Published in final edited form as:

J Clin Monit Comput. 2016 December ; 30(6): 875–888. doi:10.1007/s10877-015-9788-2.

Real Alerts and Artifact Classification in Archived Multi-signal Vital Sign Monitoring Data—Implications for Mining Big Data — Implications for Mining Big Data

Marilyn Hravnak, RN, PhD, ACNP-BC, FAAN, FCCM¹, Lujie Chen, MSc³, Artur Dubrawski, PhD³, Eliezer Bose, BSN¹, Gilles Clermont, MD², and Michael R. Pinsky, MD, Dr hc, FCCP, MCCM²

¹University of Pittsburgh Schools of Nursing (Department of Acute and Tertiary Care)

²Medicine (Department of Critical Care Medicine)

³Carnegie Mellon University Robotics Institute (Auton Lab)

Abstract

PURPOSE—Huge hospital information system databases can be mined for knowledge discovery and decision support, but artifact in stored non-invasive vital sign (VS) high-frequency data streams limits its use. We used machine-learning (ML) algorithms trained on expert-labeled VS data streams to automatically classify VS alerts as real or artifact, thereby “cleaning” such data for future modeling.

METHODS—634 admissions to a step-down unit had recorded continuous noninvasive VS monitoring data (heart rate [HR], respiratory rate [RR], peripheral arterial oxygen saturation [SpO₂] at 1/20Hz., and noninvasive oscillometric blood pressure [BP]) Time data were across stability thresholds defined VS event epochs. Data were divided Block 1 as the ML training/cross-validation set and Block 2 the test set. Expert clinicians annotated Block 1 events as perceived real or artifact. After feature extraction, ML algorithms were trained to create and validate models automatically classifying events as real or artifact. The models were then tested on Block 2.

RESULTS—Block 1 yielded 812 VS events, with 214 (26%) judged by experts as artifact (RR 43%, SpO₂ 40%, BP 15%, HR 2%). ML algorithms applied to the Block 1 training/cross-validation set (10-fold cross-validation) gave area under the curve (AUC) scores of 0.97 RR, 0.91 BP and 0.76 SpO₂. Performance when applied to Block 2 test data was AUC 0.94 RR, 0.84 BP and 0.72 SpO₂).

CORRESPONDING AUTHOR: Marilyn Hravnak, University of Pittsburgh, 336 Victoria Hall; 3500 Victoria St., Pittsburgh PA 15261, mhra@pitt.edu, Office: 412-624-6866, Cell: 724-288-4134.

COMPLIANCE WITH ETHICAL STANDARDS

The study was approved and has current active approval from the University of Pittsburgh Institutional Review Board (PRO12070002). All procedures performed in this study involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. For this type of study formal consent is not required, and was approved with consent waiver.

Conflicts of Interest: The authors have no commercial conflicts of interest to report.

CONCLUSIONS—ML-defined algorithms applied to archived multi-signal continuous VS monitoring data allowed accurate automated classification of VS alerts as real or artifact, and could support data mining for future model building.

Keywords

physiologic monitoring; machine learning; archived data; big-data; vital signs; artifact; cardiorespiratory instability

INTRODUCTION

Hospitalized patients at risk for developing cardiorespiratory instability (CRI) undergo continuous non-invasive monitoring of vital sign (VS) parameters consisting of electrocardiography, automated sphygmomanometry and pulse oximetry to estimate heart rate (HR), respiratory rate (RR), blood pressure (BP) and peripheral arterial O₂ saturation (SpO₂). Patients are at CRI risk when their VS parameters vary from pre-determined normality thresholds [1], but non-normal values can be due to monitoring artifact [2]. Artifact can be due to patient movement, while mechanical artifact or “noise” can be due to mains interference (60Hz), poor impedance, or loose sensors [3].

Data from physiologic monitors can be archived in hospital information systems, either as averaged data or in a high-frequency waveform state. There is an emerging trend in biomedical research to apply machine learning to mine these enormous “big” data repositories, either alone or coupled to other clinical and administrative data, to generate new knowledge or develop clinical decision support [4]. Big data utilization requires tools and techniques for the proper organization, efficient access, and smart analysis of huge volumes of digital data [5, 6]. Artifact in the data presents both technical and methodological challenges to meaningful data use [7]. Human experts can curate datasets by annotating events as real or artifact, but such expert annotation can be costly in terms of time, effort and economic constraints [8, 9].

