Application of Information Technology ■

A Method for Automatic Identification of Reliable Heart Rates Calculated from ECG and PPG Waveforms

Chenggang Yu, PhD, Zhenqiu Liu, PhD, Thomas McKenna, PhD, Andrew T. Reisner, MD, Jaques Reifman, PhD

Abstract **Objective:** The development and application of data-driven decision-support systems for medical triage, diagnostics, and prognostics pose special requirements on physiologic data. In particular, that data are reliable in order to produce meaningful results. The authors describe a method that automatically estimates the reliability of reference heart rates (HRr) derived from electrocardiogram (ECG) waveforms and photoplethysmogram (PPG) waveforms recorded by vital-signs monitors. The reliability is quantitatively expressed through a quality index (QI) for each HRr.

Design: The proposed method estimates the reliability of heart rates from vital-signs monitors by (1) assessing the quality of the ECG and PPG waveforms, (2) separately computing heart rates from these waveforms, and (3) concisely combining this information into a QI that considers the physical redundancy of the signal sources and independence of heart rate calculations. The assessment of the waveforms is performed by a Support Vector Machine classifier and the independent computation of heart rate from the waveforms is performed by an adaptive peak identification technique, termed ADAPIT, which is designed to filter out motion-induced noise.

Results: The authors evaluated the method against 158 randomly selected data samples of trauma patients collected during helicopter transport, each sample consisting of 7-second ECG and PPG waveform segments and their associated HRr. They compared the results of the algorithm against manual analysis performed by human experts and found that in 92% of the cases, the algorithm either matches or is more conservative than the human's QI qualification. In the remaining 8% of the cases, the algorithm infers a less conservative QI, though in most cases this was because of algorithm/human disagreement over ambiguous waveform quality. If these ambiguous waveforms were relabeled, the misclassification rate would drop from 8% to 3%.

Conclusion: This method provides a robust approach for automatically assessing the reliability of large quantities of heart rate data and the waveforms from which they are derived.

■ **J Am Med Inform Assoc.** 2006;13:309–320. DOI 10.1197/jamia.M1925.

Decision-support algorithms that automatically interpret streaming physiologic time-series data are valuable tools for a broad range of medical surveillance applications. Examples of such applications include acute monitoring of patients in intensive care, home care, and ad hoc monitoring to continuously assess the health status of personnel, such as firefighters and soldiers, who are at risk of sudden injury.[1] Advances in vital-signs monitoring software/hardware, miniaturization, storage capacity, wireless transmission, and computational power now allow recording and analysis of large quantities of physiologic data in a timely fashion. These data are invaluable for the development of triage, diagnostic, and prognostic algorithms. However, collection of time-series vital-signs data is subject to many factors that affect the quality of the data. In particular, because vital-signs data are mostly collected in a noninvasive fashion, sensor motion artifact is of significant concern when the subject is moving or being transported. Other factors that may degrade data quality include electrical interference, sensor/monitor malfunction, and poor sensor placement on the subject. If valid decision-support algorithms are to be developed, and subsequently used to monitor patients, it is critical that reliable data be distinguished from artifact. Moreover, the process of distinguishing reliable from unreliable data must be automated since the sheer volume of collected time-series vital-signs data makes post hoc manual assessment an overwhelming task, while real-time streaming data cannot be manually evaluated at all.

Heart rate (HR) is a critical vital sign that is continuously monitored during transport of trauma patients from the scene

Affiliations of the authors: Bioinformatics Cell, Telemedicine and Advanced Technology Research Center, US Army Medical Research and Materiel Command, Fort Detrick, MD (CY, ZL, TM, JR); Department of Emergency Medicine, Massachusetts General Hospital, Boston, MA (ATR); Dr. Liu is currently with the Division of Biostatistics, Greenebaum Cancer Center and Department of Epidemiology and Preventive Medicine, University of Maryland Medical Center.

The work presented here was supported by the U.S. Army Medical Research and Materiel Command, Fort Detrick, MD.

The authors express their gratitude to Col. John Holcomb and Dr. Jose Salinas of the U.S. Army Institute of Surgical Research, Fort Sam Houston, San Antonio, TX, who provided the trauma patient data.

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 U.S. Army or the U.S. Department of Defense.

Correspondence and reprints: Jaques Reifman, PhD, U.S. Army Medical Research and Materiel Command MRMC/TATRC, 504 Scott Street, Ft. Detrick, MD 21702-5012; e-mail: <jaques.reifman@us.army.mil>.

Received for review: 08/01/05; accepted for publication: 01/16/06.

of injury to the hospital. It is used as an input for existing prehospital trauma severity scores, such as the prehospital index,[2,3] and may be used for future triage scoring systems. Also, studies of heart rate variability (HRV) suggest that decreasing HRV may be associated with worsening patient status. Unfortunately, we have observed that randomly imposed noise spikes are sometimes counted as heart beats by a vital-signs monitor. These sorts of data corruption can mislead diagnosis and compromise the development and application of inductive algorithms based on the synthesis of time-series physiologic data. Therefore, it is imperative that validated HRs be available for clinical use and development of advanced automated monitoring systems.

Automated HR calculation is usually based on the identification of heart beat signals, which could be taken from the QRS complex or simply the R waves in electrocardiogram (ECG) waveforms, or the pulse waves in photoplethysmogram (PPG) waveforms,[4–6] and dependent on the count of heart beats over a period of time. Given noisy waveforms, however, true heart beat signals may be masked or noise artifacts may resemble and be counted as true heart beats. Therefore, the quality of the HR calculated from the waveform depends on the quality of the waveform, making the qualification of waveforms a necessary step in validating HRs provided by a vital-signs monitor. Here, we refer to the monitor-calculated HRs as reference HRs (HRr). Accordingly, such HRr can be categorized as unreliable when the associated waveform is determined to be of suboptimal quality. For a conservative validation method, a high standard for good-quality waveforms is preferred to minimize the possibility that bad-quality HRs are falsely categorized as good. However, an overly stringent threshold is not advisable since it will increase the chance that good-quality HRs are falsely categorized as bad and, for post hoc data analysis, will considerably reduce the amount of available good-quality HR for the development of data-driven, decision-support algorithms.

In this paper, we present an approach to automatically and systematically qualify ECG HRr and PPG HRr provided by a vital-signs monitor. We assume that the monitor also provides the corresponding waveforms from which they are derived and that the monitored individuals are alive and have been subject to a trauma injury, where arrhythmia is seldom observed. The approach numerically qualifies each sampled HRr by assigning to it a quality index (QI) that concisely expresses its reliability. The approach exploits the physical redundancy provided by ECG HRr and PPG HRr and employs an independent method for recomputing HRs from the provided waveforms. This work addresses the first and key step of automatic and systematic qualification of large amounts of time-series data of our trauma database, so that we can next address our ultimate goal: mining these data to find predictive information for some clinical outcome.

