Using Confidence Intervals to Assess the Reliability of Instantaneous Heart Rate and Respiratory Rate

Xiaoxiao Chen, Member, IEEE, Liangyou Chen, Andrew T. Reisner, and Jaques Reifman

Abstract—Physiological waveform signals collected from unstructured environments are noisy, requiring automated algorithms to assess the reliability of the derived vital signs, such as heart rate (HR) and respiratory rate (RR), before they can be used for automated decision support. We recently proposed a weighted regularized least squares method to estimate instantaneous HR (HRR) and RR (RRR), which readily provides analytically based confidence intervals (CIs). Accordingly, this method can be extended to estimate instantaneous RR (RRR). In this study, we aim to investigate whether we can use CIs to select reliable HR (HRR) and RR (RRR). We calculated HRR and RRR for 532 and 370 trauma patients, respectively, grouped the rates according to their CIs, and investigated their reliability by determining their ability to diagnose major hemorrhage. The areas under a receiver operating characteristic curve of HR (HRR) and RR (RRR) were 0.70 and 0.66, respectively. HRR was superior to the average output of the clinical monitor (p < 0.05 by DeLong’s test), while HRR was equivalent. HRR and RRR provide a new approach to systematically and automatically assess the reliability of noisy, field-collected vital signs.

I. INTRODUCTION

PHYSIOLOGICAL waveform signals collected from trauma patients during transport from the scene of an accident to a trauma center are, usually, severely contaminated with noise. Vital signs derived from such noisy waveform recordings are therefore frequently inaccurate, precluding their use in automated decision-support algorithms. To address this challenge, our group previously developed physiological data qualification algorithms that automatically assess the reliability of major vital signs, such as heart rate (HR) and respiratory rate (RR) [1], [2]. While these algorithms have been shown to match the assessments made by human experts and significantly improve the accuracy of automated decision-support algorithms [3], they have some shortcomings: they are not designed for computing instantaneous rates, require the availability of redundant sensor measurements, and are based on heuristic criteria.

To resolve these shortcomings, we recently proposed a robust method for estimating instantaneous HR from noise-laden electrocardiogram (ECG) waveforms with normal sinus rhythm (i.e., no arrhythmia) [4]. This method implements a weighted regularized least squares (WRLS) algorithm for accurate HR estimation [regularized HR (HRR)] and, importantly, provides a systematic, analytically based approach to compute confidence intervals (CIs), which reflect uncertainties in the estimated HRR. We have shown that the CIs capture the noise level in ECG waveforms: large CIs reflecting high levels of ECG noise and vice versa [4]. In addition, the method can be readily extended to estimating instantaneous RR [regularized RR (RRR)] from respiratory waveforms. In this study, we aim to investigate whether CIs can be used to select reliable HR (HRR) and RR (RRR). More specifically, we divided the CIs into three non-overlapping ranges, and compared the extent to which HR (HRR) and RR (RRR) with smaller CIs (i.e., the more reliable rates) were able to improve the detection of major hemorrhage in trauma patients.

II. METHODS

A. Respiratory Rate and Confidence Interval Estimation

The estimation of HRR and its associated CI is described in detail in [4]. Here, we summarize the estimation of RRR and its CI in an analogous manner.

Because the low-frequency respiratory signal is subject to movement artifacts and erroneous placement of sensor electrodes on the body [5], [6], the corresponding respiratory waveforms are usually characterized by low signal-to-noise ratios. Therefore, before estimating the instantaneous RR, we first denoised the respiratory waveforms with a smoothing algorithm developed by our group [2]. Second, we detected the local maxima in the denoised respiratory waveform and formed a time series of the cumulative peak occurrence times (Pi), 0 < P1 < P2 < … < PN < T, where N is the total number of the cumulative peak occurrence times and T is the length of the denoised respiratory waveform.

Third, we formulated the cumulative peak occurrence time P as an integration of the peak-to-peak interval (PPI),

\[ P = A \cdot \text{PPI} + \epsilon, \]  

where P denotes an N × 1 vector of measured cumulative peak occurrence times (in seconds), A denotes an N × N lower triangular integration matrix with all non-zero elements equal to one, and \( \epsilon \) represents an N × 1 vector of measurement uncertainties. Therefore, we computed the cumulative peak occurrence times as

\[ C = A \cdot P, \]  

where C is an N × 1 vector of cumulative peak occurrence times.