We sought to determine if using a machine learning method to automate differential recognition of artifact from true VS abnormalities in stored offline monitoring data could be accomplished. Based on clinical observation and practice, we hypothesized that a large proportion of artifact follows patterns based on each VS and its relationship to other VS, which can be described. Machine learning (ML) algorithms can learn to recognize such patterns through training on an initial primary dataset, and next use that knowledge to perform accurate classification of new, unseen examples as real or artifact [10]. ML-based automated pattern recognition has been used to successfully classify patterns in ultrasound, echocardiographic and computerized tomography images [11-13], electroencephalogram signals [14], intracranial pressure waveforms [15], and word patterns in electronic health record text [16]. We hypothesized that ML could learn and automatically classify VS artifact patterns in a manner which minimizes false positives (artifacts counted as true instability) and false negatives (true instability not captured). This approach, if applied to huge repositories of stored monitoring data, could assist in cleaning artifact from such data. To test our hypothesis that ML could successfully discriminate between real alerts and artifact

in collected physiologic data, we split 16 weeks of continuous VS monitoring data of continuous noninvasively monitored patients on a stepdown-unit (SDU) into two 8-week blocks, with Block 1 serving as a ML training/cross-validation set and Block 2 as the external test set.

MATERIALS AND METHODS

Patients and Setting

Following local Institutional Review Board approval, we collected continuous VS data streams including HR (3-lead ECG), RR (bioimpedance signaling), SpO₂ (pulse oximeter Model M1191B, Phillips, Boeblingen, Germany; clip-on reusable finger sensor), and intermittent noninvasive BP (minimum frequency two hours), from all patients over two sequential but separate 8-week time blocks in a 24-bed adult surgical-trauma SDU (Level-1 Trauma Center). Patients were temporally constrained to a single block. Demographic and clinical data were obtained from clinical and administrative databases. Block 1 yielded 308 admissions (22,588 total patient-hours of monitoring data; 81.0 mean and 60.9 median hours/patient), and Block 2 yielded 326 admissions (19,046 total patient-hours data; 78.7 mean and 51.6 median monitoring hours/patient). Total monitoring time for both blocks combined was 41,634 hours, or 4.72 years of patient monitoring hours (Table 1).

Event identification

Noninvasive VS monitoring data were recorded at 1/20Hz for HR, RR and SpO₂, and systolic (SysBP) and diastolic (DiaBP) blood pressure less frequently. VS excursion beyond local CRI thresholds (HR<40 or >140, RR< 8 or >36, SBP< 80 or >200, DBP>110, SpO₂< 85%) defined CRI events, occurring 634,137 times across both blocks. We additionally required that events persist initially for a tolerance of 40s, and a minimal duration of 4 minutes continuously or a cumulative duration of 4 out of 5 minutes (80% duty cycle) if intermittent. Tolerance of 40s suggests that if the next exceedance occurred within 40s of the current exceedance, they are the same event. Since sampling frequency is 1/20 Hz, this also means we only allow one normal reading between two abnormal readings. This, combined with the 4-minutes duration, is a prudent constraint to ensure the events period we seek to analyze contains continuously spaced abnormality with minimum interruption, and lasts sufficiently long to be reasonably considered as clinically relevant. These methods yielded 2,333 qualifying events (812 Block 1; 1,521 Block 2). For events where multiple VS crossed threshold simultaneously, the event was labeled based on the first VS crossing. The event period analyzed was from the time the first VS crossed threshold until return to the stability range, defining an "event epoch" (Figure 1).

Annotation, Rules Extraction and Featurization, and Machine Learning

We next used Block 1 data to ascertain rules defining events as artifact or real using a two-part process. Part 1 comprised artifact annotation by human experts and iterative rule derivation and numeric feature extraction (Fig. 2, Part 1), and Part 2 machine-learning model development (Fig. 2 Part 2) as described below.

