Published in final edited form as:

J Electrocardiol. 2018 ; 51(1): 68–73. doi:10.1016/j.jelectrocard.2017.08.030.

A novel ECG detector performance metric and its relationship with missing and false heart rate limit alarms

Chathuri Daluwatte, PhD^{a,*}, Jose Vicente, PhD^b, Lorian Galeotti, PhD^c, Lars Johannesen, PhD^d, David G. Strauss, MD, PhD^d, and Christopher G. Scully, PhD^a

^aOffice of Science and Engineering Laboratories, CDRH, US FDA, Silver Spring, MD, USA

^bOffice of New Drugs, CDER, US FDA, Silver Spring, MD, USA

^cOffice of Device Evaluation, CDRH, US FDA, Silver Spring, MD, USA

^dOffice of Clinical Pharmacology, CDER, US FDA, Silver Spring, MD, USA

Abstract

Purpose—Performance of ECG beat detectors is traditionally assessed on long intervals (e.g.: 30 min), but only incorrect detections within a short interval (e.g.: 10 s) may cause incorrect (i.e., missed + false) heart rate limit alarms (tachycardia and bradycardia). We propose a novel performance metric based on distribution of incorrect beat detection over a short interval and assess its relationship with incorrect heart rate limit alarm rates.

Basic procedures—Six ECG beat detectors were assessed using performance metrics over long interval (sensitivity and positive predictive value over 30 min) and short interval (Area Under empirical cumulative distribution function (AUecdf) for short interval (i.e., 10 s) sensitivity and positive predictive value) on two ECG databases. False heart rate limit and asystole alarm rates calculated using a third ECG database were then correlated (Spearman's rank correlation) with each calculated performance metric.

Main findings—False alarm rates correlated with sensitivity calculated on long interval (i.e., 30 min) ($\rho = -0.8$ and $p < 0.05$) and *AUecdf for sensitivity* ($\rho = 0.9$ and $p < 0.05$) in all assessed ECG databases. Sensitivity over 30 min grouped the two detectors with lowest false alarm rates while *AUecdf for sensitivity* provided further information to identify the two beat detectors with highest false alarm rates as well, which was inseparable with sensitivity over 30 min.

Principal conclusions—Short interval performance metrics can provide insights on the potential of a beat detector to generate incorrect heart rate limit alarms.

Keywords

ECG beat detectors; Heart rate limit alarms; False alarms; Performance metrics

*Corresponding author at: Office of Science and Engineering Laboratories, CDRH, US FDA, Silver Spring, MD, USA. cldaluwatte@mail.missouri.edu, chathuri.daluwatte@fda.hhs.gov.

Disclaimer

This article reflects the views of the authors and should not be construed to represent FDA's views or policies. The mention of commercial products, their sources, or their use in connection with material reported herein is not to be construed as either an actual or implied endorsement of such products by the Department of Health and Human Services.

Introduction

Patient monitors that include arrhythmia detection use information provided by electrocardiogram (ECG) beat detectors to trigger an alarm when an arrhythmic event is detected. Heart rate limit alarms have been identified as one of the major contributors to alarm fatigue, which refers to healthcare providers becoming desensitized to alarms due to an excessive number of false and nonactionable alarms [1,2].

Performance of ECG beat detectors are traditionally assessed using sensitivity and positive predictive value to detect beats by beat type (e.g. normal beat, atrial premature beat, premature ventricular contraction, left bundle branch block beat) [3] within an ~30 min interval on human-annotated databases. For example, in the Massachusetts Institute of Technology-Beth Israel Hospital (MIT-BIH) database signals have ~2000 beats per record within a 30 min test interval [3,4]. When evaluated using sensitivity and positive predictive value many beat detectors have shown high performance (>99%) on these databases [5–9]. While traditional sensitivity and positive predictive value scores can predict an ECG beat detector algorithm's longtime performance [10,11], how these metrics correlate with incorrect (missed + false) heart rate limit alarm rates has not been systematically examined.