These cumulative peak occurrence times, however, still contain some uncertainty, so we computed the CI for each cumulative peak occurrence time

\[ C = A \cdot P + \epsilon, \]  

where \( \epsilon \) represents the measured cumulative peak occurrence times. Using this method, we estimated the instantaneous RR and its CI in an analogous manner.
Using Confidence Intervals to Assess the Reliability of Instantaneous Heart Rate and Respiratory Rate

Physiological waveform signals collected from unstructured environments are noisy, requiring automated algorithms to assess the reliability of the derived vital signs, such as heart rate (HR) and respiratory rate (RR), before they can be used for automated decision support. We recently proposed a weighted regularized least squares method to estimate instantaneous HR (HRR), which readily provides analytically based confidence intervals (CIs). Accordingly, this method can be extended to the estimation of instantaneous RR (RRR). In this study, we aim to investigate whether we can use CIs to select reliable HRR and RRR. We calculated HRR and RRR for 532 and 370 trauma patients, respectively, grouped the rates according to their CIs, and investigated their reliability by determining their ability to diagnose major hemorrhage. The areas under a receiver operating characteristic curve of HRR and RRR with CI \( \leq 5 \) bpm were 0.70 and 0.66, respectively. RRR was superior to the average output of the clinical monitor (\( p < 0.05 \) by DeLong’s test), while HRR was equivalent. HRR and RRR provide a new approach to systematically and automatically assess the reliability of noisy, field-collected vital signs.
noise in $P$, and the PPI was estimated as the solution to an ordinary least squares (OLS) problem,

$$\text{PPI}_{\text{OLS}} = [(A^T \cdot A)^{-1} \cdot A^T] \cdot P,$$  \hspace{1cm} (2)

where $\text{PPI}_{\text{OLS}}$ represents an $N \times 1$ vector of the estimated PPI values. Consequently, the OLS solution of RR was $\text{RR}_{\text{OLS}} = 60/\text{PPI}_{\text{OLS}}$ in breaths per minute (bpm).

Next, by applying the WRLS algorithm as described in [4], we calculated the regularized PPI ($\text{PPI}_R$) as:

$$\text{PPI}_R = (W^T \cdot A^T \cdot A \cdot W + \lambda^2 \cdot L^T \cdot L)^{-1} \cdot W^T \cdot A^T \cdot A \cdot W \cdot \text{PPI}_{\text{OLS}},$$ \hspace{1cm} (3)

where $W$ denotes a diagonal $N \times N$ weighting matrix, whose elements are either zeros (represented by $10^{-5}$), for spike-like outliers detected in $\text{PPI}_{\text{OLS}}$ (and $\text{RR}_{\text{OLS}}$) via an impulse rejection filter [7], or ones for non-outliers, $L$ denotes a smoothing matrix that constrains high-frequency noise amplification in the PPI estimates and produces a smooth and consistent solution, and $\lambda$ represents a positive regularization parameter, which controls the tradeoff between the fit to the data and the smoothness of the solution. A standard choice for $L$ (and the one used here) is to use an $(N-2) \times 2$ matrix representing a second-order derivative [8]. We customized $\lambda$ for each patient. In particular, starting with $\lambda = 0$ (i.e., no regularization), we incrementally increased it until the absolute time rate of change of the estimated RRs dropped below a specified threshold of 8.0 bpm/s, which represents the average absolute time rate of change of RRs estimated from clean respiratory waveform segments in our trauma patient database [9]. Accordingly, we calculated $\text{RR}_R$ as $\text{RR}_R = 60/\text{PPI}_R$ in bpm.

Finally, we computed the CI for the estimated $\text{RR}_R$ through a standard formulation [10]:

$$\text{CI} = \text{RR}_R \pm t_{\alpha/2} \cdot \sqrt{\text{Var}(\text{RR}_R)}$$ \hspace{1cm} (4)

where $t_{\alpha/2}$ denotes a percentile of a Student’s $t$-distribution with a significance level of $\alpha$ and $\text{Var}(\text{RR}_R)$ represents the variance of $\text{RR}_R$. The derivation of $\text{Var}(\text{RR}_R)$ was analogous to the one described in [4]. Here we used $\alpha = 0.05$ and $t_{0.025} = 1.96$ for 95% CI.

