EXPLORATION OF PREHOSPITAL VITAL SIGN TRENDS FOR THE PREDICTION OF TRAUMA OUTCOMES

Liangyou Chen, PhD, Andrew T. Reisner, MD, Andrei Gribok, PhD, Jaques Reifman, PhD

ABSTRACT

Objectives. We explored whether there are diagnostically useful temporal trends in prehospital vital signs of trauma patients. **Methods.** Vital signs were monitored during transport to a level I trauma center and electronically archived. Retrospectively, we identified reliable vital signs recorded from the 0- to 7-minute interval and from the 14- to 21-minute interval during transport, and, for each subject, we computed the temporal differences between the two intervals' vital signs, the intrasubject 95% data ranges, the values during the initial 2 minutes, and the 21-minute overall means. We tested for differences between subjects with major hemorrhage versus control subjects, and computed receiver-operating characteristic (ROC) curves. We conducted sensitivity analyses, exploring alternative clinical outcomes, temporal windows, and methods of identifying reliable data. **Results.** Comparing major hemorrhage cases versus controls, there were no discriminatory differences in temporal vital sign trends. Hemorrhage cases had significantly wider intrasubject data ranges for systolic blood pressure (SBP), respiratory rate (RR), and shock index (SI) versus controls. All results were consistent in several sensitivity analyses. **Conclusions.** Our findings add to a growing body of evidence that prehospital vital sign trends over 21 minutes or less are unlikely to be diagnostically useful because of substantial nondirectional fluctuations in vital signs that would obscure any subtle, progressive temporal trends. SBP, RR, and SI values were significantly different for high-acuity patients, and had more variability. Taken together, these findings suggest that higher-acuity patients experience episodes of instability rather than gradual, steady decline. Measures that account for data variability, such as taking the average of multiple measurements, may improve the diagnostic utility of prehospital vital signs. **Key words:** vital signs; trauma; hemorrhage; trends; variability

INTRODUCTION

Progressive traumatic pathologies, such as uncontrolled hemorrhage, can cause directional changes in vital signs, such as progressive hypotension and tachycardia. These temporal trends may be of diagnostic value. This paper explores whether there are significant temporal trends in the prehospital vital signs of trauma patients and, more importantly, whether these temporal trends are diagnostically useful. There has been speculation that sophisticated temporal trend-based analysis of physiologic data may offer superior diagnostic utility,[1,2] and some evidence to support this. Rhee and colleagues reported that trauma score *change* during transport added significantly to the initial trauma score as a predictor of mortality.[3] A worsening of the shock index (SI), defined as the ratio of heart rate (HR) to systolic blood pressure (SBP),[4] from the scene to the emergency department has been associated with significantly higher acuity.[5] Indeed, well-known monitors now display vital sign trends,[6] implicitly endorsing the idea that there is value in trend information, even as temporal trend-based monitoring has shown mixed clinical value in the intensive care unit (ICU) setting.[7,8]

Conversely, there are also changes in vital signs that do not trend through time, such as variability caused by discrete events (e.g., sympathetic activation due to painful manipulation of an injured extremity or sympatholysis due to a dose of pain medication) as well as measurement errors. If such nondirectional variability is large enough, it can mask subtle temporal trends. The literature suggests that indeed there is notable temporal variation in prehospital vital signs. In a seminal paper on prehospital severity scores, Morris and colleagues found that a Revised Trauma Score that is *either* improving *or* getting worse is an indicator of high severity.[9] Even when a trauma patient is normotensive upon arrival at a receiving facility, a preceding episode of prehospital hypotension has been associated with increased severity.[10,11] This suggests that there is a population of high-severity trauma patients with abnormal vital signs but without steady,

gradual declines. In theory, if typical trauma patients show noisy, large-amplitude vital sign temporal variations, it would be difficult to determine the "true" underlying directional trend in those vital signs (at least without a sufficiently wide temporal window of observation).

In this paper, our goal was to quantify prehospital vital sign temporal trends and vital sign temporal variability in a prehospital population of trauma patients with life-threatening pathology and compare them with those of control trauma patients who had less-severe injuries. We quantitatively examined group differences, as well as the discriminatory power of the various parameters. We tried to interpret our findings in a manner that is also consistent with the aforementioned reports.[3,7,9–11] We considered some of the implications and practical strategies for interpreting prehospital vital signs, which may be valuable to the transport team and caregivers in the receiving hospital. Ultimately, transport monitors would display measurements that have been proven the most diagnostically valuable from a temporal sequence of vital signs.

