

# Automatically-Computed Prehospital Severity Scores are Equivalent to Scores Based on Medic Documentation

Andrew T. Reisner, MD, Liangyou Chen, PhD, Thomas M. McKenna, PhD, and Jaques Reifman, PhD

**Background:** Prehospital severity scores can be used in routine prehospital care, mass casualty care, and military triage. If computers could reliably calculate clinical scores, new clinical and research methodologies would be possible. One obstacle is that vital signs measured automatically can be unreliable. We hypothesized that Signal Quality Indices (SQI's), computer algorithms that differentiate between reliable and unreliable monitored physiologic data, could improve the predictive power of computer-calculated scores.

**Methods:** In a retrospective analysis of trauma casualties transported by air

ambulance, we computed the Triage Revised Trauma Score (RTS) from archived travel monitor data. We compared the areas-under-the-curve (AUC's) of receiver operating characteristic curves for prediction of mortality and red blood cell transfusion for 187 subjects with comparable quantities of good-quality and poor-quality data.

**Results:** Vital signs deemed reliable by SQI's led to significantly more discriminatory severity scores than vital signs deemed unreliable. We also compared automatically-computed RTS (using the SQI's) versus RTS computed from vital signs documented by medics. For the sub-

jects in whom the SQI algorithms identified 15 consecutive seconds of reliable vital signs data (n = 350), the automatically-computed scores' AUC's were the same as the medic-based scores' AUC's. Using the Prehospital Index in place of RTS led to very similar results, corroborating our findings.

**Conclusions:** SQI algorithms improve automatically-computed severity scores, and automatically-computed scores using SQI's are equivalent to medic-based scores.

**Key Words:** Remote triage, Monitoring, Trauma score, Vital signs, Signal quality.

*J Trauma.* 2008;65:915-923.

## Background

Prehospital severity scores for trauma casualties arose from a need to decide if a patient should be transported to a trauma center or if care at a community hospital would likely suffice. The best-known include Champion's Revised Trauma Score (RTS)<sup>1</sup> and the Prehospital Index (PHI)<sup>2</sup>; these scores are generated from formulas into which are entered a set of physiologic measurements. Although their discriminatory power is moderate,<sup>3-5</sup> prehospital severity scores can in principle be useful tools for civilian triage methodologies, for military triage<sup>6</sup> and for triage after a mass casualty incident.<sup>7</sup> Of course, severity scores can be no more accurate than the physiologic measurements input into the formulas. Unfortu-

nately, even in routine clinical practice, vital signs measurements can be unreliable,<sup>8,9</sup> and continuous monitoring is known to yield unreliable data because of motion artifact and other sources of error.<sup>10-14</sup> In uncontrolled environments, data quality is likely to be modest, and inaccurate measurements are likely to reduce the predictive power of severity scores.

## Goals of this Investigation

Recent reports described a set of signal quality indices ("SQI's"),<sup>15,16</sup> which are computer algorithms that rate the reliability of vital signs measurements made by a Propaq (Welch Allyn, Beaverton, OR) transport monitor (blood pressure, respiratory rate [RR], heart rate [HR], and oxygen saturation<sup>17</sup>). We tested if SQI's would improve the predictive power of automatically-computed severity scores calculated directly from Propaq data by eliminating unreliable vital signs. We also compared the predictive capabilities of scores computed automatically versus scores based on the vital signs documented by air ambulance caregivers ("medic-based scores"). Severity scores were evaluated as predictors of mortality and need for blood transfusion, for a population of trauma casualties transported by civilian air ambulance.

## Importance

An automated methodology for computing severity scores, tolerant of the vagaries of physiologic monitoring in uncontrolled environments, offers near-term and long-term value. Near-term, this automated method could be implemented in existing transport monitors and relieve a caregiver

Submitted for publication May 15, 2007.

Accepted for publication October 17, 2007.

Copyright © 2008 by Lippincott Williams & Wilkins

From the US Army Medical Research and Materiel Command (USAMRMC) (A.T.R., L.C., T.M.M., J.R.), Bioinformatics Cell, Telemedicine and Advanced Technology Research Center (TATRC), Fort Detrick, Maryland; and Department of Emergency Medicine (A.T.R.), Massachusetts General Hospital, Boston, Massachusetts.

Supported by the Combat Casualty Care Directorate of the U.S. Army Medical Research and Materiel Command, Fort Detrick, MD.

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 nor the US Department of Defense. This paper has been approved for public release; distribution is unlimited.

Address for reprints: Andrew Reisner, MD, Department of Emergency Medicine, Massachusetts General Hospital, Zero Emerson Place Suite 3B, Boston, MA 02114; email: areisner@partners.org.

