33rd Annual International Conference of the IEEE EMBS
Boston, Massachusetts USA, August 30 - September 3, 2011

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 (HR_R), which readily provides analytically based confidence intervals (CIs). Accordingly, this method can be extended to the estimation of instantaneous RR (RR_R). In this study, we aim to investigate whether we can use CIs to select reliable HR_R and RR_R. We calculated HR_R and RR_R 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_R and RR_R with CI ≤ 5 bpm (beats per minute for HR and breaths per minute for RR) were 0.70 and 0.66, respectively. RR_R was superior to the average output of the clinical monitor ($p < 0.05$ by DeLong's test), while HR_R was equivalent. HR_R and RR_R 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 (HR_R)] and, importantly, provides a systematic, analytically based approach to compute confidence intervals (CIs), which reflect uncertainties in the estimated HR_R. 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 (RR_R)] from respiratory waveforms. In this study, we aim to investigate whether CIs can be used to select reliable HR_R and RR_R. More specifically, we divided the CIs into three non-overlapping ranges, and compared the extent to which HR_R and RR_R 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 HR_R and its associated CI is described in detail in [4]. Here, we summarize the estimation of RR_R 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 (P_i), $0 < P_1 < P_2 < \ldots < P_N < T$, 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),

$$P = A \cdot PPI + \varepsilon, \quad (1)$$

where P denotes an $N \times 1$ vector of measured cumulative peak occurrence times (in seconds), A denotes an $N \times N$ lower triangular integration matrix with all non-zero elements equal to one, ε represents an $N \times 1$ vector of measurement

This work was supported by the Combat Casualty Care Research Program of the U.S. Army Medical Research and Materiel Command (USAMRMC), under Award #W81XWH-09-2-0147.

X. Chen is with the Bioinformatics Cell (BIC), Telemedicine and Advanced Medical Technology Research Center (TATRC), USAMRMC, Fort Detrick, MD 21702 USA (e-mail: xchen@bioanalysis.org).

L. Chen is with the BIC, TATRC, USAMRMC, Fort Detrick, MD 21702 USA (e-mail: lchen@bioanalysis.org).

A. T. Reisner is with the BIC, TATRC, USAMRMC, Fort Detrick, MD 21702 USA and with the Massachusetts General Hospital Department of Emergency Medicine, Boston, MA 02114 USA (e-mail: areisner@partners.org).

J. Reifman is a Senior Research Scientist and Director of the BIC, TATRC, USAMRMC, ATTN: MCMR-TT, 504 Scott Street, Fort Detrick, MD 21702 USA (corresponding author; phone: 301-619-7915; fax: 301-619-1983; e-mail: jaques.reifman@us.army.mil).

noise in P, and the PPI was estimated as the solution to an ordinary least squares (OLS) problem,

$$PPI_{OLS} = [(A^T \cdot A)^{-1} \cdot A^T] \cdot P, \quad (2)$$

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

Next, by applying the WRLS algorithm as described in [4], we calculated the regularized PPI (PPI_R) as:

$$PPI_R = (W^T \cdot A^T \cdot A \cdot W + \lambda^2 \cdot L^T \cdot L)^{-1} \cdot W^T \cdot A^T \cdot A \cdot W \cdot PPI_{OLS}, \quad (3)$$

where W denotes a diagonal N × N weighting matrix, whose elements are either zeros (represented by 10^{-5}), for spike-like outliers detected in PPI_{OLS} (and RR_{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 λ 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) × 2 matrix representing a second-order derivative [8]. We customized λ for each patient. In particular, starting with λ = 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 RR_R as $RR_R = 60/PPI_R$ in bpm.

Finally, we computed the CI for the estimated RR_R through a standard formulation [10]:

$$CI = RR_R \pm t_{\alpha/2} \cdot \sqrt{Var(RR_R)} \quad (4)$$

where $t_{\alpha/2}$ denotes a percentile of a Student's t-distribution with a significance level of α and $Var(RR_R)$ represents the variance of RR_R. The derivation of $Var(RR_R)$ was analogous to the one described in [4]. Here we used α = 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 HR_R and RR_R by determining the extent to which HR_R and RR_R with different CIs could distinguish major hemorrhage patients from controls. Thus, we divided the CIs into three non-overlapping ranges, CI ≤ 5, 5 < CI ≤ 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 HR_R value (or RR_R value, for the corresponding study population) in each of the three ranges.