High sensitivity and positive predictive value (>99%) over a long interval (~30 min) can be achieved by having the same number of incorrect beat detections grouped together within a short time interval or dispersed across the recording. However, missing a true arrhythmic event (missed alarm) or detecting an arrhythmic event when one is not actually present (false alarm) can be caused by a limited number of incorrect beat detections occurring within a short period. Thus these traditional metrics do not capture information about the sparsity of incorrect beat detections that may be more predictive of incorrect bradycardia or tachycardia event detections. Therefore a beat detection performance metric which also contains information about the sparsity of incorrect beat detections may provide further information useful to understand the false heart rate limit alarm rates.

In this paper we propose a metric which estimates ECG beat detector performance for long ECG recordings (e.g. 30 min) based on the distribution of incorrect beat detection over short intervals (~ 10 s) by calculating sensitivity and positive predictive value over the short intervals and assessing the distribution of sensitivity and positive predictive value over all short intervals in the test database. This metric can help developers in evaluating the performance of their algorithms. Using standard ECG test databases [3] we calculated both the proposed short interval performance metric and traditional performance metrics for 6 beat detector algorithms. We then studied the association between incorrect asystole and heart rate limit alarm rates on an ECG database with waveforms from patients in critical care units, operating rooms, and cardiac catheterization laboratories [12].

Methods

Beat detection performance metrics

Beat detection performance was assessed using sensitivity and positive predictive value with respect to human-annotated beats. Sensitivity was defined as [number of annotated beats

detected]/[number of total annotated beats]. Positive predictive value was defined as [number of annotated beats detected]/[number of total detected beats]. Using sensitivity and positive predictive value, long interval and short interval beat detection performance metrics were calculated as follows.

1. Long interval metric—As long interval performance metrics traditional gross scores for sensitivity and positive predictive value were calculated using beats within 30 min test intervals for all subjects in the database.

2. Short interval metric—The proposed short interval metric is obtained by calculating beat detection sensitivity and positive predictive value over short intervals (e.g.: 10 s) with a time overlap (e.g.: 5 s) over the 30 min test interval on all records in the database.

Histograms for short interval sensitivity and positive predictive value are then constructed (Fig. 1a and b). The histogram estimates the *probability density function (pdf)* for “beat detection performance over short interval”. Using the histogram, *empirical cumulative distribution function (ecdf)* for “performance over short interval” was constructed (Fig. 1c and d), by calculating the area under the histogram from 0% to x, for x ranging from 0% through 100%. An ideal detector which has 100% sensitivity and positive predictive value over all short intervals will have a uniform distribution at 100% on both *pdf* (Fig. 1a) and *ecdf* (Fig. 1c). When a beat detector has a high probability of having high number of incorrect beat detections over a short interval, both the *pdf* (Fig. 1b) and *ecdf* (Fig. 1d) will start to have a thicker left tail. The thickness of the left tail can be quantitatively estimated by calculating the *area under empirical cumulative distribution function (AUecdf)*. We propose to use *AUecdf* as a metric which estimates ECG beat detector performance over short intervals. A smaller *AUecdf* (closer to 1.0) is favorable, indicating a lower probability of having high number of incorrect beat detections over a short interval.

Beat detectors

We used a set of beat detector algorithms ($n = 6$) with open source implementation in Matlab which have reported ~99% gross sensitivity and positive predictive value on Massachusetts Institute of Technology-Beth Israel Hospital (MIT-BIH) database. This set consisted of three beat detectors from Physionet WFDB toolset [13–15] and beat detectors reported in Afonso et al. [9], Pan et al. [7], and Johannesen et al. [8]. The six detectors are referred using a letter from A–F. Signal analysis and beat detection for the current study was performed in Matlab R2016a (The Mathworks, Natick, MA).

Databases

Massachusetts Institute of Technology-Beth Israel Hospital (MIT-BIH) Arrhythmia Database and American Heart Association (AHA) ECG Database are recommended in the American National Standards Institute/Association for the Advancement of Medical Instrumentation/electrocardiography 57: 2012 (ANSI/AAMI/EC57: 2012) standard [3] to test an ECG beat detection algorithm. MIT-BIH consists of ECG segments from de-identified ambulatory ECG recordings (48 records from 47 patients of 30 min duration) at Boston’s Beth Israel Hospital combining uncommon but clinically important arrhythmias that would not be well represented in small samples [4]. AHA consists of de-identified ECG

segments (64 records of 3 h duration with beat annotations in the last 30 min) selected from 17 institutes worldwide focused exclusively on ventricular arrhythmia [16]. We calculated the proposed short interval performance metric as well as traditional scores using these two databases following criteria in ANSI/AAMI/EC57: 2012 to select signals and signal segments. We computed AUecdf for sensitivity and positive predictive value as described above with short interval = 10 s, overlap = 5 s, and histogram bin size = 1%.