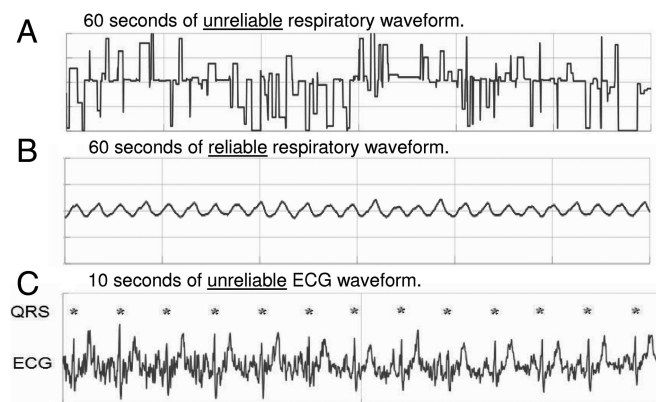
DOI: 10.1097/TA.0b013e31815eb142

of the computational task, or offer a double check when less experienced caregivers are making triage decisions. Longer-term, if it can be established that computers are able to accurately compute severity scores that are mathematically simple (e.g., RTS and PHI), the same principles may enable even more powerful automated diagnostic algorithms, for monitoring very large patient populations. Diverse efforts are already underway to develop new sensors and wireless networks for remote physiologic monitoring of disaster casualties, soldiers in combat, patients in overcrowded emergency departments, and convalescent patients at home.<sup>18-24</sup> Automatic computation of prehospital severity scores may be a first step in the development of new computer-supported triage and decision-support capabilities.

## METHODS AND MATERIALS

We held two primary hypotheses:

1. New SQI computer algorithms, designed to identify regular, consistent, and "clean" physiologic measurements, can distinguish reliable from unreliable Propaq monitor data (see Fig. 1). Reliable Propaq data identified by the SQI would be more predictive for severity scoring than data deemed unreliable.
2. Expert caregivers would only chart physiologic data that they thought to be credible. After the SQI excluded unreliable Propaq data, SQI-based automatic scores would be as predictive as medic-based scores.



**Fig. 1.** Examples of reliable and unreliable vital signs identified by signal quality indices (SQI's). (A) In this 60-second interval, the respiratory rate (RR) reported by the Propaq monitor was as high as 40 breaths per minute, but the SQI determined that the RR was unreliable because the underlying respiratory waveform was non-physiologic. (B) In this 60-second interval, the RR reported by the Propaq monitor was 26 breaths per minute. The SQI determined that this RR was reliable because the underlying respiratory waveform was reliable. (C) In this 10-second interval, the Propaq heart rate (HR) was reported as high as 120 beats per minute. The SQI determined that this HR was unreliable because it did not match the SQI's independently calculated HR, 79 beats per minute (actual QRS complexes are indicated by an asterisk, and the actual HR was 79 beats per minute).

## Setting

Our study is based on a retrospective analysis of civilian trauma patient data deposited in the Physiology Analysis System, which provides relational database and analysis tools to users over the Internet.<sup>25</sup> The original data were imported from the Trauma Vitals Database, which contains clinical data and continuous physiologic data from trauma patients transported from the scene of injury by helicopter service to the Level I unit at the Memorial Hermann Hospital in Houston, TX. Additional details about the trauma population and the Emergency Medical Service are available in.<sup>26</sup>

## Data Collection

Clinical data about the patient's hospital course, outcome (e.g., mortality), specific injuries, and specific therapy (e.g., red blood cell transfusion) were obtained by retrospective chart review.<sup>26</sup> The physiologic data were measured by a Propaq 206 monitor and its standard devices, during routine prehospital care of each casualty, and were downloaded to an attached personal digital assistant. The collected data include electrocardiogram (ECG) and photoplethysmogram (PPG) waveform signals and corresponding HRs and the RR, all calculated by Propaq's algorithms, and also the noninvasive blood pressure measurements (NIBP), including systolic blood pressure (SBP), mean arterial pressure, diastolic blood pressure, and associated NIBP HRs. The data collection was conducted with institutional review board approval, with a waiver of informed consent.

## Automatic Signal Quality Index Algorithms

The technical details of the automatic SQI algorithms for RR and HR have been reported.<sup>15,16</sup> Below, we summarize these methods, and describe a new SQI algorithm for SBP (Table 1). The SQI's rate the reliability of vital sign data with whole numbers running from least ("0") to most ("3") reliable. In this investigation, we considered vital signs rated  $\geq 2$  by the SQI as reliable.

**Table 1** Blood Pressure Signal Quality Index

IF SBP or DBP or NIBP HR changes THEN
IF ECG HR and NIBP HR are within 10 beats
AND PP > 10 mm Hg
AND abs(MAP—estimated MAP) < 10% MAP
THEN SQI = 3 (Outstanding)
ELSE IF 40 < MAP < 160
AND abs(MAP—estimated MAP) < 10% MAP
THEN SQI = 2 (Usable)
ELSE IF 40 < MAP < 160
THEN SQI = 1 (Questionable)
ELSE
SQI = 0 (Unreliable)

DBP, diastolic blood pressure; PP, pulse pressure (SBP—DBP); NIBP HR, heart rate measured by Propaq BP cuff; ECG HR, heart rate measured by Propaq electrocardiography; MAP, mean arterial pressure measured by Propaq cuff; Estimated MAP = DBP + (PP/3).

### **Respiratory Rate Signal Quality Index**

The algorithm identifies the typical height, width, and area of each potential breath and only counts those that are similar to other breaths in the same 15-second window.<sup>16</sup> The RR is deemed reliable when the following conditions are met: (a) the Propaq's computed RR is roughly equal to our computed RR, and (b) our computed RR is not very sensitive to small changes in the pneumograph baseline (in other words, each breath is prominent). The RR SQI produces a value between 0 and 3, from least to most reliable.