We evaluated the diagnostic performance of the averaged HR_R and 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 (HR_P) and RR (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 (HR_C and RR_C), which were averaged over only reliable rates whose quality index (QI) was ≥ 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 compuated CIs. Here, we labeled HR_R as reliable for those with CI ≤ 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 ≤ 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 ≤ 5 bpm was no different from that of HR_P, the improvement of RR_R with CI ≤ 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 ≤ 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 RRs. 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 ≤ 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
DEMOGRAPHICS OF THE TOTAL AND STUDY POPULATIONS

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

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 RRs) 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

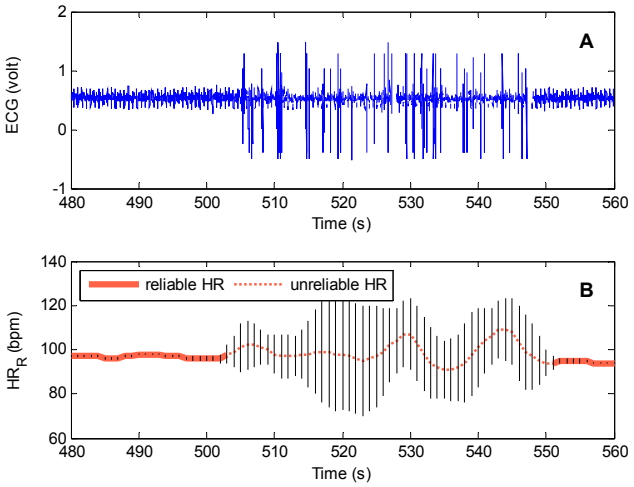

Fig. 1. (A) Electrocardiogram (ECG) waveform and (B) corresponding regularized heart rates (HR_R) and associated confidence intervals (CIs; vertical bars). Noisy waveform segments are characterized by HR_R with large CIs, whereas clean segments give rise to reliable HR_R (CIs ≤ 5 bpm).

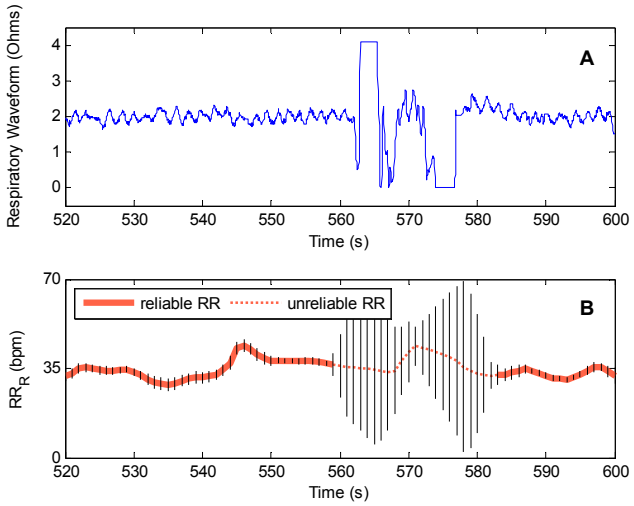

Fig. 2. (A) Respiratory waveform and (B) corresponding regularized respiratory rates (RR_R) and associated confidence intervals (CIs; vertical bars). Noisy waveform segments are characterized by RR_R with large CIs, whereas clean segments give rise to reliable RR_R (CIs ≤ 5 bpm).