Incorrect heart rate limit alarm rates for each beat detector were assessed using ECG signals from the publicly available Massachusetts General Hospital/Marquette Foundation (MGH/MF) Waveform Database [12]. MGH/MF consists of waveforms of varying length (12–86 min) from 250 patients in critical care units, operating rooms, and cardiac catheterization laboratories collected at Massachusetts General Hospital along with beat and event annotations. We used 234 ECG signals (total monitor time = 279.04 h) out of 250 patients, excluding 16 signals which we were unable to download (record numbers: 14, 26, 81, 85, 123, 146, 147, 148, 149, 163, 175, 192, 194, 196, 210, and 224). Using beat annotations available within the database, 10 s epochs with 5 s overlap were annotated as an arrhythmic epoch or not for asystole, bradycardia, and tachycardia. For the purpose of this analysis alarm epochs for each arrhythmia were defined as asystole: no heartbeats for 4 s, bradycardia: heart rate <50 beats per minute, and tachycardia: heart rate >150 beats per minute [2]. A missed alarm was declared if detected beats did not meet alarm condition during an arrhythmic epoch annotated by reference annotations. A false alarm was declared if the alarm condition was met on an epoch annotated as non-arrhythmic by reference annotations. If there was an asystole and a bradycardia event in the epoch, the epoch was marked as an asystole epoch. For each record, incorrect alarm epoch rates were then calculated by dividing the number of missed, false and incorrect (incorrect = missed + false) alarm epochs respectively by monitor time. As heart rate limit alarms and asystole have different criticality levels in clinical settings [2,17], mean alarm epoch rates for each beat detector were calculated in two categories: heart rate limit alarms (bradycardia and tachycardia) and asystole.

Statistical analysis

Using Spearman's rank correlation coefficient (ρ) at 5% level of significance, the rank order of detectors by performance metrics was compared to rank order by incorrect heart rate limit alarm epoch rates and incorrect asystole alarm epoch rates. For all performance metrics 95% confidence intervals were calculated using bootstrapping with 5000 iterations and percentile method [18]. All statistical calculations were conducted in R version 3.2.2 (R Foundation for Statistical Computing, Vienna, Austria). No adjustments were made for multiple comparisons.

Results

Table 1 summarizes the performance results for each detector on MIT-BIH and AHA databases using traditional gross scores and the proposed short interval performance metric. For all performance metrics more variability across detectors was observed on AHA database compared to MIT-BIH database for gross positive predictive value but not for gross

sensitivity, except for detector F which had considerably low gross sensitivity on AHA. Gross scores for positive predictive value on MIT-BIH and AHA were strongly correlated ($\rho = 0.9$, $p = 0.0333$) but not for sensitivity ($\rho = 0.7$, $p = 0.1361$). However AUecdf for positive predictive value were correlated on MIT-BIH and AHA ($\rho = 0.9$, $p = 0.0333$) while the correlation was marginally significant for sensitivity ($\rho = 0.8$, $p = 0.0583$).

Table 2 reports the number of incorrect heart rate limit alarm epoch rates and incorrect asystole alarm epoch rates by each beat detector on MGH/MF. All detectors had similar missed alarm epoch rates for asystole (9–11 missed epochs/h) and heart rate limit alarms (22–51 false epochs/h with three out of six detectors having ~50 missed epochs/h). However there was high variability on false alarm epoch rates for both categories: 2–24 false asystole epochs/h and 2–153 false heart rate limit alarm epochs/h. Therefore we reported the rank order correlation between false alarm epoch rates (for asystole and heart rate limit alarms) and performance metrics.

Asystole false alarm epoch rate on MGH/MF was correlated with AUecdf for sensitivity on MIT-BIH ($\rho = 0.9$, $p = 0.0188$, Fig. 2a), gross sensitivity on MIT-BIH ($\rho = -0.8$, $p = 0.0416$, Fig. 2b), AUecdf for sensitivity on AHA ($\rho = 0.9$, $p = 0.0188$, Fig. 2c), and gross sensitivity on AHA ($\rho = -0.9$, $p = 0.0416$, Fig. 2d). The same four performance metrics (AUecdf for sensitivity on MIT-BIH and AHA, gross sensitivity on MIT-BIH and AHA) were correlated with asystole incorrect alarm epoch rate on MGH/MF as well.