METHODS

Study Design

This study was based on physiologic time-series data collected from 898 trauma-injured patients during transport between August 2001 and April 2004 by medical helicopter from the scene of injury to the level I unit at the Memorial Hermann Hospital in Houston, Texas.[12] Investigational review board (IRB) approval for this data collection was given by the Memorial Hermann Hospital and the U.S. Army Medical Research and Materiel Command Office of Research Protection. Additional attribute data were collected retrospectively via chart review. The vital signs were measured by Propaq 206EL transport monitors (Protocol Systems, Beaverton, OR), downloaded to an attached personal digital assistant, and ultimately stored in our database.[13] The variables consisted of electrocardiogram (ECG), photoplethysmogram, and respiratory waveform signals recorded at 182, 91, and 23 Hz, respectively, and their corresponding monitor-calculated vital signs, recorded at 1-second intervals (HR, oxygen saturation of arterial hemoglobin [SaO_2], and respiratory rate [RR]). In addition, SBP, mean arterial pressure (MAP), and diastolic blood pressure (DBP) were collected intermittently at multiminute intervals. Pulse pressure (PP), the difference between SBP and DBP, was computed retrospectively, as was the SI, which was reported to be useful for the diagnosis of hemorrhagic hypovolemia.[4,12] The patient attribute data included demographics, injury descriptions, prehospital interventions, and hospital treatments. There were 100 attribute parameters for each patient, and these data

have undergone prior analysis.[2,12,14] For this study, we obtained de-identified patient data from Memorial Hermann Hospital.

Study Setting and Population

For each vital sign (SBP, RR, PP, HR, SaO_2, and SI), we studied the subpopulation of subjects possessing at least one reliable measurement recorded during the 0- to 7-minute interval, and at least one reliable measurement recorded during the 14- to 21-minute interval. The reason we studied different subpopulations for each vital sign was that only a small subset ($n = 97$, or 11%) of the original population possessed reliable data for *each* investigational vital sign. The reason we did not examine longer intervals is that there were relatively insufficient patient records of longer duration, as the average length of the vital sign records is about 26 minutes. We tested whether there were differences in population characteristics of the selected subpopulations versus the overall population using the chi-square test.

The reliability of prehospital vital signs has been questioned, e.g., by Low and Martin[15] and by Garner.[16] Even in-hospital vital signs are prone to erroneous measurement.[17,18] Therefore, we employed previously developed methods[19,20] that rigorously and systematically evaluate the reliability of prehospital monitor data. These automated algorithms rate each vital sign datum on an integer scale of 0 to 3. In this investigation, we studied data with a reliability level of ≥ 2:

- The HR reliability algorithm evaluates the ECG waveform and considers whether there is agreement between several different methods of computing HR. The algorithm was previously compared with blinded human experts for several hundred ECG waveform excerpts.[20] When the HR algorithm identified reliable data, in 97% of the cases, blinded human experts concurred that the waveform was clean and, in 100% of those cases, concurred with the monitor's reported HR. When the algorithm identified unreliable data, the human experts agreed 83% of the time, suggesting that the algorithm was more selective than the human experts.
- The RR reliability algorithm evaluates the impedance pneumogram, the source of the monitor's computed RR, and identifies rhythmic and clean segments.[19] RR that is computed exclusively from these clean, rhythmic segments has been shown to be statistically superior to standard measures of RR as a predictor of hospital intubation and major hemorrhage.[21]
- The blood pressure reliability algorithm compares the HR measured by an oscillometric noninvasive blood pressure cuff versus the ECG HR and also checks that the relationships between SBP, MAP, and DBP are physiologic.[22] Reliable SBP, as determined

by this algorithm, was found to be statistically superior to unreliable SBP as a predictor of major hemorrhage.[23]

- Computation of prehospital severity scores—the Revised Trauma Score[24] and the Prehospital Index,[25] which assign numerical severity scores for a trauma patient based on the patient's vital signs and mental status—as predictors of major hemorrhage and mortality, using vital signs deemed reliable by the algorithms, has been shown to be diagnostically equivalent to scores based on medic documentation and statistically superior to scores computed from vital signs that were deemed unreliable by the algorithms.[22]
- SaO_2 reliability is determined by the duration (or absence) of a clean photoplethysmographic waveform. With the use of this algorithm, the positive predictive value of prehospital hypoxia ($SaO_2 <91\%$) rises from less than 75% (for conventional SaO_2 measurements) to over 95% (for "reliable" SaO_2) as a predictor of in-hospital documentation of thoracic or intracerebral injury.[26]

Major hemorrhage was this study's primary outcome, defined as receipt of a blood transfusion within 24 hours after arrival at the hospital, along with documented injuries that were explicitly hemorrhagic. Such explicit injuries were one or more of the following: 1) laceration of solid organs, 2) thoracic or abdominal hematomas, 3) explicit vascular injury that required operative repair, or 4) limb amputation. In this primary analysis, patients who received blood but did not meet the documented injury criteria, i.e., ambiguous hemorrhagic patients, and patients who died before arrival at the hospital, were excluded from the analysis (121 patients excluded). Alternative outcome definitions, for major hemorrhage and for hospital respiratory interventions, were explored through a set of secondary sensitivity analyses, described below in the Sensitivity Analysis section.

Measurements and Data Analysis

We computed the average vital signs over 21 minutes of transport for each investigational population. We also computed the average vital sign trends, where the trend was computed as the difference between a subject's average measurements during the 0- to 7-minute interval and the subject's average measurements during the 14- to 21-minute interval. (In the sensitivity analysis, described below, we explored other manners of computing trends.) We also computed the *intrasubject* standard deviation, σ, for each 21-minute vital sign and assumed that the intrasubject 95% data range was equal to 2σ.