Figure 1 illustrates the three components of the approach. In the first component, we use the newly developed adaptive peak identification technique, termed ADAPIT, to independently compute HRs (HRc) from both ECG and PPG waveform segments corresponding to the HRr we wish to validate. ADAPIT is a computationally simple peak detection algorithm, yet robust in the presence of random, motion-induced noise spikes that are often observed in waveforms collected during transport of trauma patients. Unless

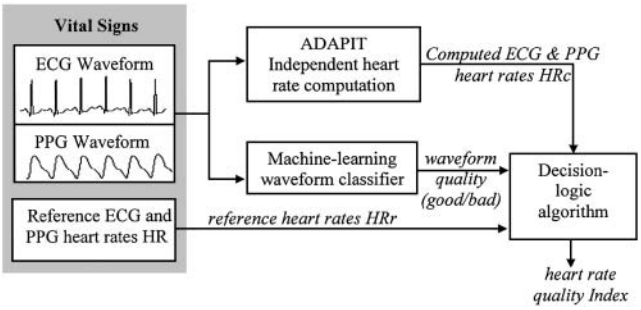

Figure 1. The three elements of the algorithm used to infer a quality index for reference heart rates provided by a vital-signs monitor.

accounted for, these noise spikes are likely to be counted as heart beats by the vital-signs monitor. Next, we separately qualify ECG waveform segments and PPG waveform segments as either good (excellent quality) or bad (suboptimal quality) through the use of a machine-learning algorithm in the form of support vector machines (SVMs).[7] In the third and final component, through a decision-logic algorithm, we combine the results of the two previous steps, the ADAPIT-computed ECG HRc and PPG HRc and the quality of their corresponding waveform segments, and compare them against ECG HRr and PPG HRr provided by a vital-signs monitor to infer a QI for the two HRr. A QI is inferred each time a HRr is provided by a vital-signs monitor and ranges from zero to three, with three representing the best-possible quality. In the absence of one of the waveforms, the decision–logic algorithm still provides a QI by assuming that the absent signal is present but possesses poor quality. Should additional HR sources be available, the approach could be extended by properly accounting for the quality of the new signal information and modifying the QI decision rules.

The approach is modular, self-contained, and independent of the data collection hardware. The waveform qualification algorithm (SVM), the HR recomputation algorithm (ADAPIT), and the QI decision rules are developed independently of each other and can be separately exchanged by functionally equivalent modules based on other methods. The three components form an effective, stand-alone system to validate reference HRs. Our approach is simply based on recorded time-series data from a vital-signs monitor, which is taken as a black box. From this point of view, the approach is independent of the data collection hardware.

Methods

In this section, we briefly describe the three components depicted in Figure 1: the HR estimation via the ADAPIT algorithm, the waveform qualification via an SVM algorithm, and the QI determination. We start by describing the data that precipitated the development of these components and that are used for the synthesis and testing of our algorithms.

Data

This study is based on physiologic time-series data collected during transport of trauma patients from the scene of injury by helicopter service to the Level I unit at the Memorial Hermann Hospital in Houston, TX.[8,9] The data were collected by ProPaq 206EL vital-signs monitors[10] on the helicopters and downloaded to an attached personal digital assistant. The data include, among other time-series data, ECG and

PPG waveform signals and their corresponding monitor-calculated HRr. The time series sampling rates are approximately 182 Hz for the ECG waveform, 91 Hz for the PPG waveform, and 1 Hz for the HRr. Complete vital-signs data for a total of 726 patients were deposited into our Physiology Analysis System,[11] which provides curated data and the ability to query and analyze discrete and time-series data over the Internet with a Web browser. The patient population is composed of 538 males and 186 females (two genders not noted), with a mean age of 37.7 years. The predominant type of injury is blunt trauma (641 patients), followed by penetrating trauma (78 patients).

Heart Rate Estimation with the ADAPIT Algorithm

The first component of our approach is the independent estimation of ECG and PPG HRs from their corresponding high-frequency waveforms. While we acknowledge that a large body of work has been developed over the past two decades,[4–6] most of the approaches are rather involved because they are designed to accommodate irregular morphologies and irregular rhythms, even though such phenomena are rarely observed in our data set of trauma victims. Due to the ambulatory nature and dynamic environment in which trauma data are collected, the major challenge is the filtering of noise and artifacts in the waveforms. Furthermore, most approaches are limited to the estimation of ECG-derived HRs through the detection and analysis of the QRS complex,[6] while we also need to estimate PPG-derived HRs. To achieve these objectives, we developed the ADAPIT algorithm. ADAPIT is a generic algorithm that, through changes in parameter settings and one computational step, is equally applicable to the estimation of HRs from both ECG and PPG waveforms and is designed to filter out noise and artifacts so they are not counted as heart beats. ADAPIT, however, may have limited ability to compute HRs in settings of highly irregular rhythms.

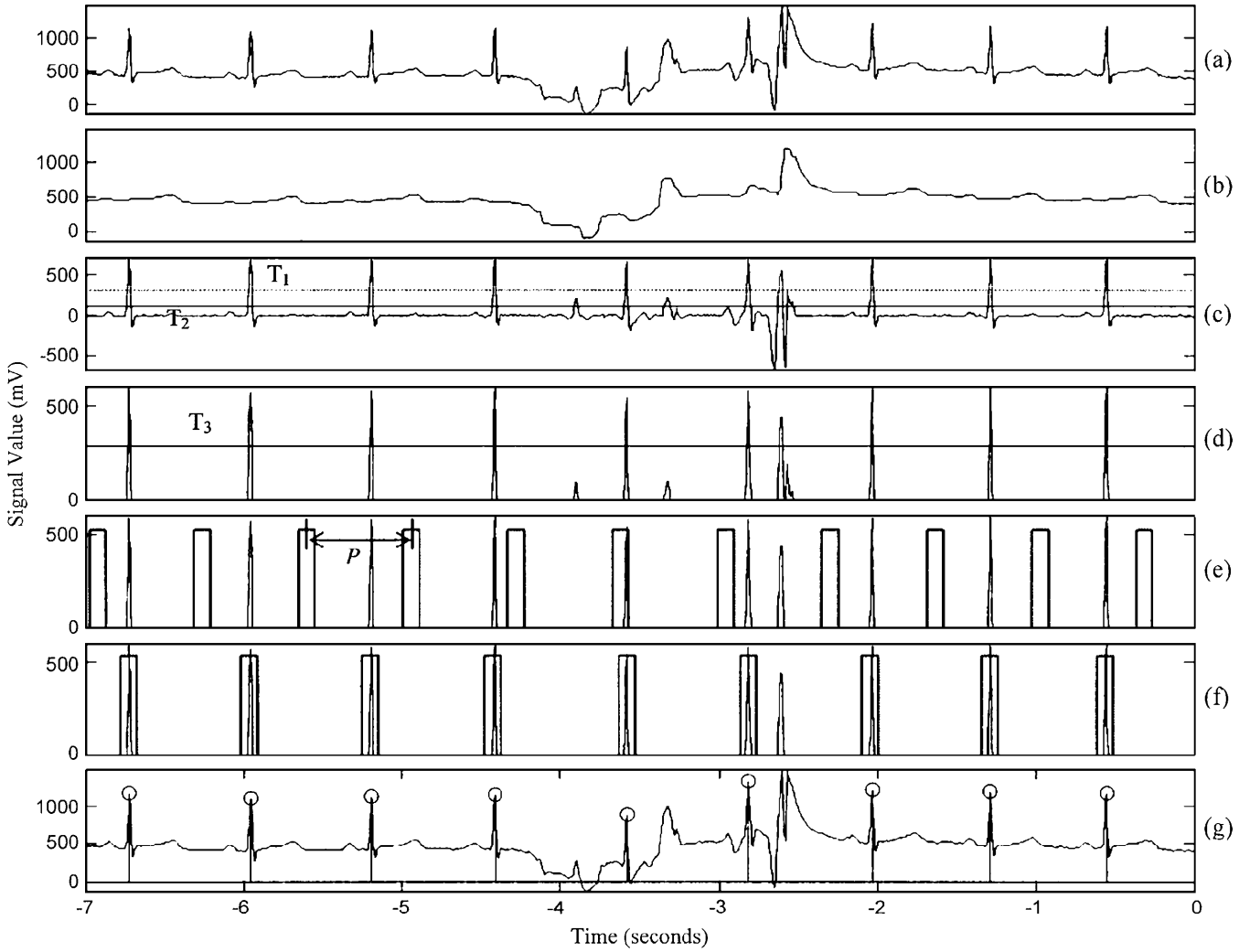

Figure 2. Illustration of the identification of heart beats by the ADAPIT algorithm. (*a*) Original 7-second ECG waveform segment. (*b*) Waveform after application of a median filter. (*c*) Difference of the original waveform in *a* minus the median-filtered waveform in *b*. The threshold T_1 defines the segment's baseline range $[-T_1, T_1]$ and the threshold T_2 provides a first cut on the lower limit of the peaks' magnitude. (*d*) The first estimates of the actual peaks and threshold T_3 (*horizontal line*) are used to eliminate small-magnitude spikes that clearly are not actual peaks. (*e*) String of markers with constant period *P*. (*f*) Best alignment between the actual peaks and markers, which is used to estimate heart rates. (*g*) The heart beats found by the ADAPIT algorithm are marked on the original electrocardiogram waveform.