### B. Study Data

In this study we used both discrete attribute data and physiological time-series data collected from 898 trauma casualties during and after transport by helicopter service from the scene of injury to the Level-I unit at the Memorial Hermann Hospital in Houston, TX [9]. The time-series variables were measured by Propaq 206EL vital-sign monitors (Welch Allyn; Skaneateles Falls, NY), downloaded to an attached personal digital assistant, and ultimately stored in our database. The physiological data include ECG waveforms (sampled at 182 Hz), respiratory waveforms (sampled at 23 Hz), their corresponding monitor-computed HR and RR (recorded at 1-s intervals), and other vital-sign data described elsewhere [11]. Patient attribute data, such as demographics, injury description, and treatments, were also collected via chart review. Data were collected and retrospectively analyzed with the approval of the local and the U.S. Army’s human subjects Institutional Review Board, Fort Detrick, MD.

### C. Outcome: Major Hemorrhage vs. Control

Our analyses required that we distinguished major hemorrhage patients from controls. Accordingly, patients with major hemorrhage were defined as those who received one or more units of packed red blood cell transfusion within 24 h upon arrival at the hospital and had a documented injury that was explicitly hemorrhagic, which was one or more of the following: (a) laceration or fracture of a solid organ, (b) thoracic or abdominal hematomas, (c) explicit vascular injury that required operative repair, or (d) limb amputation. 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. The remaining patients were labeled as controls.

### D. Data Analysis

We investigated the ability of CIs to select reliable $\text{HR}_R$ and $\text{RR}_R$ by determining the extent to which $\text{HR}_R$ and $\text{RR}_R$ with different CIs could distinguish major hemorrhage patients from controls. Thus, we divided the CIs into three non-overlapping ranges, $CI \leq 5$, $5 < CI \leq 20$, and $CI > 20$ bpm (beats per minute for HR and breaths per minute for RR), and selected different study populations for HR and RR so that each patient in each study population had at least one $\text{HR}_R$ value (or $\text{RR}_R$ value, for the corresponding study population) in each of the three ranges.

We evaluated the diagnostic performance of the averaged $\text{HR}_R$ and $\text{RR}_R$ in each range by performing univariate analysis to distinguish between major hemorrhage and control patients, constructing receiver operating characteristic (ROC) curves, and calculating the areas under the ROC curves (AUCs). We computed the ROC AUCs using DeLong’s method [12]. For comparison, we calculated ROC AUCs for the monitor-computed Propaq HR ($\text{HR}_P$) and RR ($\text{RR}_P$), which were averaged over all available data for each patient. We also computed ROC AUCs for the HR and RR calculated from our previously developed algorithms ($\text{HR}_C$ and $\text{RR}_C$), which were averaged over only reliable rates whose quality index (QI) was $\geq 2$ [1, 2]. In the analyses of the Propaq and QI-qualified rates, we used the same study populations as the ones used to assess the reliability of the CIs.

We applied the Pearson’s Chi-square test to compare the population demographics between the total population and the study sub-populations (except for the comparison of mean ages, where we used the Student’s $t$-test), and DeLong’s test to compare the ROC AUCs. We considered a $p$-value of $<0.05$ to be statistically significant.

### III. RESULTS

#### A. Population Statistics

The HR study population consisted of 470 controls and 62
hemorrhage patients, and the RR study population consisted of 323 controls and 47 hemorrhage patients. Table I shows the summary statistics of the populations. Both the HR and RR study populations had demographics similar to those of the total population, except that both study populations had lower mortality rates ($p < 0.05$), which was in accordance to our prior finding that higher-acuity casualties tend to have noisier data [2], [13], and those patients were not included because they lacked HR$_R$ and RR$_R$ with small CIs.

**B. Examples of Calculated Confidence Intervals**

Figure 1A shows 80 seconds (s) of a normalized ECG record for a typical patient in our trauma database, where the ECG was contaminated with spike noise for the segment between 505–550 s and relatively clean for the remaining segments. Figure 1B shows the corresponding HR$_R$ in beats per minute (bpm) derived with the WRLS algorithm, where the vertical bars indicate the width of the computed CIs. Here, we labeled HR$_R$ as reliable for those with CI $\leq$ 5 bpm, which coincided with the clean ECG segments.