The discriminatory performances of the investigational parameters were evaluated by constructing the receiver-operating characteristic (ROC) curves and cal-

culating the area under the curve (AUC) using a maximum-likelihood method. We used the ROCKIT freeware[27] (University of Chicago) for these analyses. ROCKIT assumes a binormal ROC model; that is, data for each of the decision outcomes (hemorrhage and control) are considered to be normally distributed. ROCKIT automatically selects multiple decision thresholds based on the distribution of the input data and estimates the parameters of the ROC curve. The curves estimated from this method are smoother than empirically evaluated ROC curves and can better represent the relationship between vital sign variables and the decision outcomes.[28,29] We performed univariate ROC analyses for each vital sign and reported the estimated AUC and the corresponding 95% confidence interval (CI). We compared ROC curves of each vital sign's initial value (the average of measurements taken during the initial 2 minutes) versus the average of *all* 21-minute vital sign measurements, employing a paired AUC test. Statistical differences between hemorrhage and control patients were compared using two-tailed unpaired Student's t-test and the Mann-Whitney U test.[30]

Sensitivity Analysis

We repeated the preceding computations 1) using different clinical outcomes; 2) using different data quality criteria; 3) using different temporal windows of analysis; and 4) excluding patients with abnormal initial vital signs. Specifically, in 1), we repeated the calculations comparing trauma patients who required major *respiratory* interventions (either hospital intubation or chest tubes) versus the control patients who did not receive these interventions. In addition, we explored an alternative definition of major hemorrhage: cases in which the patients received emergency red blood cell transfusions, regardless of their documented anatomic injuries, versus control patients who did not receive a blood transfusion. In 2), we relaxed our data inclusion criteria; this increased our population sizes. We simply required that the subjects possess nonzero vital sign data, studying the most reliable data available for each subject (as determined by our automated algorithms). Similarly, we repeated the calculations using *all* 21 minutes of data, without any filtering of the vital signs based on reliability. In 3), we investigated different temporal windows, seeking significant trend differences between major hemorrhage cases and controls, examining trends computed from time 0–5 minutes to time 5–10 minutes; from time 5–10 minutes to time 10–15 minutes; from time 0–5 minutes to time 10–15 minutes; and from the initial 2 minutes to the rest of each patient's record (all times are relative to time t = 0, when the air crew first applied the Propaq monitor to the trauma patient). Finally, in 4), we excluded patients with frankly abnormal vital signs in the initial

2 minutes, e.g., patients whose initial SBP was <80 mmHg, and repeated the calculations.

RESULTS

Scene arrival time was a median of 41 minutes (interquartile range 33–51 minutes) after the recorded incident time. Total scene time was a median of 12 minutes (interquartile range 9–20 minutes), during which interval patient monitoring with the Propaq was initiated. Table 1 details the characteristics of the study population. Relative to the overall database, the primary investigational populations had lower rates of mortality and respiratory intervention (p < 0.05, chi-square test) but no significant differences in terms of hemorrhage incidence, mechanism of injury, and gender. (We also examined multiple populations who did not possess these differences in the mortality and respiratory intervention rates versus the total population, as described in the Sensitivity Analysis section below.) Regarding the incidence of major head injury, 17% of the overall database had Abbreviated Injury Scores (AISs) ≥3 for head injury, and in subpopulations, 17–19% of those patients had head AISs ≥3, which was not significantly different.

Figure 1 illustrates the magnitudes of the average intrasubject 95% data ranges versus the average temporal trends for the corresponding subpopulations over 21 minutes of transport time. When comparing major hemorrhage cases with control cases, we found that major hemorrhage cases had significantly wider SBP, RR, and SI data variability (i.e., the intrasubject 95% data ranges) than control cases (p < 0.0001, p = 0.01, and p < 0.00001, respectively, by Student's t-test; Mann-Whitney U test yielded similar results). For the other vital signs, PP, HR, and SaO_2, there were no significant differences in the intrasubject 95% data ranges.

Vital sign trends did not offer clinically useful discriminatory power (ROC AUCs) to distinguish the major hemorrhage cases from the control cases (Table 2).

There were no significant differences in vital sign temporal trends (SBP, RR, PP, HR, SaO_2, and SI trends) between the major hemorrhage cases and the control cases (by Student's t-test and Mann-Whitney U test).

Average vital signs show discriminatory value, with ROC AUC ranging from 0.66 to 0.84 (Table 3). $SBP_{(mean)}$ and $PP_{(mean)}$ (the averages of SBP and PP measured during 21 minutes of transport) were better discriminators of hemorrhage than the initial measurements of SBP and PP, respectively (ROC AUC 0.82 versus 0.71, and AUC 0.84 versus 0.78, respectively, p < 0.05 for both comparisons). The averages of SaO_2 and SI were not statistically significantly better than their initial values, in terms of raising the AUC. All vital signs in Table 3 were significantly different in major hemorrhage versus control cases, except for $SaO_{2(2\ min)}$.

Comparing Table 2 and Table 3, we found that the average vital signs offer more discriminatory power (higher ROC AUC) than the vital sign trends (p < 0.05 in all AUC comparisons).