Part 1: Annotation, Rules Extraction and Featurization (Fig.2, Part 1)—First, clinical investigators (MP, GC, MH) reviewed the 1/20Hz time plots of events to develop a set of business rules for differentiating real from artifact based on visual distinction and clinical experience at the bedside and telemetry. These were available to the clinical reviewers (Fig.3) along with visual examples. All Block 1 VS events were provided as graphical time plots to two expert clinician reviewers separately, who after inspection annotated each as a real or artifact, with disagreements then annotated in concert to establish “ground truth.” Reviewers could use the rules and examples, but were not constrained by them. The reviewers annotated 493 (60%) of the 812 events as real, and 319 (40%) as artifact. Of artifact, 43% were RR, 40% were SpO₂, 15% were BP, and only 2% were HR (Table 1). The rules demonstrated several visually distinct artifact patterns which captured the majority events for each VS-- both variable patterns in the artifact signal values themselves and their relationship to the other VS values—which evolved into a limited set of English language (text) rules describing the most commonly encountered multi-signal patterns for artifacts. For numeric feature extraction, these text rules were translated into sets of numeric features derived from the multi-signal time series data. See Figure 4.A. for an example of a rule applied to a common SpO₂ artifact pattern, and the numeric features derived from the rule. All rules and features were then iteratively refined through a process where the events were presented in user-friendly 1- or 2-dimensional artifact plots for each VS (Figure 4.B.SpO₂ example) with respect to selected features. Events whose label seemed inconsistent with the featurization rules were re-reviewed by the experts after which they were either relabeled, new features were derived, or the text rules and in turn their features were redefined. At the end of these feature-learning steps, we identified 44 numeric features used to characterize various aspects of artifacts for each of the VS in Block 1 (Table 2), to be applied to each of the 5 VS signals (HR, RR, SpO₂, SysBP, DiaBP) and two derived signals: mean BP and BP pulse pressure (PP). The reviewers also annotated all events in Block 2 as real or artifact, to serve as the test set (Table 1)

Part 2: Machine Learning (Fig. 2 Part 2A)—For each VS, epoch features were ranked by their Area Under the Receiver Operating Characteristic Curve (AUC) score from a simple univariate logistic regression model that regressed each individual feature against the annotation labels. Figure 5 illustrates ranking of features for SpO₂ events, grouped by the associated VS signals from which the features were derived. To reduce both ML model complexity and risk of over-fitting, we filtered less informative features using an AUC score of 0.7 as an arbitrary cut-point for BP and SpO₂, and 0.85 for RR, leaving 10 features for SpO₂, 5 for RR, and 5 for BP (Table 3). Next we applied various ML algorithms to learn and validate classification models to predict the event epoch binary labels (artifact vs. real) given the features selected. ML algorithms used were: K nearest neighbors (KNN, at different K), Linear Discriminant Analysis (LDA), Naïve Bayesian classifier (NB), Logistic Regression (LR), Support Vector Machine (SVM), and Random Forest (RF) [10, 17, 18]. For computing power spectral density (PSD), we used the Lomb–Scargle periodogram method, which works for unevenly sampled data (i.e. where missing values are possible) using function available from the R package for non-linear time series analyses (nls).

For each VS type, we identified the best models based on their 10 fold cross-validated (CV) AUC scores from Block 1 data. Finally, we applied the selected models to the Block 2 event data (unseen during feature derivation and screening, model-building and model selection) to test their classification performance. The available data contained too few HR artifacts to reliably learn a model for automated adjudication of HR alerts.

RESULTS

The baseline data for Blocks 1 and 2 (demographics, monitor hours, event prevalence), and the events and their expert annotations are summarized in Table 1. Patients in Blocks 1 and 2 had similar demographics, mean Charlson Co-morbidity Index and hospital length of stay (LOS), but Block 2 patients had shorter SDU LOS. Monitoring hours per Block and per patient were similar, but events were greater in Block 2, as were distribution of events per VS. Figure 6 summarizes Block 1 AUC scores from the various ML algorithms following model training and 10-fold cross-validation (CV) AUC plots from 10-fold CV experiments on the array of algorithms for three categories of VS events. The mean AUC scores when testing the algorithms on the Block 2 test set data are also displayed. The actual data is listed in Table 4 (electronic supplement), along with standard deviations (SD) of the AUC scores (SDs for Block 1 results obtained from 100 iterations of tenfold-CV on Block 1 data; SDs for Block 2 results obtained by testing Block 2 data on models learned from 100 copies of bootstrapped Block 1 data). In Block 1, RR provided consistently best AUC scores for discriminating real and artifact alerts, ranging from an AUC of 0.970 for RF to 0.951 for K(5)NN. BP ranged from an AUC of 0.907 for LR to 0.860 for NB. SpO₂ ranged from an AUC of 0.764 for NB to 0.702 for both K(5)NN and SVM. An algorithm which performed consistently well across RR, BP and SpO₂ was LR, with AUC scores of 0.956, 0.907, and 0.750 respectively. The model degrading the least between Blocks 1 and 2 was SVM.