### **Heart Rate Signal Quality Index**

The HR SQI algorithm computes how prominently the candidate R-waves stand out from the rest of the ECG tracing, how regularly the candidate R-waves occur, and the amount of high-frequency oscillation in the ECG signal.<sup>15</sup> A machine-learning classifier is used to rate the ECG waveform as reliable or not based on those measures. The algorithm also checks that the Propaq HR matches the algorithm's own HR. The HR SQI performs similar analysis of the PPG, and considers if the ECG HR is equal to the PPG HR. On the basis of all this information, the HR SQI produces a value between 0 and 3, from least to most reliable.

### **Systolic Blood Pressure Signal Quality Index**

The Propaq monitor uses the principle of oscillometry to measure noninvasive systolic, diastolic, and mean arterial blood pressure, and also reports a HR.<sup>17,27</sup> The magnitude of the volumetric oscillations, as a function of the external cuff pressure, is used by the Propaq to establish the patient's blood pressures, and the frequency of oscillations leads to the NIBP HR. Like all oscillometric devices, it is subject to motion artifacts, because arterial pulsations cannot be readily distinguished from other motions. Our blood pressure SQI consists of several heuristic rules based on the following: is the oscillometric HR close to the ECG HR? Is the reported mean arterial pressure about 1/3 of the value between the diastolic and systolic values? And is the pulse pressure (the difference of SBP and diastolic blood pressure) physiologic? See Table 1.

## **Data Processing**

We considered data in the first 15 minutes of each Propaq electronic record. We processed data for two subject populations.

### **Reliable Versus Unreliable Data**

We identified those subjects with at least 15 consecutive seconds of reliable SBP and RR data; and 15 consecutive seconds of less reliable SBP and RR data. Reliability was determined automatically by the SQI algorithms described above. RTS was automatically-computed from (a) the mean of all reliable RR and SBP data in the initial 15 minutes of a subject's Propaq electronic record; versus (b) the mean of less reliable RR and SBP data in the initial 15 minutes. Data were extracted using the ad hoc filtering capabilities of the Physiology Analysis System,<sup>25</sup> and the computations were performed using routines we developed in the MatLab v7.1 computing environment (MathWorks; Natick, MA).

### **Signal Quality Index-Based Scores Versus Medic-Based Scores**

We compared the predictive capabilities of scores computed automatically versus scores based on the vital signs documented by air ambulance caregivers ("medic-based scores"), for records with a minimum of 15 consecutive seconds of reliable RR and SBP data (per the SQI). RTS was computed from (i) the mean of the first 15 seconds of reliable RR and SBP data in each Propaq electronic record, versus (ii) the RR and SBP data charted by the medics.

Using SBP and RR values, RTS was computed as per Champion.<sup>1</sup> For "medic-based" RTS scores, the initial documented scene SBP was not present in our database for a majority of cases, whereas the lowest medic documented field SBP was available for all cases in the study population. Therefore, the latter parameter was substituted for the former. This modification to Champion's method is considered in the discussion.

For all score calculations, the medic-charted Glasgow Coma Scale (GCS) was used because there is no accepted, automated alternative. Note that we used Champion's Triage RTS instead of RTS (both were originally described in<sup>1</sup>), because the latter weighs GCS more heavily than its SBP and RR components.

## **Outcome Measures**

Superior trauma severity scores should be more discriminatory of mortality and resource utilization. Two specific outcomes were investigated, to check that any differences between medic-based scores and automatically-computed scores were not idiosyncratic of specific outcomes. We focused on (a) prediction of mortality because, in the aftermath of a disaster or during an active military operation, it is important to identify those casualties whose prognosis is so poor that there is no justification for risking resources on them; and (b) prediction of blood transfusion, because hemorrhage is both an extremely common cause of traumatic death but often treatable, and receipt of blood transfusion is a reasonable surrogate for a major hemorrhagic abnormality.

## **Statistical Analysis**

For the severity scores, receiver operating characteristic (ROC) curves were constructed for both outcomes of interest (death and blood transfusion). The area under the ROC curve (AUC) was used as our performance metric. We used the ROCKIT freeware (University of Chicago),<sup>28</sup> which performs statistical tests-of-significance between investigational ROC curves, using a maximum likelihood estimation to fit a binormal ROC curve to data.<sup>29-31</sup> Significance of the difference between two binormal ROC curves was tested by paired comparison of the AUC's.

## **Sensitivity Analysis**

We repeated the preceding calculations using the PHI. The motivation for this sensitivity analysis was to check that any patterns between medic-based scores and the automatically-

**Table 2** Five Scoring Elements of the Modified Prehospital Index

Parameter	Range	Score
Systolic blood pressure	>100	0
	86–100	1
	75–85	2
	<75	5
Heart rate	51–120	0
	>120	3
	<50	5
Respiratory rate	“Normal” (10 < RR ≤ 29)	0
	“Labored/shallow” (RR > 29)	3
	“Below 10 needs intubation” (RR < 10)	5
Level of consciousness	“Normal” (GCS = 15)	0
	“Confused combative” (6 ≤ GCS ≤ 14)	3
	“No intelligible words” (3 ≤ GCS ≤ 5)	5
	Penetrating abdominal or chest injuries	4

computed scores were independent of specific severity scores. The RTS and PHI remain among the best-known prehospital severity scores,<sup>7,32</sup> perhaps because alternatives are not convincingly superior.<sup>33,34</sup>

We computed the PHI based on the original method of Koehler,<sup>2</sup> which yields a severity score from a formula based on the HR, SBP, “level of consciousness,” “respiratory status,” and whether or not there is penetrating thoraco-abdominal trauma. To ascertain if each casualty had penetrating thoraco-abdominal trauma, we employed an automated text search of the abbreviated injury scale fields in our database. With the above parameters, it was possible to compute PHI using the scoring system summarized in Table 2.