False heart rate limit alarm epoch rate on MGH/MF was correlated with gross sensitivity on MIT-BIH ($\rho = -0.9$, $p = 0.0188$, Fig. 3a), AUecdf for sensitivity on MIT-BIH ($\rho = 0.8$, $p = 0.0416$, Fig. 3b) and AUecdf for positive predictive value on MIT-BIH ($\rho = -0.8$, $p = 0.0416$, Fig. 2c). However, only AUecdf for sensitivity on MIT-BIH and gross sensitivity on MIT-BIH showed a significant correlation with incorrect heart rate limit alarm rate. None of the other performance metrics calculated on MIT-BIH or AHA showed significant correlations with false/incorrect alarm epoch rates for either asystole alarms or heart rate limit alarms on MGH/MF.

Discussion

The need for standard test databases to assess automated arrhythmia detector performance has been recognized and addressed for the past three decades [4,16]. While development of appropriate metrics to be used to assess performance was recognized as an important part of standard test database development [3,16], efforts have been minimal and mainly focused on ventricular arrhythmias [11,19,20]. However, heart rate limit alarms related to ECG beat detection have been shown to account for a notable number of false patient monitoring alarms [1,2]. Currently reported performance metrics, i.e. gross and/or average sensitivity and positive predictive value, provide a convenient summary of beat detector performance but may not easily extrapolate to predict performance of an ECG beat detector to produce incorrect alarm conditions [3,16]. In this study we proposed a metric which estimates ECG beat detector performance over short intervals and studied its relationship with the number of incorrect alarm rates for heart rate limit alarms and asystole alarms on a critical care unit database using 6 beat detector algorithms. This new metric can help algorithm developers

better evaluate promising improvements in their designs. We studied the same relationship using traditional performance metrics in place of short interval metric as well.

Our results showed beat detectors with similar performance assessed by traditional performance metrics on standard test databases (~99% gross sensitivity and positive predictive value on MIT-BIH) had similar missed alarm rates for both asystole and heart rate limit alarms but had high variability of false alarm rates. This might be because the standard test databases were not designed to match the prevalence of arrhythmic events, patient population, clinical settings, and artifact prevalence but were rather developed to cover a range of arrhythmias that might be encountered in a clinical environment [4,16,12]. If databases are developed or disaggregated to match arrhythmic event, patient population, clinical settings, and artifact prevalence comparable to real-world clinical prevalence, the gross scores may be able to predict the likelihood of false alarm rates of a detector as well.

We hypothesized that a performance metric that captures information of correct ECG beat detections over short intervals may provide more information on incorrect heart rate limit alarm rates. Spearman's rank correlation coefficient demonstrated a consistent relationship between detector performances reported as the short interval performance metric for sensitivity with false alarms rates for both asystole and heart rate limit alarms. A similar relationship was observed for gross sensitivity as well. The tested short interval performance metric on MIT-BIH formed three detector clusters (non-overlapping 95% confidence intervals) with two detectors in each, which corresponded to two detectors with highest, moderate and lowest false alarm epoch rates considering both asystole and heart rate limit alarms. Sensitivity gross score was able to cluster the two detectors with the lowest false alarm rates. This suggests that considering performance over short intervals may provide insights on real world performance of false heart rate limit alarm performance even while using a database that is not matched to real-world clinical prevalence in terms of arrhythmic events, patient population, clinical settings, and artifacts. It is also interesting to note that, in our study, sensitivity had superior ability to predict false alarm rates for asystole and heart rate limit alarms compared to positive predictive value. This might be because both heart rate limit alarms and sensitivity are defined with respect to reference beat annotations, as opposed to positive predictive value, which is defined with respect to detected beats (positive predictive value - possibility a detected beat is a true beat).

While all six detectors had gross/average sensitivity >98.3% and positive predictive value >98.8% on MIT-BIH, on AHA three had positive predictive value <97% (97.45%, 94.31% and 87.34%) and one had sensitivity of 93%. This might be due to the high prevalence of ventricular beats on AHA database compared to MIT-BIH database. Performance on both of these databases provides complementary evidence of a detector's performance.