Estimation from Electrocardiogram Waveforms

The ADAPIT algorithm computes an HRc at each time point (i.e., each second) t that a HRr is provided by the vital-signs monitor. This computation is performed based on a 7-second ECG waveform from time t-7 to t, which is approximately the same waveform length used by the vital-signs monitor,[10] to estimate one HRr. Figure 2 illustrates the four major steps of the algorithm to compute HRc at $t = 0$ (see Appendix 1 for additional technical details).

Step 1. ADAPIT applies a median filter (with a 55-ms window size) to the original 7-second waveform (Fig. 2a) and then subtracts the filtered signal (Fig. 2b) from the original one to yield the waveform in Figure 2c. This step de-trends the waveform, retains the amplitude of sharp R waves, and attenuates broad waves, such as the P wave and T wave.

Step 2. This step provides a first estimate of the actual peaks of the waveform through the sequential computation of two thresholds, T_1 and T_2. T_1, illustrated in Figure 2c, is taken as $2\sigma_1$, where σ_1 denotes the standard deviation of all data point values of the 7-second waveform and defines the segment's baseline range $[-T_1, T_1]$, from which the baseline standard deviation σ_2 is calculated. T_2, set to $3\sigma_2$, is used as a lower limit of the waveform amplitude for considering potential peaks. Peaks greater than T_2 are taken as the first estimate of the actual peaks (Fig. 2d).

Step 3. To eliminate small-amplitude spikes that clearly are not R waves, a threshold T_3 is defined as one half of the median amplitude of all peaks identified in Step 2 (Fig. 2d). All peaks less than T_3 are eliminated, as illustrated in Figure 2e.

Step 4. To determine actual R waves from the peaks retained in Step 3, strings of markers with period P (Fig. 2e) are iteratively generated and moved along the time line to align with the retained peaks. Through this iterative process, P is modified to range from lengths equivalent to HRs between 25 and 250 beats per minute (bpm). The string with the largest P aligned to the largest number of retained peaks is selected. Next, each unaligned marker of the selected string is allowed to move back and forth along the time line by as much as one half of P in an attempt to line up any unaligned peak (Fig. 2f). Finally, all aligned peaks, marked with circles on the original ECG waveform in Figure 2g, are assumed to be actual R waves. It should be noted that ADAPIT computes HRc based on all markers rather than the aligned peaks because an R wave could have been dropped during data collection or filtered out during the ADAPIT four-step process.

To verify ADAPIT's capability to filter out motion-induced artifacts and correctly compute HR of ambulatory trauma victims, we had a human expert visually estimate the HR of 80 seven-second, good-quality waveform samples from our database. Considering the human's estimations as the gold standard, we compare them against ADAPIT, HRr, and a well-established QRS-based detection program termed ecgpuwave.[12]

Figure 3 shows the difference between the algorithms' and the human's estimations for each of the 80 samples. The mean differences of ADAPIT, HRr, and ecgpuwave are, respectively, −0.62, 0.78, and 1.03 bpm, and the root mean square differences are 7.1, 5.1, and 7.1 bpm, respectively. These results indicate that in the process of filtering out noise, so as not to be counted as heart beats, ADAPIT tends to underestimate

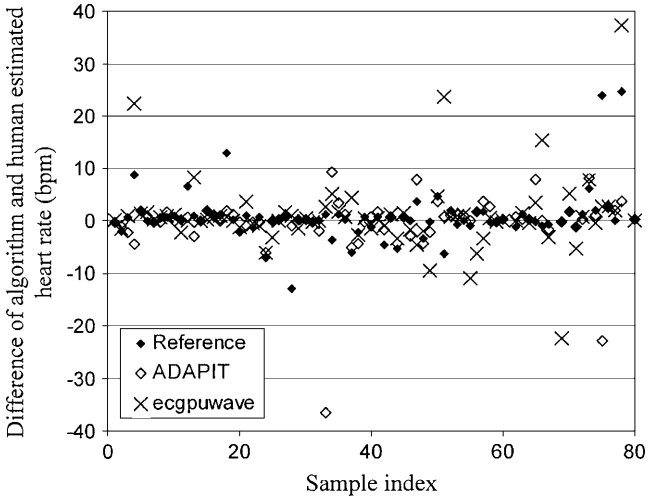

Figure 3. Difference in heart rates computed by three different algorithms (ADAPIT, reference heart rate [HRr], and ecgpuwave) and a human expert.

HRs, while the two other algorithms tend to overestimate them. This feature of ADAPIT is noticed, in particular, in waveforms with highly irregular rhythms (samples 33 and 76) and provides a lower bound estimate for the HRs that allows for a conservative consistency check (larger delta) between HRr and HRc.

Estimation from Photoplethysmogram Waveforms

ADAPIT employs the same four-step process with two small modifications in the estimation of PPG-derived HRc. First, in Step 1, the median filter window size is extended to 550 ms to preserve broad pulse waves and attenuate sharp dicrotic notches. Second, after the identification of peaks in Step 3, each peak is smoothed with a moving-average filter of window size equal to 110 ms. This additional filtering is needed to smooth out the broad and often distorted pulse waves and reduce the ambiguity in detecting the exact time of a heart beat, assumed to occur when the smoothed pulse wave reaches its maximum.

Waveform Qualification

This component of the approach implements our premise that the reliability of HRr is highly dependent on the quality of the underlying waveforms from which they are derived. A machine learning classifier, implemented by an SVM, automates the categorization of waveforms by attempting to mimic the performance of human experts who rely on visual inspection and the application of some implicit or explicit rules of thumb. A classifier "learns" these rules by finding coefficients that optimize the "correlations" between a set of waveform-extracted features and waveform quality obtained from manually categorized waveform samples.

Figure 4 illustrates the four steps in the development of a machine-learning classifier: (1) manually categorize sample waveform segments, (2) define candidate waveform features that distinguish good/bad waveforms, (3) select the most informative features, and (4) train and test the classifier. Once trained and given input features, the classifier categorizes waveform segments as being good or bad.

Manual Waveform Categorization

To develop the SVM classifier, human experts visually examined and categorized 7-second waveform segments for 362

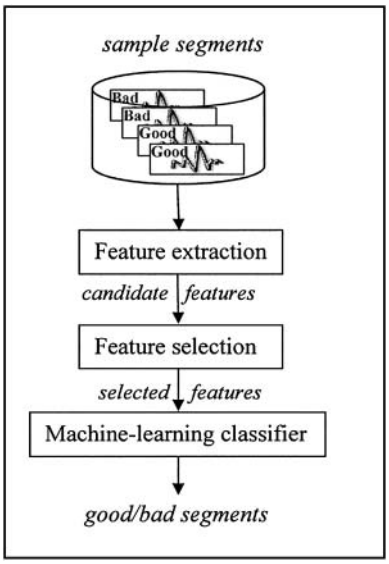

Figure 4. The development of machine-learning classifiers requires (1) manual categorization of good/bad waveform-segment samples, (2) definition and extraction of candidate waveform features, (3) selection of the most discriminatory features, and (4) training and testing of the machine-learning classifier. Once trained and given input features, the classifier categorizes waveform segments as being good or bad.