Note that PHI requires HR data, in addition to SBP and RR data. RTS only requires SBP and RR data. PHI study groups are smaller than their respective RTS groups, because the exclusion criteria are applied to subjects’ HR data as well as their SBP and RR data.

## RESULTS

There were a total of 660 case records in the database. The size of the study groups, their demographics, and outcomes are tabulated in Table 3. All groups had similar distributions.

### Validation of Signal Quality Indices

In subjects with 15 seconds or more of both reliable (high SQI) and unreliable (worse SQI) vital sign data, the reliable vital signs produced an RTS score significantly more predictive of blood transfusion than unreliable vital signs ( $p = 0.0003$ ). Reliable data also yielded consistent trends toward improved prediction of mortality using RTS, and blood transfusion and mortality using PHI (+0.04, +0.09, and +0.11 for ROC AUC’s, respectively), though these were

**Table 3** Population Characteristics

Population Size	Male (%)	Age (year ± SD)	ISS (mean ± SD)	Blunt Injury	Head Injury*	GCS <8	Abdominal Injury*	Thorax Injury*	Field Intubated	Field Hypotension	DOA	PRBC Transfusion	Mortality
Subjects with at least 15 consecutive seconds of reliable and at least 15 consecutive seconds of unreliable Propaq data													
RTS	187	74	39 ± 16	24 ± 22	190 (91)	51 (27)	27 (14)	37 (20)	36 (19)	35 (19)	7 (4)	40 (21)	15 (8)
PHI	103	74	38 ± 15	23 ± 22	92 (89)	30 (29)	16 (16)	21 (20)	23 (22)	21 (20)	5 (5)	22 (21)	9 (9)
Subjects with at least 15 consecutive seconds of reliable Propaq data													
RTS	350	74	38 ± 16	26 ± 24	304 (87)	92 (26)	68 (19)	76 (22)	84 (24)	59 (17)	10 (3)	67 (19)	29 (8)
PHI	348	74	38 ± 16	25 ± 24	302 (87)	89 (26)	64 (18)	76 (22)	80 (23)	57 (16)	8 (2)	67 (19)	29 (8)
Overall database	660	74	38 ± 15	26 ± 25	587 (89)	178 (27)	129 (20)	127 (19)	149 (23)	104 (16)	26 (4)	128 (19)	68 (10)

Values within the parentheses indicate percentage.

\* Based on search of the abbreviated injury scale fields.

RTS, Triage Revised Trauma Score (requires SBP and RR); PHI, prehospital index (requires SBP, RR, and HR); DOA, dead on arrival; PRBC, packed red blood cell.



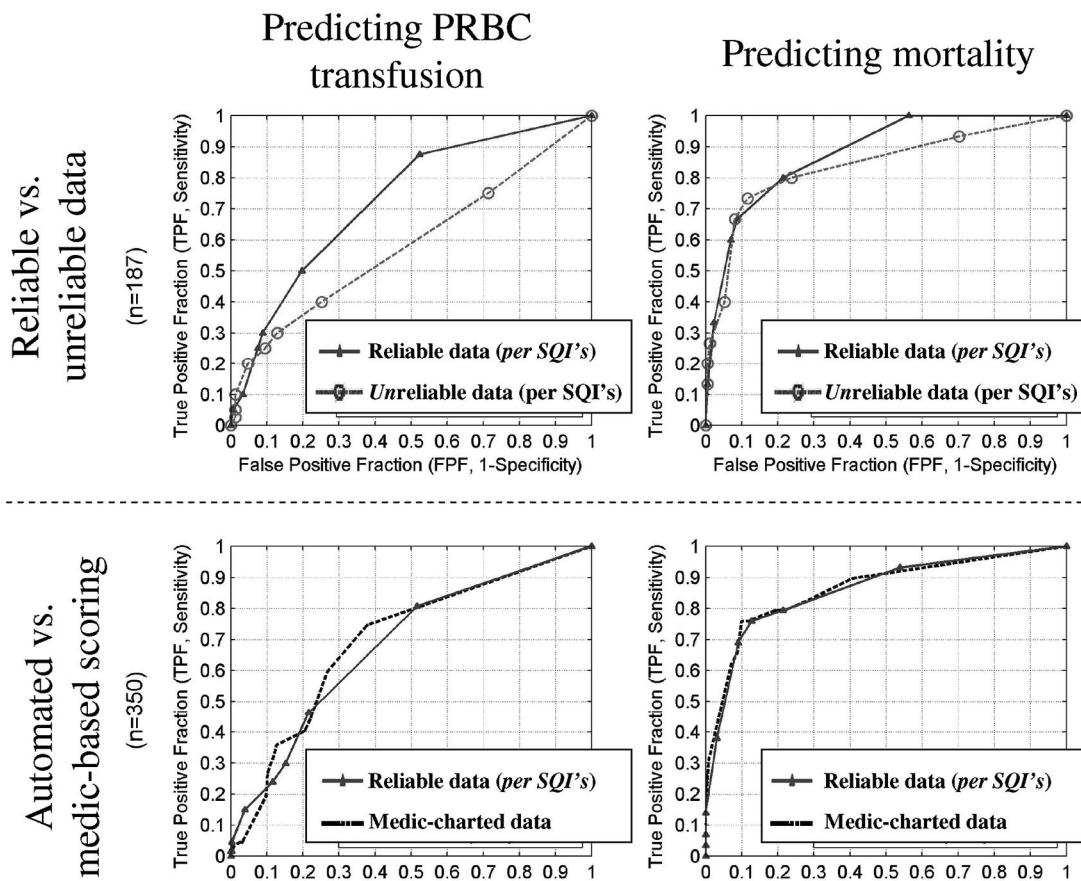
**Table 4** Reliable Propaq Data Versus Unreliable Propaq Data: Differences in ROC AUC for Prediction of Mortality and PRBC Transfusion With Automatically-Computed Severity Scores