Figure 7 presents ROC diagrams illustrating application of the best performing Block 1 model on the Block 2 test set. In both diagrams, the positives refer to real alerts while negatives refer to artifact. Figure 7A illustrates the tradeoff between True Positive Rate (Sensitivity) versus False Positive Rate (1.0–Specificity), and is relevant to ascertain the model's ability to identify real instability. The second, Figure 7B illustrates the tradeoff between True Negative Rate (Specificity) and False Negative Rate (1.0–Sensitivity) to ascertain the model's ability to identify artifact. Clearly, defining the appropriate sensitivity and specificity is not only a function of algorithm performance but also the cost of misclassifying a real event (false negative) against accepting too much as artifact. Even for SpO₂ – the most challenging VS to classify--our artifact identification algorithms remained robust at a false negative rate of 30%. Thus of our total of 354 expert-annotated SpO₂ artifact in the Block 2 test set, our algorithm would have correctly identified 70% or 248 of them as artifact events.

DISCUSSION

Our key findings are that 1) many common VS monitoring artifacts in stored off-line data obtained in SDU care can be captured using a modestly sized set of patterns unique to each VS parameter and its relationship to other VS, 2) VS artifact patterns can be captured by

numeric features, which can in turn be used to discriminate VS events as real or artifact, and 3) ML algorithms can perform well in classifying events as real or artifact in offline unseen data. We envision these findings to be of importance in enabling mining of large monitoring data repositories, since artifact recognition and extraction can occur without the use of time-intensive and expensive human annotation. Automating artifact rejection from data repositories holds promise to enable building prediction models based upon an important degree of reliability for true instability.

We believe ours is the first study examining artifact in offline continuous noninvasive monitoring of multiple VS signals in patients outside the ICU. There is precedent for differentiating between artifact and real events in the large publically-available Physionet Multitparameter Intelligent Monitoring in Intensive Care (MIMIC) I and II databases of surface ECG and invasive monitor parameters [19, 20], which was developed to provide large volumes of offline ICU data to serve as a data mining and clinical decision support platform [21]. Waveform information is stored at 125 Hz, and trend parameters at a resolution of once per second, as well as a record of all bedside monitor-generated alarms (monitor sense of arrhythmia or inoperable condition like sensor disconnect) from a single monitor manufacturer (Philips Healthcare). Alarms and inoperable alerts are not verified as correct by humans, except for monitor-generated alarms for extreme bradycardia, tachycardia, ventricular tachycardia and fibrillation, and asystole, which were retrospectively reviewed offline by two independent experts, and discrepancies adjudicated by a third expert. Although MIMIC search engines permit users to determine which signals exist for which patients, user guides caution that there is no guarantee of quality. To avoid requesting noisy data, users can apply a set of signal quality indices for ECG and BP data, based on a combination of statistical measures in time and frequency domains, as well as a multiple-channel ECG performance and correlation to past data. Nevertheless, these are for screening and not annotation, and some artifacts are not well represented [21]. Using MIMIC II, Aboukhalil et al. [22] developed an algorithm to identify ECG alarms for critical arrhythmias when the annotated ECG signal (real or artifact) was featurized and considered in relationship to the invasive arterial BP signal. For example, extreme bradycardia evolved as a function of the maximum negative error allowed between the ECG calculated HR and the intra-arterial BP-derived HR, and the number of beats used to calculate the HR. Other investigators also describe handling noisy signals in the MIMIC databases [23-25].

There are a variety of ways that online artifact alerts at the bedside have been addressed by the monitoring industry, including analog and digital filters using temporal or spatial averaging techniques to address mains interference, electromagnetic emissions and muscle signals on the ECG [26]. Adaptive filtering algorithms have been proposed [27-30], along with alert-related filters that recover signals at defined intervals in cases of periodic noise patterns [31, 32], and signal reconstruction [33]. However the problem we seek to address is different as it applies to already collected offline data to be cleaned for mining purposes. Such a system should be agnostic to propriety artifact labeling approaches across monitor types within or across hospitals, lending to greater generalizability.

There is also precedent for ML artifact detection applied in smaller local datasets of high frequency physiologic ICU monitoring data. Tsien et al. examined neonatal ICU data and

used decision trees to first detect artifact in multiple signals, and then LR to differentiate between real alerts and artifact with mixed results, although better results were obtained when neural networks were applied [34]. There are also studies using ML of archived offline ICU signal data in developing outcome risk models. Several used filtering to screen artifact from data streams of ECG and/or arterial BP signals for modeling for cardiac arrest or mortality [35, 36]. Others used data streams from ICU and OR databases to predict impending events (35) or provide decision support [37, 38], but did not explicate how the problem of artifact was addressed. Finally, some addressed the problem by developing data repositories in real time wherein they videotaped the entire LOS, and then had experts view the videotapes and annotate bedside alarm conditions as artifact when based upon observed patient movement or clinician activity (turning the patient, changing sensors, drawing blood, etc.) [39, 40]. Although this direct observation method for artifact versus real event discrimination can increase accuracy, the tradeoff is in limited monitoring hours, as well as significant human annotation time.