ECG samples and 388 PPG samples randomly selected from different patients. Of these, 194/168 ECG samples and 180/208 PPG samples were categorized as good/bad based on the following rules:

> *An ECG segment is ranked as bad (suboptimal) if more than one expected R wave is not observed or if the R wave is indistinguishable from noisy peaks. Otherwise it is ranked as good. A PPG segment is ranked as bad (suboptimal) if more than one expected pulse wave is not observed or if any one pulse wave peak cannot be distinguished from a dicrotic notch. Otherwise it is ranked as good.*

These rules express the hypothesis that if more than one heart beat signal in a 7-second waveform segment is ambiguous, the HR calculated from such segment may be inaccurately extrapolated. The rules are conservative by design so that the inductively constructed classifiers are equally conservative and attempt to ensure that even if the classifier produces occasional false good waveform evaluations, those false good waveforms will still be of sufficient quality for estimation of HRs.

Candidate Waveform Features

A key phase in the development of machine-learning classifiers involves the definition and extraction of candidate features that can be used as class discriminators. For the characterization of waveforms as good or bad, we define three features in the frequency domain from ECG waveforms and three features in the time domain from ECG and PPG waveforms. Their definitions are presented in Appendix 2.

Similar to the ADAPIT algorithm, we extract features from 7-second waveform segments that immediately precede each HRr we wish to qualify. The three frequency-domain features, high-frequency energy (HFE), low-frequency energy (LFE), and their ratio LFE/HFE, are obtained by applying the discrete-time fast Fourier transform[13] to the ECG time-series data. These features are designed to exclude ECG frequency components that are associated with a QRS complex, while capturing high- and low-frequency component characteristics that may be attributed to noise and baseline drifts and shifts.

The first time-domain feature is the fraction of aligned waves FW, which provides a measure of temporal regularity of potential heart beat signals. The second time-domain feature is a specific signal-to-noise ratio SN, which provides a measure of the distinctiveness of potential heart beat signals above the baseline. The pulse-wave variability (PV), extracted from PPG waveform segments, is the third time-domain feature and provides a measure of the variability of the time interval between two adjacent pulse waves.

Feature Selection

The goal of automatic feature selection is to choose and retain a subset of salient features from the original list of candidate features such that the process of pattern discovery by the machine-learning classifier is implemented in a reduced space without degrading its performance. The underlying philosophy is to retain features that can clearly characterize or discriminate the quality of the waveforms and eliminate features that are redundant, and hence, do not contribute additional information. Here, we employ information entropy[14,15] as a measure of discriminatory power of the features. The most discriminatory (informative) feature has the lowest entropy.

Our previously developed Rule Generator (RG) program[14,15] is used to compute entropies of candidate ECG and PPG waveform features. The RG program also defines patterns formed by these features and populated by the previously characterized samples to discriminate good/bad waveforms. The features that characterize the most discriminatory patterns, defined as the patterns that discriminate the largest number of samples, are selected as the most informative. Through this procedure, we find that HFE, FW, and SN are the most discriminatory features for ECG waveform classification and that FW and PV are the most informative features for PPG waveform classification.

Support Vector Machine Classifier

In this study, we employ our previously developed version of an SVM algorithm[16] to classify ECG and PPG waveforms. The SVM, a recently proposed supervised machine-learning algorithm,[7] has been shown to be an effective classifier in a wide variety of applications, including the categorization of ECG data.[17–20] As a supervised-learning algorithm, the development (or "training") of an SVM requires a set of input/output training samples, where the inputs consist of a list of discriminatory features, such as the three ECG features and two PPG features selected in the previous section, and the outputs consist of labeled binary classes, good and bad. Once trained to implicitly "learn" the "rules" embedded in the training samples, given the values of the input features, extracted from a waveform segment that we wish to classify, the SVM automatically categorizes the segment as good or bad. An in-depth description of SVMs can be found in Vapnik.[7]

We trained and tested an SVM classifier through a cross-validation procedure employing the manually categorized

waveform samples (362 ECG and 388 PPG), where at each of 200 cross-validation repetitions 70%–30% of the samples were used for training-testing the classifier. For all simulations, we used the same SVM model with a linear kernel function and at the end of the 200 simulations computed average performance measures, such as sensitivity and specificity, for the classifier. We did not attempt to optimize the SVM classifier. Classifier sensitivity provides a measure of the incorrectly classified (i.e., missed) bad waveform segments, whereas classifier specificity provides a measure of false hits, i.e., the fraction of good segments classified as bad.

Averaged over the 200 cross-validation repetitions, the SVM yielded 93% sensitivity and 96% specificity for the ECG waveforms, and 91% sensitivity and 88% specificity for the PPG waveforms. The slightly worse performance for the PPG waveforms reflects the increased difficulty in classifying this waveform due to a lack of more distinct characteristics of its profile. Considering that very conservative rules were used to categorize bad waveform segments, the 93% and 91% sensitivity result can be taken as conservative estimates of the classifier's ability to correctly categorize truly bad waveforms. Indeed, for those segments assigned bad quality by human experts that the SVM misclassified, our visual estimates of the HRs are compared and agree with those HRr provided by the vital-signs monitor. This indicates that misclassification of bad waveforms by the classifier may still lead to correct estimation of HRc.

Quality Index Determination

The final component of the algorithm is the numerical qualification of the ECG HRr and PPG HRr provided by the vital-signs monitor. The qualification combines the independent estimation of ECG HRc and PPG HRc from redundant sources, their reference values HRr, and the results of the waveform SVM classifier to assign a QI that concisely expresses the reliability of each HRr provided by a vital-signs monitor. A QI of 3 indicates that both ECG HRr and the PPG HRr are "highly" reliable, a QI of 2 indicates that ECG HRr is "fairly" reliable, a QI of 1 indicates that PPG HRr is "somewhat" reliable, and a QI of 0 indicates that neither HRr is reliable. The qualification algorithm assumes that good-quality HRr should come from high-quality waveforms and should be consistent with our independently calculated HRc. Another implicit assumption is that the data originate from live patients.

Table 1 describes the rules used to generate the four QIs. The entries in the second and third columns indicate the quality of the two waveforms. The entries in the fourth and fifth columns indicate whether the ECG HRr and PPG HRr, respectively, are consistent with their corresponding HR computed by ADAPIT. HRr and HRc are consistent with each other when the discrepancy $\epsilon_1 < 5\%$, with

$$\epsilon_1 = \frac{|\mathrm{HRr} - \mathrm{HRc}|}{0.5\,(\mathrm{HRr} + \mathrm{HRc})}. \quad (1)$$

The entries in the last column indicate whether all four HRs are consistent. Consistency is achieved when the discrepancy $\epsilon_2 < 10\%$, where ϵ_2 is defined as the ratio of the largest absolute difference among the six possible pairwise comparisons and the average HR over the four values. The table entries denoted with a dash indicate that consistency is not required.

Table 1 ■ Rules Describing the Four Quality Indices

	Waveform Quality		Heart Rate Consistency		
Quality Index	ECG	PPG	ECG	PPG	All Four
3	Good	Good	Yes	Yes	Yes
2	Good	Good	Yes	No	–
	Good	Bad	Yes	–	–
1	Good	Good	No	Yes	–
	Bad	Good	–	Yes	–
0	All other cases that do not match conditions above				

– represents that consistency is not required.
ECG = electrocardiogram; PPG = photoplethysmogram.

For example, a QI of 3 is inferred when both ECG and PPG waveforms are classified by the SVM as having good quality, ECG HRr and PPG HRr are consistent with their corresponding ADAPIT-computed HR, and all four HRs are consistent with each other. This rule infers that the two HRr, which originate from redundant sources (ECG vs. PPG) and are independently calculated (vital-signs monitor vs. ADAPIT), are in agreement and, with high confidence, correctly represent the actual HR. In this case, either ECG HRr or PPG HRr could be used to represent the actual HR. A QI of 2 is inferred in two possible scenarios. First, when both ECG and PPG waveforms have good quality and ECG HRr is consistent with ECG HRc. Second, when the ECG waveform has good quality, the PPG waveform has bad quality, and ECG HRr is consistent with ECG HRc. In essence, this rule expresses the situation where the ECG provides reliable information and ECG HRr alone should be used to represent the actual HR. Conversely, a QI of 1 indicates the situation where the PPG provides reliable information and PPG HRr alone should be used. Note that for equivalent requirements in the rules for a QI of 1 and QI of 2, we assign a higher confidence for the ECG HRr. This is to reflect that the ECG waveform is generally more reliable, possesses distinctive features that facilitate characterization, and is the one often used as the gold standard for HR computation. Finally, by exclusion, when none of the above conditions are satisfied, we assign a QI of 0 to indicate that neither HRr should be used to represent the actual HR. In the absence of one of the waveforms or their derived HRs, the decision–logic algorithm still provides a QI by assuming that the absent signal is present but possesses poor quality.