	RTS (n = 187)		PHI (n = 103) [Sensitivity Analysis]	
	PRBC Transfusion	Mortality	PRBC Transfusion	Mortality
ROC AUC using unreliable Propaq data (per SQI)	0.58	0.84	0.63	0.76
$\Delta$ ROC AUC using reliable Propaq data (per SQI)	+0.15 ( $p = 0.0003$ )	+0.04 (NS)	+0.09 (NS)	+0.11 (NS)

Statistical comparisons between automatically-computed severity scores using two sources of vital signs data, for subjects with comparable quantities of reliable and unreliable data (e.g., subjects with at least 15 consecutive seconds of reliable vital signs data; and 15 consecutive seconds of less reliable vital signs data).

Propaq data were automatically classified as “reliable” or “unreliable” by SQI’s.

PRBC, packed red blood cells; NS, not significant.



**Fig. 2.** Predicting packed red blood cell transfusion (PRBC’s, left column) and mortality (right column) with Triage Revised Trauma Score using different vital signs inputs. Top row: Reliable versus unreliable data. SQI’s were used to automatically select subjects with both reliable and unreliable Propaq data (187 total, 40 received RBC, 15 died). For each subject, RTS was automatically computed from the reliable and from the unreliable data. The AUC’s are given in Table 4. Bottom row: Automatically-computed versus medic-based RTS. Subjects had at least 15 seconds of continuous, reliable Propaq data (350 total, 67 received RBC, 29 died). The two sources of vital signs are the reliable Propaq data (determined by the SQI), and the medic charted vital signs. Corresponding AUC’s are reported in Table 5.

not statistically significant. These results are tabulated in Table 4, and the ROC curves are shown in Figure 2.

**Signal Quality Index-Based Scores Versus Medic-Based Scores**

In 350 subjects, the SQI algorithms identified at least 15 consecutive seconds of reliable Propaq SBP and RR data.

ROC AUC’s for automatically-computed severity scores using the initial 15 seconds of reliable vital sign data were comparable to medic-based scores’ ROC AUC’s (Table 5). These ROC curves are illustrated in Figure 2. There were no significant differences between SQI-based scores and medic-based RTS scores. Similarly, there were no differences when the computations were repeated for the PHI (Table 5).

**Table 5** Automatically-Computed Versus Medic-Based Severity Scores: Differences in ROC AUC for Prediction of Mortality and PRBC Transfusion Using Different Sources of Vital Signs Data

	RTS (n = 350)		PHI (n = 348) (Sensitivity Analysis)	
	PRBC Transfusion	Mortality	PRBC Transfusion	Mortality
ROC AUC for medic-based severity scores	0.70	0.86	0.74	0.85
ΔROC AUC when automatically computed with reliable Propaq data	-0.03 (NS)	-0.00 (NS)	-0.02 (NS)	+0.03 (NS)

Statistical comparisons between severity scores computed from medic-charted vital signs data versus scores automatically-computed from reliable Propaq data (with at least 15 consecutive seconds of reliable vital signs data, automatically determined by SQI algorithms).

PRBC, packed red blood cells; NS, not significant.

## DISCUSSION

### Automated Differentiation of Reliable Versus Unreliable Vital Signs

These findings support the hypothesis that automatically-computed severity scores are improved by SQI algorithms, which can identify and exclude unreliable physiologic data. Reliable vital signs identified by the SQI's lead to significantly better RTS prediction of blood transfusion ( $p = 0.0003$ ). Reliable data also yield consistent trends toward improved prediction of mortality using RTS, and of blood transfusion and mortality using PHI (+0.04, +0.09, and +0.11 for ROC AUC's, respectively). The lack of statistical significance in these comparisons is likely a result of the small sample sizes (only fifteen died), and all the results are consistent with the highly significant finding. The statistically significant difference is all the more striking for two reasons. First, we used the same (medic's) GCS for all automatically-computed scores, which had the effect of making severity scores based on reliable versus unreliable data more similar. Second, we used the mean of all unreliable vital signs, which to some degree "averages out" random errors in individual measurements. Overall, these results are consistent with prior reports evaluating the effectiveness of SQI algorithms, where 97% of HR values deemed reliable by the SQI algorithm matched the HR measured by blinded human review,<sup>15</sup> and the RR SQI algorithm was similarly shown to identify reliable RR values.<sup>16</sup>