TABLE II
DIAGNOSTIC ABILITY OF REGULARIZED HRS AND RRS WITH DIFFERENT CIS TO DIAGNOSE MAJOR HEMORRHAGE

Vital Signs	Data Selection	ROC AUC (95% CI)
HR_R	CI ≤ 5 bpm	0.70 (0.62-0.77)
	5 < CI ≤ 20 bpm	0.70 (0.62-0.77)
	CI > 20 bpm	0.66 (0.58-0.73)[ab]
HR_P	All available data	0.70 (0.63-0.77)
HR_C	QI ≥ 2	0.72 (0.64-0.78)
RR_R	CI ≤ 5 bpm	0.66 (0.57-0.73)[a]
	5 < CI ≤ 20 bpm	0.62 (0.54-0.70)
	CI > 20 bpm	0.60 (0.52-0.67)[b]
RR_P	All available data	0.54 (0.45-0.63)[b]
RR_C	QI ≥ 2	0.71 (0.61-0.79)[a]

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 ≤ 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_P were without bias, so any errors were filtered out by taking an average value over several minutes. By contrast, RR_P 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 ≤ 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

[1] C. Yu, Z. Liu, T. McKenna, A. T. Reisner, and J. Reifman, "A method for automatic identification of reliable heart rates calculated from ECG and PPG waveforms," *J. Am. Med. Inform. Assoc.*, vol. 13, no. 3, pp. 309-320, May/Jun, 2006.

[2] L. Chen, T. McKenna, A. Reisner, and J. Reifman, "Algorithms to qualify respiratory data collected during the transport of trauma patients," *Physiol. Meas.*, vol. 27, pp. 797-816, Jun, 2006.

[3] L. Chen, A. T. Reisner, A. Gribok, T. M. McKenna, and J. Reifman, "Can we improve the clinical utility of respiratory rate as a monitored vital sign?" *SHOCK*, vol. 31, no. 6, pp. 574-580, 2009.

[4] A. V. Gribok, X. Chen, and J. Reifman, "A robust method to estimate instantaneous heart rate from noisy electrocardiogram waveforms," *Ann. Biomed. Eng.*, vol. 39, no. 2, pp. 824-834, Feb, 2011.

[5] K. P. Cohen, W. M. Ladd, D. M. Beams, W. S. Sheers, R. G. Radwin, W. J. Tompkins, and J. G. Webster, "Comparison of impedance and inductance ventilation sensors on adults during breathing, motion, and simulated airway obstruction," *IEEE Trans. Biomed. Eng.*, vol. 44, no. 7, pp. 555-566, Jul, 1997.

[6] N. D. Khambete, B. H. Brown, and R. H. Smallwood, "Movement artifact rejection in impedance pneumography using six strategically placed electrodes," *Physiol. Meas.*, vol. 21, no. 1, pp. 79-88, Feb, 2000.

[7] R. A. Thuraisingham, "Preprocessing RR interval time series for heart rate variability analysis and estimates of standard deviation of RR intervals," *Comput. Methods Programs Biomed.*, vol. 83, no. 1, pp. 78–82, Jun, 2006.

[8] J. O. Ramsay and B. W. Silverman, *Functional Data Analysis*, New York, NY: Springer, 2005.

[9] J. B. Holcomb, J. Salinas, J. M. McManus, C. C. Miller, W. H. Cooke, and V. A. Convertino, "Manual vital signs reliably predict need for life-saving interventions in trauma patients," *J. Trauma Injury Infect. Crit. Care*, vol. 59, no. 4, pp. 821-828, Oct, 2005.

[10] N. R. Draper and H. Smith, *Applied Regression Analysis*, New York, NY: Wiley, 1966.

[11] L. Chen, T. M. McKenna, A. T. Reisner, A. Gribok, and J. Reifman, "Decision tool for the early diagnosis of trauma patient hypovolemia," *J. Biomed. Inform.*, vol. 41, no. 3, pp. 469-478, Jun, 2008.

[12] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach," *Biometrics*, vol. 44, no. 3, pp. 837-845, Sep, 1988.

[13] A. T. Reisner, L. Chen, T. M. McKenna, and J. Reifman, "Automatically-computed prehospital severity scores are equivalent to scores based on medic documentation," *J. Trauma*, vol. 65, no. 4, pp. 915–23, Oct, 2008.