Sensitivity Analysis

Our major findings were insensitive to the following: 1) alternative clinical outcomes (major respiratory interventions and the alternative definition of major hemorrhage); 2) different data quality criteria (relaxing the data quality inclusion requirements and thus increasing our population sizes; when we excluded fewer subjects, the mortality and the respiratory intervention rates of the study population were *the same* as those of the total population); 3) different temporal windows of analysis (examining trends computed from time 0–5 minutes to time 5–10 minutes; from time 5–10 minutes to time 10–15 minutes; from time 0–5 minutes to time 10–15 minutes; and from the initial 2 minutes to the rest of each patient's record); and 4) the exclusion of patients with frankly abnormal vital signs in the initial 2 minutes. Again, there were no changes to any findings.

TABLE 1. Characteristics of the Investigational Populations

Population*	No. Patients	Male Gender†	Penetrating Mechanism	Major Hemorrhage‡	Respiratory Intervention	Mortality
Total	898	660 (73%)	101 (11%)	90 (10%)	266 (30%)	94 (10%)
SBP trends	345	251 (73%)	33 (10%)	41 (12%)	78 (23%§)	14 (4%§)
RR trends	149	118 (79%)	14 (9%)	21 (14%)	37 (25%)	8 (5%)
PP trends	345	251 (73%)	33 (10%)	41 (12%)	78 (23%§)	14 (4%§)
HR trends	221	163 (74%)	20 (9%)	27 (12%)	54 (24%)	11 (5%§)
SaO_2 trends	240	176 (73%)	18 (8%)	17 (7%)	57 (24%)	14 (6%§)
SI trends	169	122 (72%)	18 (11%)	19 (11%)	38 (22%)	7 (4%§)

*For each vital sign, eligible subjects had at least one reliable measurement recorded during the 0- to 7-minute interval and another reliable measurement recorded during the 14- to 21-minute interval.

†There was no assigned gender for four patients in the database.

‡Excluded from analysis were 121 patients who received hospital blood but did not have documented injuries that were unambiguously hemorrhagic.

§Statistically significant versus the total population's rates. In the sensitivity analyses, we analyzed patient populations whose mortality and respiratory intervention rates were the same as the overall population's; details are available in the text.

HR = heart rate (in beats·min^{-1}); PP = pulse pressure (in mmHg); RR = respiratory rate (in breaths·min^{-1}); SaO_2 = oxygen saturation of arterial hemoglobin (in %); SBP = systolic blood pressure (in mmHg); SI = shock index (in beats·min^{-1} $mmHg^{-1}$).
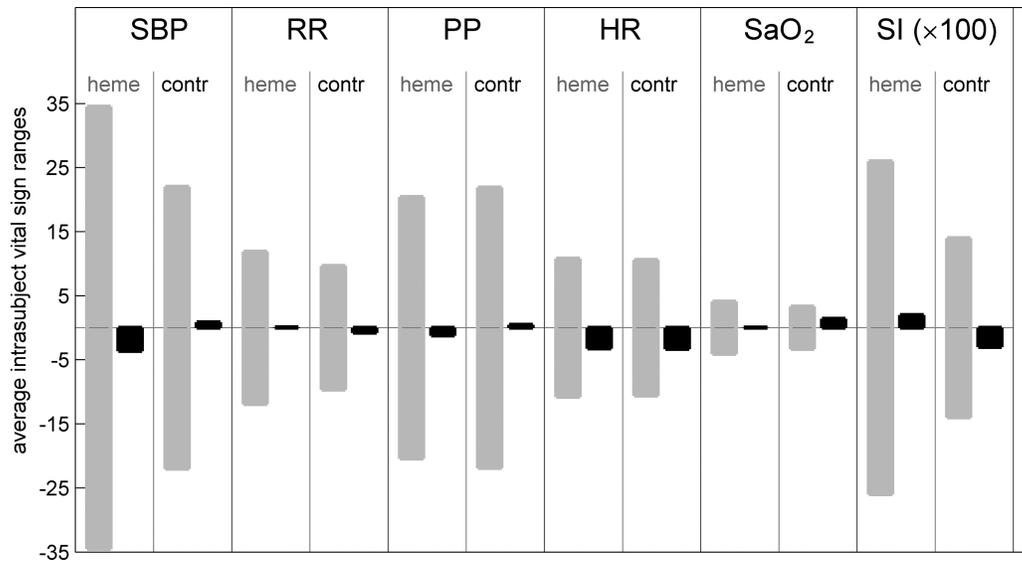
FIGURE 1. Major hemorrhage cases ("heme") versus control cases: intrasubject 95% vital sign ranges (*gray bars*) versus vital sign trends (*black bars*) averaged over all subjects in the corresponding study populations (systolic blood pressure [SBP] in mmHg; pulse pressure [PP] in mmHg; heart rate [HR] in beats·min^{-1}; respiratory rate [RR] in breaths·min^{-1}; oxygen saturation of arterial hemoglobin [SaO$_2$] in %; shock index [SI] in beats·min^{-1} mmHg^{-1} · multiplied by 100 [so the scale will be comparable for all illustrated vital signs]).