### Potential Clinical Effectiveness

Whenever prehospital severity scores are deemed useful, automated computation could be applied. Using only 15 consecutive seconds of reliable data, SQI-based scores are equivalent to medic-based scores, from data documented by experienced prehospital medics. Near-term, this automated methodology could be implemented into any transport monitor to produce severity scores. This could relieve a caregiver of the computational task, or offer a double check when less experienced caregivers (e.g., basic life support) are making difficult triage decisions (e.g., nonurban settings with significant transport times). Given an automated functionality, it would be more practical for emergency medical systems to routinely incorporate physiologic scores in their triage methodologies (presumably in conjunction with other criteria,

given the limitations of triage by RTS or PHI alone<sup>3-5</sup>). Longer-term, since we have demonstrated that computers can accurately compute severity scores that are mathematically simple (e.g., RTS and PHI), we speculate that computers could automatically apply even more diagnostically powerful and complicated algorithms to interpret biosignals and their patterns of change over time.<sup>35-38</sup>

Our "automatically-computed" method would be even more efficient if it did not require full GCS scoring by the caregiver. Full GCS scoring yields 13 ordinal levels (3 through 15), whereas the triage RTS score only requires 5 distinct mental status levels (1 through 5 points), and the PHI only 3 (0, 3, or 5 points). Prior work on the GCS suggests that the individual motor component of the GCS may be as good as, or better than, the full GCS.<sup>39</sup> The APVU (Alert/Pain/Verbal/Unresponsive) system also involves merely four mental status categories. This all suggests that triage may be performed using a simpler method with fewer distinct mental status levels. In an exploratory analysis, we computed severity scores using individual components of the GCS score rather than the complete three-component GCS score. The ROC AUC's were very similar ( $\Delta$ ROC AUC's  $-0.01 \pm 0.01$ , for predictions of mortality and transfusion with modified RTS computed from individual GCS components). The optimal implementation of automatically-computed severity scores might only require that caregivers input a simple four- or five-point mental status assessment, such as the GCS motor score, or the APVU score. Our results, in keeping with prior research, suggest that such a modification will not reduce the discriminatory power of prehospital triage scoring. In the even longer-term, a wholly automated methodology for mental status assessment would of course be desirable.

### Availability of Reliable Data

In 310 of 660 subjects, the SQI algorithms could not identify 15 consecutive seconds of high-quality SBP and RR data. In practice, this could limit automatically-computed severity scores, so this requires further discussion. First, note that the algorithm determines data reliability autonomously, and it could certainly provide an indication when sufficient reliable data have been collected and, conversely, when an automatically-computed severity score is based on unreliable

data. If notified that the data are unreliable, the caregiver may be able to rectify the situation, e.g., replacing loose ECG leads, holding the patient still for a 15-second interval, etc. In the meantime, severity scores based on unreliable data provide some information, though not as much as scores based on reliable data (Table 4). Future SQI algorithm enhancements, so that the SQI algorithms are more sensitive to reliable vital signs, may offer a modest benefit in broadening the clinical usefulness of automatically-computed severity scores. In prior reports, when the HR SQI algorithm identified unreliable data, blinded human experts concurred most of the time (e.g., 83% sensitivity for reliable data).<sup>15</sup> The algorithm was quite specific (97%) for data reliability.

It is possible that, for the most severely injured, it might be difficult to obtain reliable Propaq data. But when we examined the relationship between misclassification rate (false positives plus false negatives) versus a metric of severity (Injury Severity Score [ISS]) using an RTS cuff-off  $\leq 9$  as a predictor for mortality and blood transfusion, we found that the least injured casualties and the most severe casualties were less likely to be misclassified, because their vital signs tended to be normal or abnormal, respectively. The middle range casualties were more likely to have borderline physiologic profiles, which are most difficult to classify. For example, in the middle range of injury severity (ISS, 26–50), misclassification rates were 30% (mortality) and 50% (blood transfusion) higher than the misclassification rate for the most severe casualties (ISS, 51–75).

## Remote Monitoring and Triage

One motivation for this study was to assess the feasibility of automated clinical scoring for remote triage and other remote monitoring applications. Retrospective analysis of the Propaq data is similar to the remote monitoring problem: only the raw electronic data are considered, not other cues and information available to a caregiver at the patient's side. The United States Department of Defense has an active program to develop specialized light-weight physiologic sensors and wireless networks for remote monitoring of military combatants.<sup>20</sup> In military operations, the remote determination of which combat casualties have only mild injuries or, at the other extreme, which are unsalvageable, would be very useful, to prioritize casualty care, and because approaching casualties to assess their state during ongoing combat endangers medics. There are also civilian efforts to develop analogous devices for remote monitoring of disaster casualties and even patients convalescing at home.<sup>18–24</sup> Remote triage biosensors and algorithms would be useful in responding to civilian mass casualty incidents, when there are too few personnel available to evaluate and reevaluate a large number of casualties. Mental status might be assessed remotely by motion sensors (for approximate GCS motor scoring), two-way radio (for approximate GCS verbal scoring), or wholly novel signal processing or sensor solutions. For civilian mass casualty

responses, mental status could be assessed with a brief, caregiver interaction.