Similarly, Figure 2 (A and B) shows 80 s of a normalized respiratory waveform record for a different patient in our trauma database and the associated RR$_R$ and CI estimates. We identified as reliable RR$_R$ those with CIs $\leq$ 5 bpm, which corresponded to the relatively clean respiratory waveforms outside of the 560–580 s segment. Conversely, the RR$_R$ within this noise-corrupted segment were characterized by large CIs, indicating their unreliable nature.

**C. Confidence Interval Performance Evaluation**

Table II summarizes the ROC AUCs of HR$_R$ and RR$_R$ for the three CI ranges in the detection of major hemorrhage in trauma patients. For comparison, it also includes the ROC AUCs of HR$_P$ and RR$_P$ and of the reliable HR$_C$ and RR$_C$. In general, ROC AUCs of HR$_R$ and RR$_R$ increased with smaller CIs. While the HR$_R$ result for CI $\leq$ 5 bpm was no different from that of HR$_P$, the improvement of RR$_R$ with CI $\leq$ 5 bpm over RR$_P$ was statistically significant. The results of our previously developed algorithms were consistently, but not statistically significantly, better than those obtained with regularized rates for CI $\leq$ 5 bpm.

**IV. DISCUSSION AND CONCLUSIONS**

In this study, we explored the utility of statistically based CIs to assess the reliability of field-collected HRs and RR. We evaluated the ability of the CIs to yield reliable rates by using HRs and RRs with different CIs to diagnose major hemorrhage in trauma patients, knowing that more reliable HR and RR estimates would offer better diagnostic value [3]. Our major finding was that HR$_R$ and RR$_R$ computed from smooth and clean waveforms (assessed by CI $\leq$ 5 bpm) were statistically significantly more diagnostic than those from noisy or arrhythmic waveforms (assessed by CI $> 20$ bpm), for diagnosing major hemorrhage. This suggested that the regularized rates with smaller CIs were physiologically more informative (i.e., more reliable) and provided superior clinical information for trauma patients, where arrhythmia was seldom observed.

### Table I

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Total Population</th>
<th>HR Study Population$^a$</th>
<th>RR Study Population$^a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population size</td>
<td>898</td>
<td>532</td>
<td>370</td>
</tr>
<tr>
<td>Male</td>
<td>660$^b$ (73%)</td>
<td>394 (74%)</td>
<td>279 (75%)</td>
</tr>
<tr>
<td>Female</td>
<td>234 (26%)</td>
<td>137 (26%)</td>
<td>91 (25%)</td>
</tr>
<tr>
<td>Mean age, yr</td>
<td>38 (SD 15)</td>
<td>38 (SD 15)</td>
<td>38 (SD 15)</td>
</tr>
<tr>
<td>Blunt injury</td>
<td>778$^c$ (87%)</td>
<td>476 (89%)</td>
<td>326 (88%)</td>
</tr>
<tr>
<td>Penetrating injury</td>
<td>101$^c$ (11%)</td>
<td>49 (9%)</td>
<td>38 (10%)</td>
</tr>
<tr>
<td>Mortality</td>
<td>94 (10%)</td>
<td>34 (6%)</td>
<td>22 (6%)</td>
</tr>
<tr>
<td>Major hemorrhage</td>
<td>97 (11%)</td>
<td>62 (12%)</td>
<td>47 (13%)</td>
</tr>
</tbody>
</table>

HR, heart rate; RR, respiratory rate; SD, standard deviation.

$^a$HR (or RR) Study Population is the subset of patients found to have regularized HRs (or RR) from each of the three confidence interval ranges

$^b$Four patients had no assigned gender in the total population

$^c$Nineteen patients had no assigned mechanism of injury

---

![Fig. 1. (A) Electrocardiogram (ECG) waveform and (B) corresponding regularize respiratory waveform record for a different patient.](image)

![Fig. 2. (A) Respiratory waveform and (B) corresponding regularized respiratory waveform record for a different patient.](image)

