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 subjects 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.

**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)1 and the Prehospital Index (PHI)2; these scores are generated from formulas into which are entered a set of physiologic measurements. Although their discriminatory power is moderate,3–5 prehospital severity scores can in principle be useful tools for civilian triage methodologies, for military triage6 and for triage after a mass casualty incident.7

Of course, severity scores can be no more accurate than the physiologic measurements input into the formulas. Unfortu-
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.18–24 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 asterix, and the actual HR was 79 beats per minute).](image)

**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.25 The original data were imported from the Trauma Vitals Database, which contains clinical 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.26

**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.26 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.15,16 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 ≥2 by the SQI as reliable.

<table>
<thead>
<tr>
<th>Table 1 Blood Pressure Signal Quality Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>IF SBP or DBP or NIBP HR changes THEN</td>
</tr>
<tr>
<td>IF ECG HR and NIBP HR are within 10 beats</td>
</tr>
<tr>
<td>AND PP &gt; 10 mm Hg</td>
</tr>
<tr>
<td>AND abs(MAP—estimated MAP) &lt;10% MAP</td>
</tr>
<tr>
<td>THEN SQI = 3 (Outstanding)</td>
</tr>
<tr>
<td>ELSE IF 40 &lt; MAP &lt; 160</td>
</tr>
<tr>
<td>AND abs(MAP—estimated MAP) &lt;10% MAP</td>
</tr>
<tr>
<td>THEN SQI = 2 (Usable)</td>
</tr>
<tr>
<td>ELSE IF 40 &lt; MAP &lt; 160</td>
</tr>
<tr>
<td>THEN SQI = 1 (Questionable)</td>
</tr>
<tr>
<td>ELSE</td>
</tr>
<tr>
<td>SQI = 0 (Unreliable)</td>
</tr>
<tr>
<td>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).</td>
</tr>
</tbody>
</table>
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. 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. 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. 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, 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. 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), 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), which performs statistical tests-of-significance between investigational ROC curves, using a maximum likelihood estimation to fit a binormal ROC curve to data. 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-
computed scores were independent of specific severity scores. The RTS and PHI remain among the best-known prehospital severity scores,7,32 perhaps because alternatives are not convincingly superior.33,34

We computed the PHI based on the original method of Koehler,2 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>
<thead>
<tr>
<th>Parameter</th>
<th>Range</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Systolic blood pressure</td>
<td>$&gt;100$</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>86–100</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>75–85</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>$&lt;75$</td>
<td>5</td>
</tr>
<tr>
<td>Heart rate</td>
<td>51–120</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>$&gt;120$</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>$&lt;50$</td>
<td>5</td>
</tr>
<tr>
<td>Respiratory rate</td>
<td>“Normal” ($10 &lt; RR &lt; 29$)</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>“Laborated/shallow” ($RR &gt; 29$)</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>“Below 10 needs intubation” ($RR &lt; 10$)</td>
<td>5</td>
</tr>
<tr>
<td>Level of consciousness</td>
<td>“Normal” (GCS = 15)</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>“Confused combative” ($6 \leq GCS \leq 14$)</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>“No intelligible words” ($3 \leq GCS \leq 5$)</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Penetrating abdominal or chest injuries</td>
<td>4</td>
</tr>
</tbody>
</table>
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 4 Reliable Propaq Data Versus Unreliable Propaq Data: Differences in ROC AUC for Prediction of Mortality and PRBC Transfusion With Automatically-Computed Severity Scores**

<table>
<thead>
<tr>
<th></th>
<th>RTS (n = 187)</th>
<th>PHI (n = 103) [Sensitivity Analysis]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PRBC Transfusion</td>
<td>Mortality</td>
</tr>
<tr>
<td>ROC AUC using unreliable Propaq data (per SQI)</td>
<td>0.58</td>
<td>0.84</td>
</tr>
<tr>
<td>ΔROC AUC using reliable Propaq data (per SQI)</td>
<td>+0.15 (p = 0.0003)</td>
<td>+0.04 (NS)</td>
</tr>
</tbody>
</table>

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.
matched the HR measured by blinded human review, and 97% of HR values deemed reliable by the SQI algorithm. Reports evaluating the effectiveness of SQI algorithms, where measurements. Overall, these results are consistent with prior studies to some degree "averages out" random errors in individual measurements. 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 (ΔROC AUC’s = 0.01 ± 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.

**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). 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.

**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, and the RR SQI algorithm was similarly shown to identify reliable RR values.

**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). 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.

**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. 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 (ΔROC AUC’s = 0.01 ± 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. Routinely incorporate physiologic scores in their triage methodologies (presumably in conjunction with other criteria, given the limitations of triage by RTS or PHI alone). 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.

**Discriminatory Power of Prehospital Triage Scoring**

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.
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). 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 ≤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. In military operations, the remote determination of which combat casualties have only mild injuries or, at the other extreme, which are unsavable, 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. 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. 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. 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’ ΔROC AUC’s averaged 0.00 ± 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.

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 (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

33. Ornato J, Mlinek EJ Jr, Craven EJ, Nelson N. Ineffectiveness of the trauma score and the CRAMS scale for accurately triaging