These numerous sensitivity analyses were consistent with the major findings of Table 2 and Table 3. To summarize, in *all* sensitivity analyses we found the following:

- For all vital signs, trends were nondiscriminatory (i.e., ROC AUCs were not significantly better than 0.50, consistent with Table 2).
- For all vital signs, their 21-minute averages were significantly different in major hemorrhage (or major respiratory intervention) versus control cases (i.e., consistent with Table 3).
- For all vital signs, the magnitude of the 95% data range (averaged over all subjects) was considerably larger than the magnitude of any temporal trend (i.e., consistent with Fig. 1).

TABLE 2. Major Hemorrhage Cases versus Controls: Vital Sign Temporal Trends Are Nondiscriminatory

	Major Hemorrhage	Control	ROC AUC (95% CI)*
SBP trend	−3.6 (SD 24.69)	0.9 (SD 15.89)	0.55 (0.45–0.66)
RR trend	0.1 (SD 6.44)	−0.8 (SD 5.76)	0.55 (0.41–0.68)
PP trend	−1.2 (SD 12.00)	0.5 (SD 14.67)	0.53 (0.44–0.62)
HR trend	−3.2 (SD 7.52)	−3.3 (SD 9.10)	0.50 (0.39–0.61)
SaO$_2$ trend	0.0 (SD 2.63)	1.4 (SD 5.56)	0.57 (0.41–0.72)
SI trend	0.02 (SD 0.24)	−0.03 (SD 0.12)	0.51 (0.36–0.67)

*Area under the receiver-operating characteristic curve and the 95% confidence interval.
AUC = area under the curve; CI = confidence interval; HR = heart rate (in beats·min^{-1}); PP = pulse pressure (in mmHg); ROC = receiver-operating characteristic; RR = respiratory rate (in breaths·min^{-1}); SaO$_2$ = oxygen saturation of arterial hemoglobin (in %); SBP = systolic blood pressure (in mmHg); SD = standard deviation; SI = shock index (in beats·min^{-1}·mmHg^{-1}).

- For SBP, RR, and SI, there were significantly wider ranges (i.e., more data variability) in major hemorrhage (or major respiratory intervention) versus control cases (i.e., consistent with Fig. 1).
- Using the 21-minute average of four prehospital vital signs (RR, HR, SaO$_2$, and SI) versus their initial values yielded improvements in ROC AUCs, anywhere from +0.02 to +0.10, although these improvements were not statistically significant.

TABLE 3. Initial Vital Sign "(2min)" versus Average Vital Sign during 21 Minutes of Transport "(mean)" for the Major Hemorrhage Cases versus Control Cases

	Major Hemorrhage	Control	ROC AUC (95% CI)*
SBP$_{(2min)}$	115 (SD 26)	134 (SD 23)	0.71 (0.61–0.79)†
SBP$_{(mean)}$	111 (SD 19)	135 (SD 21)	0.82 (0.74–0.87)†
RR$_{(2min)}$	32 (SD 9)	27 (SD 6)	0.70 (0.55–0.83)
RR$_{(mean)}$	30 (SD 9)	25 (SD 6)	0.66 (0.51–0.79)
PP$_{(2min)}$	43 (SD 13)	58 (SD 16)	0.78 (0.70–0.84)†
PP$_{(mean)}$	41 (SD 11)	58 (SD 14)	0.84 (0.77–0.90)†
HR$_{(2min)}$	109 (SD 20)	92 (SD 19)	0.75 (0.65–0.83)
HR$_{(mean)}$	107 (SD 20)	90 (SD 17)	0.75 (0.64–0.84)
SaO$_{2(2min)}$	96 (SD 3)	97 (SD 11)	0.72 (0.59–0.82)
SaO$_{2(mean)}$	97 (SD 2)	99 (SD 3)	0.75 (0.61–0.86)
SI$_{(2min)}$	0.97 (SD 0.29)	0.72 (SD 0.21)	0.77 (0.65–0.86)
SI$_{(mean)}$	0.93 (SD 0.22)	0.70 (SD 0.19)	0.79 (0.67–0.88)

*Area under the receiver-operating characteristic curve and the 95% confidence interval.
†p < 0.05, comparing vital signs from the initial 2 minutes versus the average of all 21-minute measurements.
AUC = area under the curve; CI = confidence interval; HR = heart rate (in beats·min^{-1}); PP = pulse pressure (in mmHg); ROC = receiver-operating characteristic; RR = respiratory rate (in breaths·min^{-1}); SaO$_2$ = oxygen saturation of arterial hemoglobin (in %); SBP = systolic blood pressure (in mmHg); SD = standard deviation; SI = shock index (in beats·min^{-1} mmHg^{-1}).