In summary, the highest reliability (QI = 3) is achieved only when the redundantly measured and independently computed HRs corroborate each other, while the lower reliability levels (QI = 2 and QI = 1) only require agreement between independent computations from the same source.

Results and Discussion

We demonstrate the performance of our algorithm through the analysis of two examples illustrated in Figures 5 and 6. Figure 5 shows an example where the ECG waveform is mostly noisy over a 40-second interval and the PPG waveform is partially noisy. The thick horizontal bars on the top panel of the figure indicate segments of bad-quality waveforms determined by the SVM classifier. The middle panel shows a deviance of the ECG HRr that is considerably larger than the other three HRs, which, in contrast, are more consistent. This suggests that noise spikes counted by the vital-signs monitor as heart beats are filtered out by ADAPIT in spite of

the bad quality of the waveform. Based on our decision logic, the bad quality of the ECG waveform for most of the segment limits the value of the QI, illustrated in the lower panel, to ≤1.

The lack of distinguishable pulse waves in the PPG waveform between 106 and 116 seconds causes our algorithm to provide a lower estimate for PPG HRc (middle panel, Fig. 5). Surprisingly, the PPG HRr remains unaffected during this time interval. A possible explanation is that the vital-signs monitor outputs extrapolated HRr values based on previous records when it cannot detect pulse waves. The use of such HR is inappropriate, as it does not correctly reflect the quality of the PPG waveform. This discrepancy between PPG HRr and PPG HRc is caught by our method, which correctly assigns a QI of 0 for this time segment where both waveforms are judged to have bad quality and neither of the two HRr provides a reliable estimate. The algorithm provides a QI of 1 for the remaining time intervals, where the PPG waveform possesses good quality and there is good agreement (<5% discrepancy) between reference and ADAPIT-calculated PPG HRs.

The data in Figure 6 provide another example in which the use of redundant signal sources and independent computation of HR provide a powerful method to assess HR reliability. It shows a case where the HRr are clearly overestimated in spite of the pristine nature of the waveforms. The ECG HRr and PPG HRr are about 15 bpm greater than their ADAPIT counterparts, which are around 143 bpm. Manual counting of the heart beats confirms the accuracy of the ADAPIT calculation. This example suggests that the reliability of the HRr cannot be inferred solely through the determination of waveform quality. A discrepancy >5% between ECG HRr and ECG HRc and between PPG HRr and PPG HRc causes the algorithm to infer a QI of 0 for the entire interval because neither of the two HRr is reliable.

We evaluate our methodology against 173 randomly selected samples reviewed by two human experts. Each sample consists of one ECG HRr, one PPG HRr, and their associated, simultaneously recorded 7-second waveforms. The experts are asked to duplicate the decision logic in Table 1 by visually classifying the quality of the waveforms using the rules described in the "Manual Waveform Categorization" section above to estimate ECG and PPG HRs and to provide a QI for each data sample. Of the 173 samples, 158 (91%) were assigned the same QI by the two experts. We use these consensus samples as the gold standard against which we tested our methodology.

Table 2 compares the QI assignment of our method against the human experts for the 158 samples. Our method agrees with the experts' QI assignment in 135 (85%) of the samples (shaded diagonal entries in the table), overestimates the experts in 13 (8%) of the samples (entries above the diagonal),

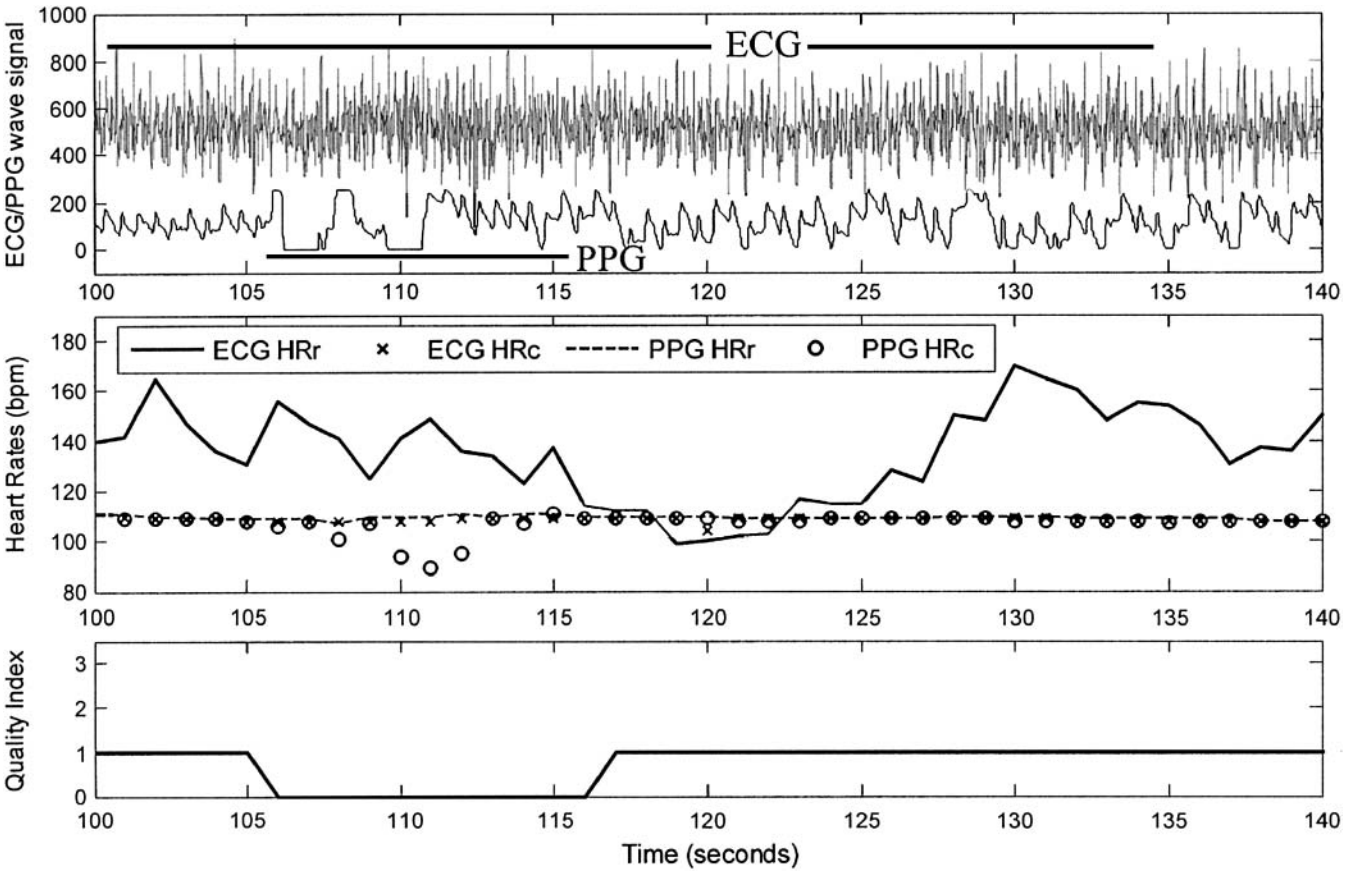

Figure 5. Validation of reference heart rates (HRr) from low-quality electrocardiogram (ECG) and photoplethysmogram (PPG) waveforms (patient 23). (*Top*) The thick bars indicate the bad-quality regions in the ECG and PPG waveforms, as determined by the waveform qualification algorithm. (*Middle*) ECG HRr is considerably larger than the other HRs and the PPG HRr is almost constant during times where the PPG waveform has bad quality. (*Bottom*) Quality index (QI) values, which, due to the bad-quality of the ECG waveform, never reach values >1.

