices: recommendations of the European Society of Hypertension. BMJ 2001;322:531-6.

[11] Clayton DG, Webb RK, Ralston AC, et al. A comparison of the performance of 20 pulse oximeters under conditions of poor perfusion. Anaesthesia 1991;46:3-10.

[12] Clemmer TP, Orme Jr JF, Thomas FO, et al. Outcome of critically injured patients treated at level I trauma centers versus full-service community hospitals. Crit Care Med 1985;13:861-3.

[13] Hedges JR, Feero S, Moore B, et al. Comparison of prehospital trauma triage instruments in a semirural population. J Emerg Med 1987;5: 197-208.

[14] Koehler JJ, Malafa SA, Hillesland J, et al. A multicenter validation of the prehospital index. Ann Emerg Med 1987;16:380-5.

[15] Baxt WG, Berry CC, Epperson MD, et al. The failure of prehospital trauma prediction rules to classify trauma patients accurately. Ann Emerg Med 1989;18:1-8.

[16] Emerman CL, Shade B, Kubincanek J. A comparison of EMT judgment and prehospital trauma triage instruments. J Trauma 1991;31:1369-75.

[17] Bickell WH, Wall Jr MJ, Pepe PE, et al. Immediate versus delayed fluid resuscitation for hypotensive patients with penetrating torso injuries. N Engl J Med 1994;331:1105-9.

[18] Dutton RP, Mackenzie CF, Scalea TM. Hypotensive resuscitation during active hemorrhage: impact on in-hospital mortality. J Trauma 2002;52:1141-6.

[19] Buist MD, Moore GE, Bernard SA, et al. Effects of a medical emergency team on reduction of incidence of and mortality from unexpected cardiac arrests in hospital: preliminary study. BMJ 2002;324:387-90.

[20] Dacey MJ, Mirza ER, Wilcox V, et al. The effect of a rapid response team on major clinical outcome measures in a community hospital. Crit Care Med 2007;35:2076-82.

[21] Hillman K, Chen J, Cretikos M, et al. Introduction of the medical emergency team (MET) system: a cluster-randomised controlled trial. Lancet 2005;365:2091-7.

[22] Kenward G, Castle N, Hodgetts T, et al. Evaluation of a medical emergency team one year after implementation. Resuscitation 2004;61:257-63.

[23] Jones AE, Brown MD, Trzeciak S, et al. The effect of a quantitative resuscitation strategy on mortality in patients with sepsis: a meta-analysis. Crit Care Med 2008;36:2734-9.

[24] Peng R, Chang C, Gilmore D, et al. Epidemiology of immediate and early trauma deaths at an urban level I trauma center. Am Surg 1998;64:950-4.

[25] Sauaia A, Moore FA, Moore EE, et al. Epidemiology of trauma deaths: a reassessment. J Trauma 1995;38:185-93.

[26] Cooke WH, Salinas J, Convertino VA, et al. Heart rate variability and its association with mortality in prehospital trauma patients. J Trauma 2006;60:363-70.

[27] McKenna TM, Bawa G, Kumar K, et al. The physiology analysis system: an integrated approach for warehousing, management and analysis of time-series physiology data. Comput Methods Programs Biomed 2007;86:62-72.

[28] Chen L, McKenna TM, Reisner AT, et al. Algorithms to qualify respiratory data collected during the transport of trauma patients. Physiol Meas 2006;27:797-816.

[29] Reisner AT, Chen L, McKenna TM, et al. Automatically-computed prehospital severity scores are equivalent to scores based on medic documentation. J Trauma 2008;65:915-23.

[30] Yu C, Liu Z, McKenna T, et al. A method for automatic identification of reliable heart rates calculated from ECG and PPG waveforms. J Am Med Inform Assoc 2006;13:309-20.

[31] Lumley T, Diehr P, Emerson S, et al. The importance of the normality assumption in large public health data sets. Annu Rev Public Health 2002;23:151-69.

[32] DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44:837-45.

[33] Chen L, McKenna TM, Reisner AT, et al. Decision tool for the early diagnosis of trauma patient hypovolemia. J Biomed Inform 2008;41: 469-78.

[34] Luna GK, Eddy AC, Copass M. The sensitivity of vital signs in identifying major thoracoabdominal hemorrhage. Am J Surg 1989;157:512-5.

[35] McGee S, Abernethy III WB, Simel DL. The rational clinical examination. Is this patient hypovolemic? JAMA 1999;281:1022-9.

[36] West JG, Murdock MA, Baldwin LC, et al. A method for evaluating field triage criteria. J Trauma 1986;26:655-9.

[37] Chen L, Reisner AT, Gribok A, et al. Exploration of prehospital vital sign trends for the prediction of trauma outcomes. Prehosp Emerg Care 2009;13:286-94.

[38] Chen L, Reisner AT, Gribok A, et al. Can we improve the clinical utility of respiratory rate as a monitored vital sign? Shock 2009;31:574-80.

[39] Shapiro NI, Kociszewski C, Harrison T, Chang Y, Wedel SK, Thomas SH. Isolated prehospital hypotension after traumatic injuries: a predictor of mortality? J Emerg Med 2003;25:175-9.

[40] Lipsky AM, Gausche-Hill M, Henneman PL, et al. Prehospital hypotension is a predictor of the need for an emergent, therapeutic operation in trauma patients with normal systolic blood pressure in the emergency department. J Trauma 2006;61:1228-33.

[41] King RW, Plewa MC, Buderer NM, et al. Shock index as a marker for significant injury in trauma patients. Acad Emerg Med 1996; 3:1041-5.

[42] Zarzaur BL, Croce MA, Fischer PE, et al. New vitals after injury: shock index for the young and age × shock index for the old. J Surg Res 2008;147:229-36.

[43] Runcie CJ, Reeve WG, Reidy J, et al. Blood pressure measurement during transport. A comparison of direct and oscillotonometric readings in critically ill patients. Anaesthesia 1990;45:659-65.
[44] Clemmer TP, Orme Jr JF, Thomas F, et al. Prospective evaluation of the CRAMS scale for triaging major trauma. J Trauma 1985;25:188-91.
[45] Morris Jr JA, Auerbach PS, Marshall GA, et al. The Trauma Score as a triage tool in the prehospital setting. JAMA 1986;256:1319-25.
[46] D'Silva JL, Gill D, Mendel D. The effects of acute haemorrhage on respiration in the cat. J Physiol 1966;187:369-77.
[47] Kenney RA, Neil E. The contribution of aortic chemoceptor mechanisms to the maintenance of arterial blood pressure of cats and dogs after haemorrhage. J Physiol 1951;112:223-8.
[48] Landgren S, Neil E. Chemoreceptor impulse activity following haemorrhage. Acta Physiol Scand 1951;23:158-67.
[49] Winder CV. Combination of hypoxic and hypercapnic stimulation at the carotid body. Am J Physiol 1942;136:200-6.
[50] Hravnak M, Edwards L, Clontz A, et al. Defining the incidence of cardiorespiratory instability in patients in step-down units using an electronic integrated monitoring system. Arch Intern Med 2008;168: 1300-8.
[51] Chen X, Chen L, Reisner AT, et al. Using confidence intervals to assess the reliability of instantaneous heart rate and respiratory rate: The 33rd Annual EMBS Conference, IEEE Engineering in Medicine and Biology Society and Biomedical Society. Boston, MA; 2011.