Several minor findings from the sensitivity analyses, however, were different from the primary results in Table 2 and Table 3:

- For cases requiring respiratory interventions versus their controls, there was more variability in SaO_2 ($p < 0.05$).
- In Table 3 (the primary findings), using the 21-minute average of RR versus the initial RR value did *not* improve the ROC AUC. When we changed the definition of major hemorrhage, using the 21-minute average of RR versus the initial RR value again did *not* improve the ROC AUC. However, for *all* other sensitivity analyses, using the 21-minute average of RR versus the initial RR value did (nonsignificantly) increase the ROC AUCs.
- In all sensitivity analyses, using the 21-minute average of SBP or PP versus the initial SBP or PP *did* increase the ROC AUCs. However, this improvement in ROC AUC was not statistically significant in a minority of the sensitivity analyses.
- When we analyzed all subjects in the database (i.e., not excluding subjects for failing to meet the data quality criteria), we found a significant association between the drop of PP by 7 mmHg over the entire transport versus hemorrhage (44% of major hemorrhage cases versus 28% of controls, $p = 0.02$ by the chi-square test). In the same larger population, we found a significant association between the drop of SBP by 10 mmHg over the entire transport and hemorrhage (28% of major hemorrhage cases versus 10% of controls, $p < 0.01$). We did not identify this association in the primary analysis or when examining other temporal windows, data quality criteria, or the other clinical endpoints.

DISCUSSION

We found that vital sign temporal trends were diagnostically weak and did not discriminate between sicker trauma patients and their controls, although the average of vital signs measured over 21 minutes showed higher discriminatory value than the initial vital signs. This may be, in large part, because there was considerable up-and-down, nondirectional temporal variability in the typical patient record relative to the magnitude of any temporal trends (illustrated in Fig. 1). Such vital sign variability during transport can "mask" any underlying temporal trend, making it difficult to estimate a vital sign's true temporal trajectory. Temporal trends were not statistically different between the hemorrhage and the control cases, and we did not find them useful for the discrimination of major hemorrhage (i.e., having low ROC AUCs).

We found that trauma patients with major injuries have lower SBPs and higher RRs (which is not surpris-

ing). We also found that higher-acuity subjects have significantly more variability (e.g., wider 95% intrasubject ranges) in SBP, RR, and SI. We conjecture that sick trauma patients suffered episodic decompensation (e.g., hypotensive or hypoxic intervals) as well as episodic recoveries, manifested as large swings in vital signs that increased the overall data range (i.e., vital sign variability) even in the absence of strong directional trends overall. We speculate that this variability might have been caused by any combination of true physiologic variability (e.g., fluctuations in pain, fear, relaxation, or pharmacologic intervention) as well as measurement errors.

If there are no major temporal trends, then taking the average of multiple serial vital signs can eliminate some of the random variability and yield a better estimate of the subject's "true" underlying vital signs, and may improve diagnostic classification. In this investigation, using the average of all SBP or all PP measurements during the 21 minutes of transport significantly improved discrimination (ROC AUCs) versus the initial values and the vital sign trends. Moreover, in the primary analysis and all sensitivity analyses, there were nonsignificant improvements in AUCs for the other average vital signs (SaO_2 and SI) versus vital signs measured in the initial 2 minutes. In summary, taking the average of serial prehospital SBPs, and perhaps other vital signs, may offer improvements in discriminatory capability.

It is important to consider how generalizable these findings are. Do our findings relate to other serious outcomes, and to clinical settings beyond air medical transport? Regarding the former, we report that there are major fluctuations in vital signs (e.g., the 95% range for SBP was ±34 mmHg for major hemorrhage cases and ±22 mmHg for control cases). These major fluctuations were found throughout a wide set of sensitivity analyses (different outcomes, different data reliability criteria, etc.). Because our patient records are characterized by such fluctuations, we argue that it is unlikely that any outcomes definition or data selection method will yield trend information that is sensitive and specific for any diagnosis. Simply put, it will be nearly impossible to identify progressive, gradual trends when the data fluctuate so dramatically.

The second question relates to the generalizability of our findings to clinical settings beyond air medical transport. Is the vital sign variability in this data set representative of other prehospital experiences? We found an absence of diagnostic trends, but we did find that sicker patients showed more vital sign variability. This is entirely analogous to what was found by Morris et al.[9] in an analysis of urban trauma patients during prehospital ground transport. Morris and colleagues found diagnostic equivalency between Trauma Scores computed earlier in time and those computed later, as predictors of a high Injury Severity Score. This

implies that there were no powerful trends, i.e., that higher-acuity patients did not develop progressively abnormal SBP and RR (determinants of the Trauma Score). Moreover, Morris et al. reported that patients with *either* improvement *or* deterioration in their Trauma Score were at increased risk of death. This implies that there must have been more vital sign variability in high-severity cases. Rhee and colleagues[3] commented on this finding of Morris et al. as follows: "No explanation could be found for these provocative results." Our findings suggest an explanation: Prehospital vital signs are highly variable, and even more so for the sickest patients. Gradual changes in vital signs caused by progressive pathologies, e.g., hemorrhage, are too weak to be diagnostically useful.