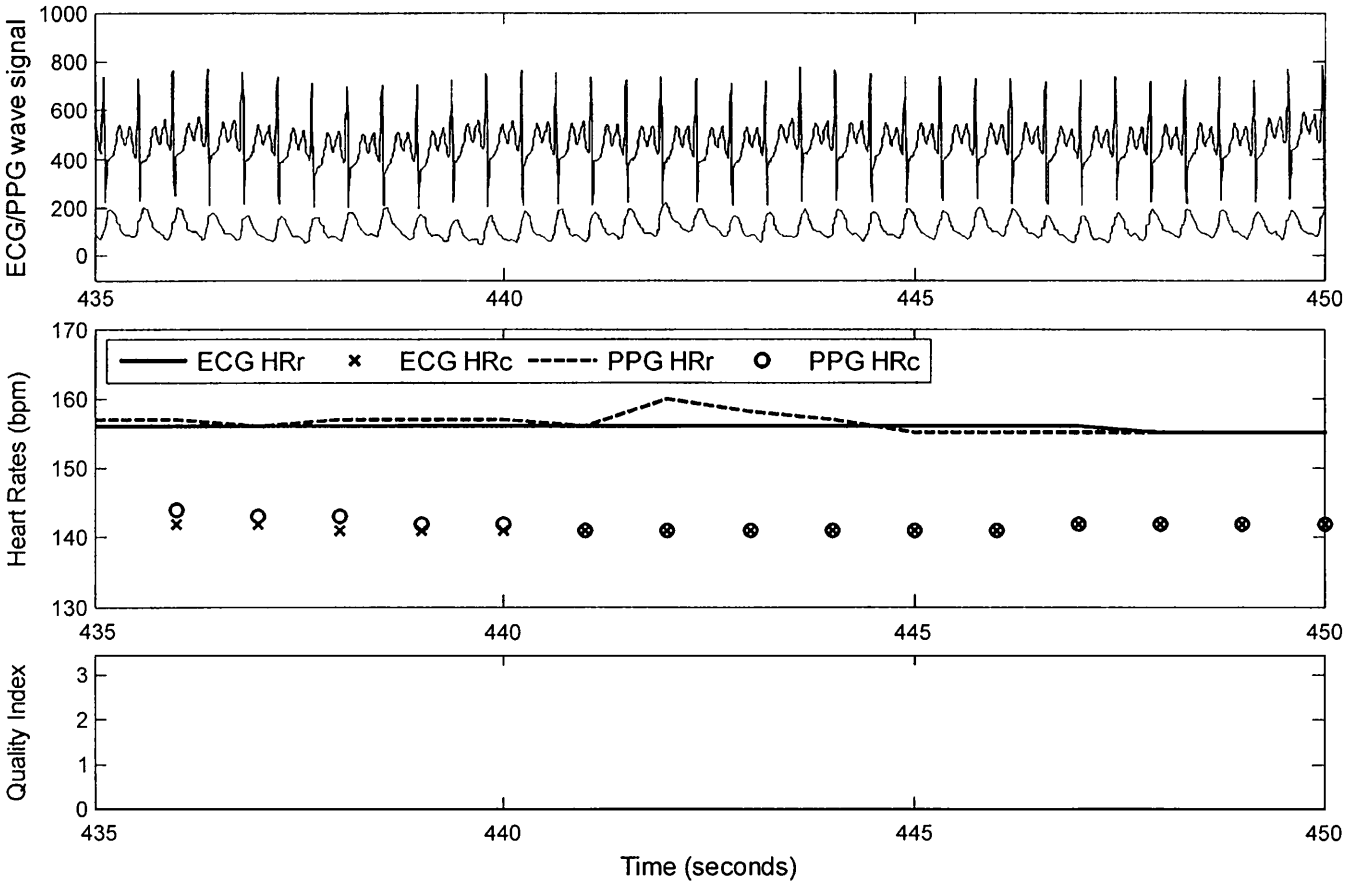

Figure 6. Incorrect electrocardiogram reference heart rate and photoplethysmogram reference heart rate computed from pristine waveforms (patient 128).

and underestimates the experts in 10 (6%) of the samples (entries below the diagonal). The fraction of exact matches (85%) is slightly lower than those observed between the two human experts (91%) in the original 173 samples, and in 92% of the cases, the method's HR qualifications either match or are more conservative than the human's qualifications.

It is important to note that although we designed the algorithm to be conservative, so that when it rates a HRr as reliable, the user would have confidence that the HRr is trustworthy and can be used to develop decision–support algorithms, the results in Table 2 seem to contradict our intent, as the method's overestimation rate (8%) is slightly higher than its underestimation rate (6%). This results from the method's categorization of bad-quality waveforms as good. However, because the SVM waveform classifier was developed based on very conservatively categorized waveforms, when it occasionally produces false good waveform evaluations, those false good waveforms are generally still of sufficient quality for accurate HR estimation. Indeed, when we further examined the 13 overestimated samples, we found that in at least nine samples, the experts' estimated HRs agree with the HRr within 5%, indicating that it is possible to obtain accurate HR from suboptimally classified waveforms. This suggests that the method's actual overestimation rate could be as low as 3% (4/158).

We also evaluate the sensitivity (the performance measure of interest) of the two components of the method, the SVM waveform classifier and the ADAPIT HRc, against the experts' evaluation for the same 158 samples (Table 3). ECG HRr and PPG HRr are separately evaluated. The human experts evaluated a sample as good when the waveform was of good quality and the human's estimated HR was within 5% of the corresponding HRr. In the second column, the waveform quality inferred by the SVM is used as the

Table 2 ■ Comparision of Human Expert Versus Algorithm Assignment of Heart Rate Quality for 158 Samples

	QI	Program			
QI		**0**	**1**	**2**	**3**
Human	**0**	78	1	6	0
	1	2	2	0	0
	2	5	1	42	6
	3	1	0	1	13

Table 3 ■ Comparison of the Contribution of the Two Components of the Heart Rate Qualification Algorithm

	Waveform Quality		HR Consistency		Both Components	
	Bad	Good	Bad	Good	Bad	Good
			ECG HRr			
Human						
Bad cases	77	12	29	60	83	6
Good cases	4	65	2	67	7	62
Sensitivity	87% (77/89)		33% (29/89)		93% (83/89)	
			PPG HRr			
Human						
Bad cases	124	15	68	71	131	8
Good cases	2	17	2	17	4	15
Sensitivity	89% (124/139)		49% (68/139)		94% (131/139)	

HR = heart rate; ECG = electrocardiogram; HRr = reference heart rate; PPG = photoplethysmogram.

indicator of the HRr quality. In the third column, consistency (i.e., agreement within 5%) between HRc and HRr is used as the sole indicator of the HRr quality, and in the last column, both waveform quality and HR consistency are used to infer the HRr quality. The results indicate that sensitivity is highest when we employ both components of the method. Waveform quality alone provides a slightly lower sensitivity, as in a few cases inconsistent HRr are associated with waveforms categorized as good by the SVM classifier. Consistency between HRc and HRr provide considerably lower sensitivity, as although the two derived HRs are in agreement, due to the conservative nature of the rules used to categorize good-quality waveforms, the human experts deem their corresponding waveforms to be bad.