Limitations

One limitation of this study is that we only looked at one short interval performance metric with short interval set to 10 s. We did not study how the relationship changes with the duration of the short interval nor with other performance metrics which can capture sparsity of incorrect beat detections over time. It has been reported that widening alarm parameters

was one of the promising interventions to reduce false heart rate limit alarms but this came with mixed safety outcomes [21]. In our study we limited the analysis to one set of parameters to define heart rate limit alarm criteria (<50 bpm for bradycardia and >150 bpm for tachycardia) but did not study how changing these parameters affects the association.

Conclusions

Beat detection performance calculated as sensitivity on standard human annotated databases was related with the false alarm rates for heart rate limit alarms and asystole alarms. Considering performance over short intervals can provide further insights on real world performance of ECG beat detector algorithms to produce false heart rate alarms.

Acknowledgments

This project was supported in part by U.S. Food and Drug Administration's Medical Countermeasures Initiative, Critical Path Initiative, Office of Women's Health and appointments to the Research Participation Programs at the Oak Ridge Institute for Science and Education through an interagency agreement between the Department of Energy and FDA.

References

1. Drew BJ, Harris P, Zègre-Hemsey JK, Mammone T, Schindler D, Salas-Boni R, et al. Insights into the problem of alarm fatigue with physiologic monitor devices: a comprehensive observational study of consecutive intensive care unit patients. *PLoS One*. 2014; 9(10):e110274. [PubMed: 25338067]
2. Graham KC, Cvach M. Monitor alarm fatigue: standardizing use of physiological monitors and decreasing nuisance alarms. *Am J Crit Care*. 2010; 19(1):28–34. [PubMed: 20045845]
3. ANSI/AAMI EC57. Testing and reporting performance results of cardiac rhythm and ST segment measurement algorithms. 2012
4. Moody GB, Mark RG. The impact of the MIT-BIH arrhythmia database. *IEEE Eng Med Biol Mag*. 2001; 20(3):45–50. [PubMed: 11446209]
5. Hamilton PS, Tompkins WJ. Quantitative investigation of QRS detection rules using the MIT/BIH arrhythmia database. *IEEE Trans Biomed Eng*. 1986; BME-33(12):1157–65.
6. Zong W, Moody GB, Jiang D. A robust open-source algorithm to detect onset and duration of QRS complexes. *Comput Cardiol*. 2003
7. Pan J, Tompkins WJ. A real-time QRS detection algorithm. *IEEE Trans Biomed Eng*. 1985; BME-32(3):230–6.
8. Johannesen L, Vicente J, Galeotti L, Strauss DG. Ecglib: library for processing electrocardiograms. *Comput Cardiol*. 2013
9. Afonso VX, Tompkins WJ, Nguyen TQ, Shen L. ECG beat detection using filter banks. *IEEE Trans Biomed Eng*. 1999; 46(2):192–202. [PubMed: 9932341]
10. Moody GB, Mark RG. How can we predict real-world performance of an arrhythmia detector. *Comput Cardiol*. 1983
11. Wang JYJ, Helfenbein ED. Long-term performance evaluation of arrhythmia algorithms using unannotated databases. *Comput Cardiol*. 1988
12. Welch J, Ford P, Teplick R, Rubsamen R. The Massachusetts General Hospital-Marquette Foundation hemodynamic and electrocardiographic database—comprehensive collection of critical care waveforms. *J Clin Monit*. 1991; 7(1):96–7.
13. Silva I, Moody GB. An open-source toolbox for analysing and processing physionet databases in matlab and octave. *J Open Res Softw*. 2014; 2(1)
14. Moody GB, Mark RG, Goldberger AL. PhysioNet: a web-based resource for the study of physiologic signals. *IEEE Eng Med Biol Mag*. 2001; 20(3):70–5. [PubMed: 11446213]

15. Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, et al. Physiobank, physiotoolkit, and physionet. *Circulation*. 2000; 101(23):e215–20. [PubMed: 10851218]
16. Hermes, R., Oliver, G. Ambulatory electrocardiographic recording. Chicago: Yearbook Med Pub; 1981. Use of the American Heart Association database.
17. Whalen DA, Covelle PM, Piepenbrink JC, Villanova KL, Cuneo CL, Awtry EH. Novel approach to cardiac alarm management on telemetry units. *J Cardiovasc Nurs*. 2014; 29(5):E13–22. [PubMed: 24365870]
18. Chernick, MR. Bootstrap methods: a guide for practitioners and researchers. 2. New York: Wiley; 2007.
19. Wang J. Proposed new requirements for testing and reporting performance results of arrhythmia detection algorithms. *Comput Cardiol*. 2013
20. Greenwald SD, Albrecht P, Moody GB, Mark RG. Estimating confidence limits for arrhythmia detector performance. *Comput Cardiol*. 1985
21. Paine CW, Goel VV, Ely E, Stave CD, Stemler S, Zander M, et al. Systematic review of physiologic monitor alarm characteristics and pragmatic interventions to reduce alarm frequency. *J Hosp Med*. 2016; 11(2):136–44. [PubMed: 26663904]

Fig. 1.

a) Histogram with bin size 1%, which estimates the probability density function (*pdf*) for “Sensitivity (Se) over short interval” for an ideal detector. b) Histogram with bin size 1%, which estimates the probability density function (*pdf*) for “Sensitivity (Se) over short interval” for a non-ideal detector. The non-ideal detector is an ECG beat detector algorithm with gross Sensitivity of 82.8% on Massachusetts Institute of Technology-Beth Israel Hospital (MIT-BIH) database. c) Empirical cumulative distribution function (*ecdf*) for “Sensitivity (Se) over short interval” derived using histogram 1a. d) Empirical cumulative distribution function (*ecdf*) for “Sensitivity (Se) over short interval” using histogram 1b.

Fig. 2.

Scatter plots for significant correlations asystole false alarm epoch rate on MGH/MF database showed with beat detection performance metrics: a) AUecdf: *Area Under empirical cumulative distribution function for Sensitivity (Se) over short interval* on MIT-BIH, b) Gross Sensitivity (Se) on MIT-BIH, c) AUecdf: *Area Under empirical cumulative distribution function for Sensitivity (Se) over short interval* on AHA and d) Gross Sensitivity (Se) on AHA. Error bars represent 95% confidence interval calculated using bootstrapping.

Fig. 3.

Scatter plots for significant correlations false heart rate (HR) limit alarm epoch rates on MGH/MF database showed with beat detection performance metrics: a) Gross Sensitivity (Se) on MIT-BIH, b) AUecdf: *Area Under empirical cumulative distribution function for Sensitivity (Se) over short interval* on MIT-BIH, c) AUecdf: *Area Under empirical cumulative distribution function for positive predictive value (PPV) over short interval* on MIT-BIH. Error bars represent 95% confidence interval calculated using bootstrapping.

Table 1

Beat detection performance metrics (Sensitivity: Se, Positive Predictive Value: PPV, AUecdf: Area Under empirical cumulative distribution function for “performance over short interval”) on MIT-BIH arrhythmia database and AHA ECG database.

Detector	MIT-BIH arrhythmia database				AHA ECG database			
	Gross score		AUecdf		Gross score		AUecdf	
	Se	PPV	Se	PPV	Se	PPV	Se	PPV
A	98.36	99.91	2.43	1.14	99.34	99.81	1.90	1.42
B	99.15	99.55	1.98	1.46	99.44	97.45	1.79	2.80
C	98.59	99.67	2.13	1.28	99.28	99.60	1.94	1.62
D	99.78	98.82	1.64	2.26	99.68	94.31	1.56	4.67
E	99.72	99.71	1.57	1.54	99.88	98.22	1.25	2.58
F	98.86	99.31	2.38	1.85	93.00	87.34	8.25	12.34

Table 2
Incorrect heart rate limit and asystole alarm epoch rates on MGH/MF Waveform Database for each beat detector.

Detector	Incorrect alarm epoch rate (epochs per hour)					
	Asystole			Heart rate limit (bradycardia, tachycardia)		
	Missed	False	Total	Missed	False	Total
A	11	18	28	41	153	194
B	10	6	16	36	28	63
C	9	8	17	22	65	86
D	11	2	13	50	6	56
E	11	3	14	51	2	53
F	11	24	34	51	16	66