A major challenge to remote physiologic monitoring will be data quality. Our results suggest that SQI algorithms can distinguish reliable from unreliable physiologic data that the use of SQI algorithms can improve automatically-computed severity scores, and that automatically-computed severity scores are equivalent to medic-based scores for the majority of subjects in whom SQI algorithms identified 15 consecutive seconds of reliable vital sign data. Yet standard monitoring devices (e.g., ECG leads, pulse oximetry probe, NIBP cuff) are vulnerable to patient motion artifact and to improper placement.<sup>8–14</sup> A rushed air ambulance ride is unlikely to be conducive to pristine measurements. Therefore, the incidence of unreliable vital sign data is not surprising. Even in controlled in-hospital settings, false alarms due to erroneous physiologic measurements are an ubiquitous nuisance.<sup>10</sup> Remote triage will require novel monitoring hardware capable of reliable measurements in very challenging environments, because it seems likely that data quality would suffer further during remote operation unsupervised by caregivers.

## Limitations

We made minor modifications in how RTS and PHI were calculated for the “medic-based” scores. For instance, we used the “lowest-documented SBP” instead of the initial scene SBP, because the former was available for all case records. This may alter the ROC AUC's of the “medic-based scores.” However, we doubt these modifications altered our results, for three reasons. First, we investigated variations in how to compute RTS from a time-series of Propaq RR and SBP data, and the results were essentially unchanged. Specifically, we recomputed the RTS using the highest measured Propaq RR and SBP, and we recomputed the RTS using the lowest measured Propaq RR and SBP. We also recomputed RTS using data from the first half of each subjects' record, and from the second half. These exploratory permutations in how to compute RTS did not change our ROC AUC's by any appreciable amount (RTS permutations'  $\Delta$ ROC AUC's averaged  $0.00 \pm 0.01$ , for predictions of mortality and blood transfusion). This suggests that the method of selecting vital signs from a prehospital time series does not notably affect the ROC AUC of RTS. This temporal insensitivity, of early versus later triage scores in the majority of trauma casualties, has been previously described.<sup>35,36</sup>

Second, our results showed consistent trends between RTS and PHI (and between prediction of mortality and prediction of blood transfusion). This consistency between different severity scores provides some assurance that the results are valid. Last, the ROC curves for the medic-based severity scores are comparable to prior reports<sup>3–5</sup> (note that our RTS and PHI predict mortality better than blood transfusion; this is not surprising since GCS is more likely to predict massive head injury than hemorrhage).



## CONCLUSION

Using only 15 consecutive seconds of reliable Propaq data (as identified automatically by computer algorithms), automatically-computed severity scores are equivalent to medic-based scores, in terms of predicting mortality and red blood cell transfusion. This suggests that near-term, an automated methodology could be implemented into a transport monitor to produce severity scores based on physiologic data. The methodology might be further simplified, without loss of discriminatory power, using an abbreviated assessment of mental status. Longer-term, since we have demonstrated that computers can distinguish reliable versus unreliable data and accurately compute severity scores that are mathematically simple (e.g. RTS and PHI), we speculate that computers could automatically apply even more diagnostically powerful and complicated algorithms to interpret biosignals and their patterns of change over time. A major issue will be data quality, as many records did not contain the requisite 15 consecutive seconds of reliable Propaq data. Near-term, the computer could notify caregivers when the data are unreliable, so that the caregiver may rectify the situation. Longer-term, better biosensors that output more reliable measurements may enable automated analysis of physiologic data for large numbers of unsupervised subjects.

## ACKNOWLEDGMENTS

We are grateful to the University of Texas Health Science Center, and to COL John Holcomb and Dr. Jose Salinas, of the U.S. Army Institute of Surgical Research, for providing access to the Vital Signs (Trauma) database.