Moreover, our findings are consistent with the findings of Shapiro and colleagues[11] and of Lipsky and colleagues,[10] who independently reported that, among patients who arrived normotensive in the emergency department, one or more episodes of preceding hypotension were associated with higher acuity (Shapiro et al. studied air ambulance patients and used hypotension as a predictor of mortality, and Lipsky et al. studied ground transport patients and used hypotension as a predictor of emergent surgical intervention). If there are no major temporal trends during prehospital transport (as we report here), then measurements made at the end of transport are *not* more useful than preceding prehospital measurements. Though we did not find any gradual progressive trends, we found that our higher-acuity cases had lower average SBP and wider SBP fluctuations, which, in practice, meant that our high-acuity patients were prone to *nonsustained* episodes of frank hypotension, consistent with the reports of Shapiro et al. and Lipsky et al. Taken together, our findings, along with the reports of Morris et al., Shapiro et al., and Lipsky et al., suggest that prehospital vital sign variability is characteristic of high-acuity patients in myriad prehospital arenas. Our findings, together with what was reported by Morris et al., suggest that gradual progressive trends may be too weak to be of diagnostic value.

Of course, it is inevitable that major uncontrolled hemorrhage will ultimately cause a decrease in SBP. For that reason, we expect that in our data set there are probably real, but *weak*, differences in vital sign trends between major hemorrhage and controls. In one sensitivity analysis (in which we studied the largest possible population without any subject exclusions on the basis of their vital sign reliability), we did find that 28% of hemorrhage cases had a reduction in SBP after 2 minutes, compared with only 10% of the control cases. However, this association, even if not an artifact of repeated statistical testing, would be neither sensitive nor specific as a diagnostic tool, and it does not challenge our consistent finding that, given just 21 minutes of data in a highly uncontrolled prehospital air-ambulance environment, temporal trends in vital signs do not appear to be diagnostically useful. The up-and-down fluctuations in the vital signs appear to mask any gradual temporal trends.

LIMITATIONS

There are likely factors that weakened the discriminatory value of early temporal trends within this data set. Prehospital interventions may have "stabilized" patients (e.g., putting pressure on external bleeding, giving intravenous fluid and supplemental oxygen before or during air ambulance transportation) and so reduced temporal trends over the 21 minutes. To reduce the effects of "physiologic noise" (episodic physiologic changes due to moments of pain, fear, etc.), it may help to make additional measurements that account for the sources of physiologic variability. It is conceivable that additional information about analgesia, movement, agitation, fluid boluses, etc., might improve the discriminatory value of vital sign trends, if incorporated together in a multivariate predictive model. (We do not think that prehospital intravenous fluid administration was a major factor, however. In a prior analysis of these data, we developed a multivariate regression model to predict major hemorrhage.[31] We found that inputting the volume of fluid as a predictor of hemorrhage did not significantly add new information; i.e., it did not improve the multivariate classifier over and above the vital signs.)

In terms of our analytic methodology, the use of automated algorithms to identify reliable vital signs is a potential limitation, because if the algorithms are inaccurate, they may accept erroneous data and confound our results. The validity of these algorithms was reviewed in detail in the Methods section. This automated methodology was advantageous for several reasons. It enabled us to run, methodically, through a set of sensitivity analyses, in which we alternatively tightened and relaxed the reliability criteria for the vital sign data that we analyzed. In all permutations of our computations, we always found the same results, which reinforced our primary findings. Note that stringently selecting patients based on data reliability was a double-edged sword: it increased the reliability of the vital sign data that were included, but it also reduced the available data for analysis. In our primary analysis, a large number of the patients in the database were excluded because of insufficient data quality and quantity. This reduced the study's power to detect subtle differences in vital sign trends. For this reason, it was important to run those secondary analyses, in which we loosened the criteria, and we even performed one set of analyses without any restrictions on data quality. Such varied analyses are not possible if one examines only the vital signs documented by caregivers, and vital signs that are directly measured

and documented by caregivers have been shown to be quite imperfect.[15−18]

Perhaps hemodynamic deterioration occurs in abrupt transitions rather than in a steady fashion. Indeed, during the first hour after hemorrhage, transcapillary fluid shifts of up to 1 or 2 liters can occur,[32] counteracting slow hemorrhagic volume loss. Certain neurohumeral compensations are also activated during this time frame. Such physiologic compensations may (temporarily) counteract the progressive effects of traumatic pathology during prehospital transport, stabilizing early vital signs. Our negative results may not apply to longer temporal windows and other more controlled clinical settings. Indeed, trend-based monitoring has shown at least some mixed value in the ICU setting, where there is less measurement noise and longer observation intervals.[6−8]

CONCLUSIONS

Our findings add to a growing body of evidence that prehospital vital sign trends over 21 minutes or less are unlikely to be diagnostically useful. We found that nondirectional, up-and-down fluctuations in vital signs obscured any subtle, progressive temporal trends during air-ambulance transport. These substantial fluctuations were observed even when we analyzed the cleanest, most reliable physiologic data within each patient record, e.g., ECG segments without motion artifact, which suggests that the variability is in large part physiologic. Measures that account for variability through time, such as taking the average of serial measurements rather than relying only on the initial measurements, may offer some improvements in the discriminatory value of prehospital vital signs.