6941
TABLE II

<table>
<thead>
<tr>
<th>Vital Signs</th>
<th>Data Selection</th>
<th>ROC AUC (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HR$_R$</td>
<td>CI $\leq$ 5 bpm</td>
<td>0.70 (0.62-0.77)</td>
</tr>
<tr>
<td></td>
<td>5 $&lt;$ CI $\leq$ 20 bpm</td>
<td>0.70 (0.62-0.77)</td>
</tr>
<tr>
<td></td>
<td>CI $&gt;$ 20 bpm</td>
<td>0.66 (0.58-0.73)$^a$</td>
</tr>
<tr>
<td>HR$_P$</td>
<td>All available data</td>
<td>0.70 (0.63-0.77)</td>
</tr>
<tr>
<td>HR$_C$</td>
<td>QI $\geq$ 2</td>
<td>0.72 (0.64-0.78)</td>
</tr>
<tr>
<td>RR$_R$</td>
<td>CI $\leq$ 5 bpm</td>
<td>0.66 (0.57-0.73)$^a$</td>
</tr>
<tr>
<td></td>
<td>5 $&lt;$ CI $\leq$ 20 bpm</td>
<td>0.62 (0.54-0.70)</td>
</tr>
<tr>
<td></td>
<td>CI $&gt;$ 20 bpm</td>
<td>0.60 (0.52-0.67)$^b$</td>
</tr>
<tr>
<td>RR$_P$</td>
<td>All available data</td>
<td>0.54 (0.45-0.63)$^b$</td>
</tr>
<tr>
<td>RR$_C$</td>
<td>QI $\geq$ 2</td>
<td>0.71 (0.61-0.79)$^a$</td>
</tr>
</tbody>
</table>

HR$_R$ / RR$_R$, regularized heart rate / respiratory rate; HR$_P$ / RR$_P$, Propaq heart rate and respiratory rate; HR$_C$ / RR$_C$, heart rate / respiratory rate from prior reliability algorithms that assess morphology of the source waveforms [1], [2]; CI, confidence interval; QI, quality index; ROC AUC, area under the receiver operating characteristic curve.

$^a$Significantly different (p $<$ 0.05 by DeLong’s test) from HR$_P$ / RR$_P$.

$^b$Significantly different (p $<$ 0.05 by DeLong’s test) from HR$_R$ / RR$_R$ when CI $\leq$ 5 bpm.

When compared to the Propaq vital signs, HR$_R$ yielded no improvement, while RR$_R$ was significantly more diagnostic. We believe this is because measurement errors in HR$_R$ were without bias, so any errors were filtered out by taking an average value over several minutes. By contrast, RR$_R$ errors tended to be falsely elevated (i.e., motion artifact was counted falsely as a breath). As a result, the average RR$_P$ yielded a significantly worse ROC AUC than that of RR$_R$ when CI $\leq$ 5 bpm.

When compared to the vital signs from our previously developed QI algorithms, HR$_R$ and RR$_R$ were slightly (though not statistically) worse. The prior algorithms apply a set of heuristic rules involving the shape, timing, and frequency characteristics of the source waveforms (ECG and respiratory waveform) to determine when the measurements are reliable [1], [2]. By contrast, reliability for HR$_R$ and RR$_R$ are based entirely on the timing of the heartbeats / breaths, that is, the difference between the OLS solution and the regularized rates where larger differences yield larger CIs [4]. Another reason is because HR$_R$ and RR$_R$ are instantaneous rates, while HR$_C$ and RR$_C$ are average rates over 7- and 15-s data segments, respectively, which helps in further suppressing high-frequency noise in the estimated rates and improving estimation accuracy. Nevertheless, we believe that a slight decrement in performance of the proposed algorithm over the previously developed QI algorithms may be an acceptable trade off for certain applications, because the older QI algorithms are based on heuristic criteria and require redundant sensor measurements, while the proposed algorithm is statistically based and requires no additional information other than the original waveform.

As biosensors become ubiquitous in everyday life, it is important that we continue to develop algorithms which can improve our ability to automatically assess the reliability of vital signs while simultaneously attempting to develop more reliable sensors for physiological data collection. For both civilian and military applications, it is particularly important to infer reliable values of HRs and RRs collected from austere, unstructured environments, such as a battlefield, during the transport of trauma patients, in-home care of elderly patients, and in the monitoring of active individuals during physical activity, where the original physiological waveforms are prone to be contaminated with noise artifacts. The study presented here suggests that statistical CIs can be used as a systematic, analytical approach to automatically assess the reliability of field-collected HRs and RRs.

DISCLAIMER

The opinions and assertions contained herein are the private views of the authors and are not to be construed as official or as reflecting the views of the U.S. Army or of the U.S. Department of Defense. This paper has been approved for public release with unlimited distribution.

REFERENCES