## REFERENCES

- Champion HR, Sacco WJ, Copes WS, Gann DS, Gennarelli TA, Flanagan ME. A revision of the Trauma Score. *J Trauma*. 1989; 29:623–629.
- Koehler JJ, Baer LJ, Malafa SA, Meindertmsa MS, Navitskas NR, Huizenga JE. Prehospital Index: a scoring system for field triage of trauma victims. *Ann Emerg Med*. 1986;15:178–182.
- Baxt WG, Berry CC, Epperson MD, Scalzitti V. The failure of prehospital trauma prediction rules to classify trauma patients accurately. *Ann Emerg Med*. 1989;18:1–8.
- Hedges JR, Feero S, Moore B, Haver DW, Shultz B. Comparison of prehospital trauma triage instruments in a semirural population. *J Emerg Med*. 1987;5:197–208.
- Eichelberger MR, Gotschall CS, Sacco WJ, Bowman LM, Mangubat EA, Lowenstein AD. A comparison of the trauma score, the revised trauma score, and the pediatric trauma score. *Ann Emerg Med*. 1989; 18:1053–1058.
- Burkle FM Jr, Newland C, Orebaugh S, Blood CG. Emergency medicine in the Persian Gulf War—Part 2. Triage methodology and lessons learned. *Ann Emerg Med*. 1994;23:748–754.
- Hodgetts TJ. Triage: a position statement. European Union Core Group on Disaster Medicine. 2002.
- Edmonds ZV, Mower WR, Lovato LM, Lomeli R. The reliability of vital sign measurements. *Ann Emerg Med*. 2002;39:233–237.
- Jones DW, Appel LJ, Sheps SG, Roccella EJ, Lenfant C. Measuring blood pressure accurately: new and persistent challenges. *JAMA*. 2003;289:1027–1030.
- Kacmarek RM. Alarms. In: Tobin MJ, ed. *Principles and Practice of Intensive Care Monitoring*. NY: McGraw-Hill, Inc; 1998:133–140.
- Goldman JM, Petterson MT, Kopotic RJ, Barker SJ. Masimo signal extraction pulse oximetry. *J Clin Monit Comput*. 2000;16:475–483.
- Amoore JN. A simulation study of the consistency of oscillometric blood pressure measurements with and without artifacts. *Blood Press Monit*. 2000;5:69–79.
- Portet F, Hernandez AI, Carrault G. Evaluation of real-time QRS detection algorithms in variable contexts. *Med Biol Eng Comput*. 2005;43:379–385.
- Kaiser W, Findeis M. Artifact processing during exercise testing. *J Electrocardiol*. 1999;32:212–219.
- 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. *JAMA*. 2006;13:309–320.
- Chen L, McKenna TM, Reisner AT, Reifman J. Algorithms to qualify respiratory data collected during transport of trauma patients. *Physiol Meas*. 2006;27:797–816.
- Propaq Encore Vital Signs Monitor Reference Guide. Beaverton: Protocol Systems, Inc; 1998.
- Field MJ, Grigsby J. Telemedicine and remote patient monitoring. *JAMA*. 2002;288:423–425.
- Wendelken SM, McGrath SP, Blike GT. A medical assessment algorithm for automated remote triage. In: *Proceedings of the 25th Annual Engineering in Medicine and Biology Conference*. Cancun, Mexico; 2003.
- Hoyt RW, Reifman J, Coster TS, Buller MJ. Combat medical informatics: present and future. *Proc AMIA Symp*. 2002:335–339.
- Seo J, Choi J, Choi B, Jeong DU, Park K. The development of a noninvasive home-based physiologic signal measurement system. *Telemed J E Health*. 2005;11:487–495.
- Grossman P. The LifeShirt: a multi-function ambulatory system monitoring health, disease, and medical intervention in the real world. *Stud Health Technol Inform*. 2004;108:133–141.
- Pino E, Ohno-Machado L, Wiechmann E, Curtis D. Real-time ECG algorithms for ambulatory patient monitoring. *AMIA Annu Symp Proc*. 2005:604–608.
- Yamakoshi K. Unconstrained physiologic monitoring in daily living for health care. *Front Med Biol Eng*. 2000;10:239–259.
- 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;86:62–72.
- Holcomb JB, Niles SE, Miller CC, Hinds D, Duke JH, Moore FA. Prehospital physiologic data and lifesaving interventions in trauma patients. *Mil Med*. 2005;170:7–13.
- Drzewiecki G. Noninvasive arterial blood pressure and mechanics. In: *The Biomedical Engineering Handbook*. Vol 1. 2nd ed. [http://www.radiology.uchicago.edu/krl/KRL\\_ROC/software\\_index6.htm](http://www.radiology.uchicago.edu/krl/KRL_ROC/software_index6.htm). Accessed June 2005.
- ROCKIT [computer program]. Chicago: University of Chicago; 2005.
- McNeil BJ, Hanley JA. Statistical approaches to the analysis of receiver operating characteristic (ROC) curves. *Med Decis Making*. 1984;4:137–150.
- Metz CE, Herman BA, Shen JH. Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data. *Stat Med*. 1998;17:1033–1053.
- Metz CE, Herman BA, Roe CA. Statistical comparison of two ROC-curve estimates obtained from partially-paired datasets. *Med Decis Making*. 1998;18:110–121.
- Kennedy K, Aghababian RV, Gans L, Lewis CP. Triage: techniques and applications in decision making. *Ann Emerg Med*. 1996;28:136–144.
- Ornato J, Mlinek EJ Jr, Craren EJ, Nelson N. Ineffectiveness of the trauma score and the CRAMS scale for accurately triaging



- patients to trauma centers. *Ann Emerg Med.* 1985;14:1061–1064.
34. Emerman CL, Shade B, Kubincanek J. Comparative performance of the baxt trauma triage rule. *Am J Emerg Med.* 1992;10:294–297.
35. 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;256:1319–1325.
36. Rhee KJ, Willits NH, Turner JE, Ward RE. Trauma score change during transport: is it predictive of mortality? *Am J Emerg Med.* 1987;5:353–356.
37. Cooke WH, Salinas J, Convertino VA, et al. Heart rate variability and its association with mortality in prehospital trauma patients. *J Trauma.* 2006;60:363–370.
38. Schoenberg R, Sands DZ, Safran C. Making ICU alarms meaningful: a comparison of traditional versus trend-based algorithms. *Proc AMIA Symp.* 1999:379–383.
39. Healey C, Osler TM, Rogers FB, et al. Improving the Glasgow Coma Scale score: motor score alone is a better predictor. *J Trauma.* 2003;54:671–678; discussion 678–680.