References

1. Chambrin MC. Alarms in the intensive care unit: how can the number of false alarms be reduced? Crit Care. 2001;5:184–8.
2. Holcomb JB, Salinas J, McManus JM, Miller CC, Cooke WH, Convertino VA. Manual vital signs reliably predict need for life-saving interventions in trauma patients. J Trauma. 2005;59:821–8.
3. Rhee KJ, Willits NH, Turner JE, Ward RE. Trauma Score change during transport: is it predictive of mortality? Am J Emerg Med. 1987;5:353–6.
4. Birkhahn RH, Gaeta TJ, Terry D, Bove JJ, Tloczkowski J. Shock index in diagnosing early acute hypovolemia. Am J Emerg Med. 2005;23:323–6.
5. Eastridge B, Salinas J, Baskin T, McManus J, Wade C, Holcomb J. Trend in shock index after injury predicts complications, life saving interventions, and mortality. Academic Emergency Medicine. 2007;14(S1):S138.
6. Giuliana K, Johannessen A, Crockett S. Clinical simulation: caring for a critically ill patient with sepsis. Critical Care. 2007;11(S2):P74.
7. Cunningham S, Deere S, Symon A, Elton RA, McIntosh N. A randomized, controlled trial of computerized physiologic trend monitoring in an intensive care unit. Crit Care Med. 1998;26:2053–60.
8. Schoenberg R, Sands DZ, Safran C. Making ICU alarms meaningful: a comparison of traditional vs. trend-based algorithms. Proc AMIA Symp. 1999:379–83.
9. Morris JA Jr, Auerbach PS, Marshall GA, Bluth RF, Johnson LG, Trunkey DD. The Trauma Score as a triage tool in the prehospital setting. JAMA. 1986;256:1319–25.
10. 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.
11. 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.
12. 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.
13. McKenna TM, Bawa G, Kumar K, Reifman J. The physiology analysis system: an integrated approach for warehousing, management and analysis of time-series physiology data. Comput Methods Programs Biomed. 2007;86(1):62–72.
14. 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.
15. Low RB, Martin D. Accuracy of blood pressure measurements made aboard helicopters. Ann Emerg Med. 1988;17:604–12.
16. Garner DC. Noise in medical helicopters. JAMA. 1991;266:515–6.
17. Jones DW, Appel LJ, Sheps SG, Roccella EJ, Lenfant C. Measuring blood pressure accurately: new and persistent challenges. JAMA. 2003;289:1027–30.
18. 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.
19. Chen L, McKenna TM, Reisner AT, Reifman J. Algorithms to qualify respiratory data collected during the transport of trauma patients. Physiol Meas. 2006;27:797–816.
20. 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;13:309–20.
21. Chen L, Reisner AT, Gribok A, McKenna TM, Reifman J. Can we improve the clinical utility of respiratory rate as a monitored vital sign? Shock. 2009;31(6):574–80.
22. Reisner AT, Chen L, McKenna TM, Reifman J. Automatically-computed prehospital severity scores are equivalent to scores based on medic documentation. J Trauma. 2008;65:915–23.
23. Chen L. Blood pressure data quality exercise. In personal correspondence with: Reisner AT, Gribok A. Reliable systolic blood pressure yielded a significantly better area under the receiver-operating characteristic curve (0.78 vs. 0.76, p < 0.05) than standard systolic blood pressure in 565 control and 71 hemorrhage patients. November 26, 2007.
24. Champion HR, Sacco WJ, Copes WS, Gann DS, Gennarelli TA, Flanagan ME. A revision of the Trauma Score. J Trauma. 1989;29:623–9.
25. Koehler JJ, Baer LJ, Malafa SA, Meindertsma MS, Navitskas NR, Huizenga JE. Prehospital Index: a scoring system for field triage of trauma victims. Ann Emerg Med. 1986;15:178–82.
26. Reisner AT. Interrelationships between oxygen saturation, traumatic pathology, and photoplethysmogram waveform quality in a prehospital population. In personal correspondence with:

Chen L, Gribok A. The odds ratio for predicting respiratory-related lifesaving interventions increases as the length of "verified" hypoxia (oxygen saturation $\leq$91% with a sustained, good-quality photoplethysmographic waveform) increases. September 5, 2007.

27. ROCKIT. Available from: http://xray.bsd.uchicago.edu/krl/KRL_ROC/software_index.htm. Accessed March 1, 2008.

28. Metz CE, Herman BA, Roe CA. Statistical comparison of two ROC-curve estimates obtained from partially-paired datasets. Med Decis Making. 1998;18:110–21.

29. Obuchowski NA. ROC analysis. Am J Roentgenol. 2005;184:364–72.

30. Glantz SA. Primer of Biostatistics. Sixth Edition. Columbus, OH: McGraw-Hill Companies, Inc., 2005.

31. Chen L, McKenna TM, Reisner AT, Gribok A, Reifman J. Decision tool for the early diagnosis of trauma patient hypovolemia. J Biomed Inform. 2008;41:469–78.

32. Prist R, Rocha-e-Silva M, Scalabrini A, et al. A quantitative analysis of transcapillary refill in severe hemorrhagic hypotension in dogs. Shock. 1994;1:188–95.