Our method was also applied to qualify the records of the 726 trauma patients deposited in our Physiology Analysis System,[11] where, on average, each record consists of approximately 25-minutes of time-series data per variable. Table 4 summarizes the qualification results for the HRr and the 7-second waveform segment, one segment at a time, for the entire time-series record. Approximately half of the ECG waveforms (48%) and a third of the PPG waveforms (30%) have good quality. A significant portion of the bad-quality waveform is attributed to motion artifacts that occur during patient transport to the trauma center, especially for the sickest patients requiring life-saving interventions. This is particularly true for PPG waveforms that are easily degraded by movement or slippage of the sensor clipped on the patient's finger. Also of note is the small incidence of unreliable HRs with associated good-quality waveforms, which is about 4% for both ECG HRr and PPG HRr. This observation is corroborated by the high sensitivity of the waveform qualification in the second column of Table 3. This suggests that, for high-quality waveforms, there is good agreement between HRs estimated by the vital-signs monitor and ADAPIT, and reflects the very stringent rules that define good-quality waveforms. If we relax the rules, there will be an increase in the fraction of waveforms qualified as good, increasing the fraction of unreliable HRs associated with these good-quality waveforms and shifting the classification burden to the consistency check between HRc and HRr. We also find that 21% of the HRr have the highest possible quality (QI of 3) and that very rarely

Table 4 ■ Percentage of Good-Quality Time-series Data in the Database as Assessed by the Validation Algorithm

Condition	% of Time-series Data
Good-quality ECG waveforms	48
Reliable ECG HRr (QI = 3 or QI = 2)	44
Good quality PPG waveforms	30
Reliable PPG HRr (QI = 3 or QI = 1)	26
Very reliable HRr (QI = 3)	21
Criteria for QI = 3 are met, except that all four heart rates do not match within 10%	0.05

ECG = electrocardiogram; HRr = reference heart rate; QI = quality index; PPG = photoplethysmogram.

(0.05%) the four computed HRs do not match within 10% when the first four criteria for a QI of 3 are satisfied.

Conclusions

Validated HR data allow investigators to select high-quality records for data mining, diagnosis and prognosis of trauma patients, and the development of advanced trauma scoring methods. This paper presents an effective, systematic, and automated method for validating vital-signs monitor HRr derived from ECG and PPG waveforms, where each sampled HRr is assigned a numerical QI that concisely expresses its reliability. The method exploits the physical redundancy provided by these two distinct signal sources as well as the use of independent methods for separately estimating HRs from these sources. Moreover, it can be readily modified if only one source of waveform data is available and it is not tied to any specific vital-signs monitor. The method especially focuses on HR data collected during transport of trauma patients when ECG and PPG waveforms are more likely to be deteriorated by movement artifacts.

The reliability of ECG and PPG HRr is highly dependent on the quality of the underlying waveforms from which the HRs are derived. It is possible to develop a machine-learning classifier to distinguish bad-quality waveforms, so that the associated HRs are also classified as bad (or at least of questionable) quality. Our study suggests that SVM classifiers using features extracted from the time domain and the frequency domain are capable of assessing the quality of ECG and PPG waveforms as good or bad with sensitivity and specificity around 90%. In our study, very stringent rules are applied by human experts to characterize waveform quality used to develop the SVM classifier, which tends to ensure a high HRr reliability when waveforms are assessed as having good quality. This leads to conservative HR qualification results, which are desired for the development of data-driven decision-support algorithms and models. However, extreme conservatism may eliminate usable waveforms and constrain the size of the data set, precluding the development of data-driven algorithms.

The quality of HRr is not always tied to the quality of the underlying waveforms (Fig. 6). The ability to compare the reference HR with one calculated by an independent method provides another level of assurance, with agreement between the two suggesting good-quality data. Randomly imposed noise spikes are often observed in ECG waveforms collected during transport of trauma patients, and, if not properly filtered out, these spikes may be counted as heart beats. The

ADAPIT algorithm presented here is designed to identify real heart beats from randomly imposed noise by assuming quasi-constant intervals between real heart beats. While this assumption has been effective in filtering out random noise, it also precludes the correct estimation of HR in settings of highly irregular rhythms, which is rarely observed in our data set of trauma victims. The algorithm, however, does allow for sufficient variation between adjacent intervals such that modest physiologic HR variability is not falsely identified as noise.

Our approach splits the qualification burden between two tasks, the assessment of waveform quality and the comparison of HRs through redundant and independent means, where the split is regulated by the stringency of the rules used to develop the classifier. For example, the very stringent rules used in this study shift the burden toward the waveform qualification. Our results show that when ECG and PPG waveforms are categorized as having good quality, 90% of the associated HRr are deemed reliable. This implies that the comparison of HRs is responsible for filtering out only 10% of the data, and indicates that, for good-quality waveforms, the vital-signs monitor HR predictions compare well with ours. However, as we relax the classifier rules, the burden shifts away from the waveform qualification and toward the HR comparisons, as a larger percentage of the data will be filtered out through HR comparisons. Our decision logic combines these two elements in assessing the reliability of HRs recorded by vital-signs monitors and succinctly expresses them through a QI.

References ■

1. Hoyt R, Reifman J, Coster T, Buller M. Combat medic informatics: present and future. Proc AMIA Annu Symp. 2002, pp. 335–9.
2. Senkowski CK, McKenney MG. Trauma scoring system: a review. J Am Coll Surg. 1999;189:491–503.
3. Koehler JJ, Baer LJ, Malafa SA, Meindertsma MA, Navitskas NR, Huizenga JE. Prehospital Index: a scoring system for field triage of trauma victims. Ann Emerg Med. 1986;5:178–82.
4. Nagin VA, Selishchev SV. Implementation of algorithms for identification of QRS-complexes in real-time ECG systems. Biomed Eng. 2001;35:304–9.
5. Neves JN, Owall V, Sornmo L. QRS detection for pacemakers in a noisy environment using a time lagged artificial neural network. Circuits Syst. 2001;3:596–9.
6. Kohler BU, Henning C, Orglmeister R. The principles of software QRS detection. IEEE Eng Med Biol Mag. 2002;21:42–57.
7. Vapnik V. Statistical learning theory. New York: John Wiley & Sons; 1998.
8. Holcomb JB, Salinas J, McManus JM, Miller CC, Cooke WH, Convertino VA. Manual vital-signs reliably predict need for life saving interventions in trauma patients. J Trauma Injury Infect Crit Care. 2005;59:821–8.
9. Cooke WH, Salinas J, Convertino VA, Ludwig DA, Hinds D, Duke JH, Moore FA, Holcomb JB. Heart rate variability and its association with mortality in pre-hospital trauma patients: a pilot study. J Trauma Injury Infect Crit Care. 2006;60:363–70.
10. Propaq Encore Reference Guide. Beaverton, OR: Welch Allyn; 1998, Available at: http://www.monitoring.welchallyn.com/products/portable/propaqencore.asp. Accessed December 2004.
11. Reifman J, Gunawardena J, Liu Z. Physiology analysis system. Presented at the 4th IEEE International Symposium on Signal Processing and Information Technology, December 18–21, 2004, Rome, Italy.
12. Jané R, Blasi A, Garcìa J, Laguna P. Evaluation of an automatic threshold based detector of waveform limits in Holter ECG with the QT database. Comput Cardiol. 1997;24:295–8, Available at: http://www.physionet.org/physiobank/database/qtdb/eval/. Accessed December 2004.
13. Signal processing toolbox for use with MATLAB. Natick, MA: The MathWorks; 2002.
14. Reifman J, Lee JC. Reactor diagnostic rule generation through statistical pattern recognition. Nucl Sci Eng. 1991;107:291–314.
15. Reifman J, Lee JC. Comparison of two inductive learning methods: a case study in failed fuel identification. Presented at the 8th Power Plant Dynamics, Control and Testing Symposium, May 27–29, 1992; Knoxville, TN.
16. Yu C, Zavaljevski N. ActiveSVM user's manual. Argonne, IL: National Laboratory; 2003.
17. Jankowski S, Oreziak A. Learning system for computer-aided ECG analysis based on support vector machines. Int J Bioelectromagnet. 2003;5:175–6.
18. Osowski S, Hoai LT, Markiewicz T. Support vector machine-based expert system for reliable heartbeat recognition. IEEE Trans Biomed Eng. 2004;51:582–9.
19. Li P, Chan KL, Chan YW. A mixed SVM-based hierarchical learning approach for abnormal ECG beat recognition. Presented at the 1st International Conference on Bioengineering, September 8–10, 2004, Singapore.
20. Zimmerman MW, Povinelli RJ. On improving the classification of myocardial ischemia using Holter ECG data. Comput Cardiol. 2004;31:377–80.

Appendix 1

Heart Rate Estimation with the ADAPIT Algorithm

ADAPIT Estimation of Electrocardiogram-Derived Heart Rate

Step 1. Given the 7-second ECG waveform segment (Fig. 2a), ADAPIT first applies a median filter (Fig. 2b) to remove baseline drifts, preserve R waves, and attenuate broad waves, such as the P wave and the T wave. Next, it subtracts the median-filtered signal from the original waveform (the waveform in Figure 2c) that serves as the starting point for the next step. The selection of a correct window size for the median filter is critical for preserving the sharp QRS complex and attenuating broad waves in the subtracted signal. A rule of thumb is to choose a window size of length close to the average width of typical R waves, which generally ranges from 40 to 100 ms.[6] Here, we use 55 ms, equivalent to ten sampled points at 182 Hz.

Step 2. At the second step, ADAPIT provides a first-estimate of the actual peaks of the waveform, i.e., the R waves in the case of ECG, through the sequential computation of two thresholds. The first threshold T_1 (Fig. 2c) is taken as $2\sigma_1$, with σ_1 denoting the standard deviation of all data point values that make up the waveform over the 7-second segment. The waveform values in the range $[-T_1, T_1]$ around zero define the segment's baseline range from which the baseline standard deviation σ_2 is calculated. The second threshold T_2 is set to $3\sigma_2$. Those peaks with magnitude greater than T_2 (i.e., the peaks in Figure 2d) are taken as the first estimates of the actual peaks and are used as the starting point of the next step.

Step 3. At the third step, ADAPIT employs another threshold, T_3, to eliminate small-magnitude spikes that clearly are not actual peaks, i.e., spikes that are not part of a QRS complex. T_3 (Fig. 2d) is set at one half of the median magnitude of all peaks identified in Step 2 over the 7-second segment. The N_p peaks of magnitude greater than T_3 ($N_p = 10$, in this case) are kept and taken to the last step of the algorithm.

Step 4. In this last step, ADAPIT uses an adaptive iterative approach to discard ambiguous spikes, identify QRS complexes, and compute the HR at time zero. The iterative approach starts by first generating a string of N_r markers of constant period P, with P initially set to 240 ms, corresponding to an assumed maximum HR of 250 bpm and $N_r = 29$ markers in the 7-second segment. Figure 2e shows the case for $N_r = 11$. Next, the string of N_r markers is allowed to move along the time line in order to optimize their alignment with the N_p peaks identified in Step 3. This is achieved when the fraction of aligned waves,

$$\mathrm{FW} = \frac{N_a}{N_p + N_r - N_a}, \tag{A1}$$

is maximized, where N_a denotes the number of aligned waves. FW attains its maximum of 1.0 when $N_p = N_r = N_a$ and its minimum of 0.0 when N_a is zero. Next, the period P of the markers is increased by 11 ms, and the alignment is optimized by finding the maximum FW corresponding to the updated P. This process is repeated until P reaches its maximum value of 2,396 ms, corresponding to an assumed minimum HR of 25 bpm. Then, the string of markers with period P^* corresponding to the maximum FW over the range of P values [240, 2,396] is selected, and each unaligned marker of the string is allowed to move back and forth along the time line by as much as one half of P^* in an attempt to line up the misaligned peaks (Fig. 2f), after which FW is recomputed. This adjustment allows the algorithm to account for expected heart beat variations.

Figure 2g shows the heart beats found by ADAPIT, which are marked with circles on the original ECG waveform. ADAPIT computes HR based on the number of markers N_r in the segment, rather than the number of aligned peaks N_a. This avoids potential errors, should an actual QRS complex be dropped in the data collection process or filtered out during the ADAPIT four-step process.

ADAPIT Estimation of PPG-Derived Heart Rate

ADAPIT employs the same four-step process with two small modifications in the estimation of PPG-derived HRs. First, in step 1, the window size of the medium filter is extended to 550 ms, equivalent to 50 sampled points at about 91 Hz. This widening preserves the broad pulse waves associated with the heart beats and attenuates the sharp dicrotic notches. Second, after the algorithm identifies the N_p peaks of magnitude greater than T_3 in step 3, each peak is smoothed with a moving-average filter of window size equal to 110 ms. This additional filtering is needed to smooth out the broad and often distorted pulse waves and reduce the ambiguity in detecting the exact time of a heart beat assumed to occur when the smoothed pulse wave reaches its maximum.

Appendix 2

Definitions of Candidate Waveform Features

The three frequency-domain features are obtained by applying the discrete-time fast Fourier transform to the ECG time-series data to compute the power spectral density PSD (f), which describes how the power of a time series is distributed as a function of frequency f.[13] Accordingly, we define the low-frequency energy (LFE) feature and the high-frequency energy (HFE) feature,

$$\mathrm{LFE} = \int_0^{f_L} PSD(f)\ df \quad \text{and} \tag{A2}$$

$$\mathrm{HFE} = \int_{f_H}^{f_S} PSD(f)\ df, \tag{A3}$$

by integrating the PSD(f) over the low- and high-frequency ranges, respectively, where the low-frequency cutoff f_L is set to 1 Hz and the cutoff for the high frequencies is set to f_H = 40 Hz and f_S = 91 Hz (corresponding to half of the ECG sampling frequency).

These features are designed to exclude ECG frequency components that are associated with a QRS complex, which typically range from about 10 to 25 Hz,[6] and capture low- and high-frequency components associated with potential artifacts. The HFE captures high-frequency noise and the LFE characterizes baseline drifts and shifts. For instance, a large LFE is characteristic of bad-quality waveforms. The third

frequency-domain feature is defined by the ratio LFE/HFE, which is intended to characterize the rate of attenuation of the PSD from low to high frequencies.

The first time-domain feature is the fraction of aligned waves FW defined in Equation A1 at the end of step 4 of the ADAPIT HR estimation. FW provides a measure of regularity of the frequency of heart beats computed from ECG and PPG waveforms. The SN of a waveform is the second time-domain feature. Based on statistics computed in step 2 of the ADAPIT algorithm, SN is defined as the log ratio of the median value μ over the waveform data points above the threshold T_2 corresponding to the points making up the upper region of the QRS complex (or pulse-waves in PPG waveforms), and the standard deviation σ_2 of the waveform values that define the ''noise'' component of the waveform around the baseline range $[-T_1, T_1]$. Thus,

$$\mathrm{SN} = \frac{1}{2}\log\frac{\mu}{\sigma^2}, \tag{A4}$$

is greater than zero, as $\mu > \sigma_2$, and attains larger values as the QRS complexes (or pulse waves in PPG waveforms) become more distinguishable from the baseline.

The third time-series feature provides a measure of the variability of the time interval or period P between two adjacent pulse waves in a 7-second PPG segment. Disregarding disease-driven HR arrhythmia, we observed that pulse-wave variability PV can be used as a waveform discriminator, with good/bad waveforms having small/large values of PV. Accordingly, we define

$$\mathrm{PV} = \frac{\sigma HRc}{\overline{HRc}} \tag{A5}$$

where $\overline{HRc}$ and σHRc denote the mean value and standard deviation, respectively, of the ADAPIT-computed HR over the 7-second PPG segment, with HRc taken as $60/P$ with period P expressed in seconds.