In contrast, our approach seeks to develop algorithms to automate accurate annotation of VS events directly within the archived local parent database as real or artifact to clean the repository, using pattern recognition and machine learning of VS data that is independent of proprietary monitor manufacturer-assigned alarm states. Our data demonstrate that artifact in offline VS data were, for the most part, not difficult to classify. The majority of RR artifact was captured by two primary patterns (RR alert in the absence of a HR signal, indicating unreliable bioimpedance capture for RR calculation [91%], and sparse RR signal in concordance with a dense HR signal [9%]). Although most SpO₂ artifact was captured by 3 patterns (Figure 4.B.), they proved the most difficult to classify with ML given the current set of features and models. This may be because of more potential causes for SpO₂ artifact (signal most prone to motion artifact, ambient light, thickness of skin or nails, skin discoloration, changes in perfusion [41]), making patterns less repetitive and therefore more difficult to classify. Nevertheless, our artifact detection algorithms make some headway in differentiating between SpO₂ real alerts vs. artifact, and would have correctly identified 70% or 248 artifacts in Block 2 data as artifact events.

The ultimate goal for classification is 100% accurate identification of artifact with no misclassification of true events but, as seen in Figure 7, the AUCs are not perfect, and cutoff thresholds must be chosen. When applying this method with the goal of eliminating artifact from stored data to develop models predicting true instability, identifying the real instability is more relevant, and threshold choice plays a role in determining the extent of purity of the resultant classified true instability. Setting a higher threshold will yield fewer instability events but of higher confidence—an approach more acceptable when there is an abundance of data. But in a small data set, researchers may be willing set a lower AUC potentially accepting more artifact so that more instability events can be modeled, but confidence in the models would be lower. In practice, the best threshold choice is determined by the robustness of the algorithm and the amount of available data.

Much past work uses curated datasets for prediction modeling, but available datasets are few, do not apply to all populations, have been assembled at expense, and may reflect proprietary monitor characteristics limiting generalizability. The newer challenge is to leverage raw data

accumulating in local repositories using local monitors and clinical record capture for predictive and analytic capabilities [42]. We see the need to provide for both automated annotation of physiologic monitoring data agnostic to proprietary monitor alarms, and integrate additional time-tagged biological, clinical and administrative data at the local level in order to improve data mining of instability causes and treatment [6]. Although using bedside charted and validated VS are possible, they are intermittent and do not capture important information about regulatory mechanisms and interrelationships that are only found within granular data streams and waveforms [43, 44], the tradeoff being that granular continuous data sources are riddled with artifact [36]. It is imperative that valid methodologies be developed to handle the problem of artifact detection, annotation and elimination in large sets of granular data. Only in this manner can data mining for clinical decision support be accurate. The approach we describe holds promise for meeting this challenge.

In future research, better methods to resolve any inconsistency in the human adjudication process such as use of a larger pool of defined experts or even “crowd sourcing” [45, 46], use of an active ML method to select the most difficult cases for annotation [8, 9], and better approaches to extract features that reflect the expert rules with high fidelity and consistency, may prove beneficial in improving the positive predictive value and sensitivity of the discrimination in the unseen data.

Our study has limitations. First, our “ground truth” definition of real versus artifact events was defined by two experts with extensive clinical experience in differentiation of real and artifactual waveforms at the bedside. They may have applied an incorrect label even in agreement of such, or missed noting key features important in their discrimination. Thus, our ground truth may not be as objective as a method using direct patient observation, or a larger expert review panel. However, our annotation method was similar to that used to annotate MIMIC [22], and we were able to annotate data for a number of patients and monitoring hours many magnitudes greater than studies using direct patient observation. In future study, better methods to improve objectivity and inconsistency resolution in the human adjudication process, and better approaches to extract features reflecting the expert rules with high fidelity may help improve robustness of ground truth, and in turn the positive predictive value and sensitivity of discrimination in the unseen data. Thus, our work provides a good start in developing a framework and approach to inform future exploration. Second, for some VS the prevalence of artifacts was rare as compared to real alerts. For example, HR artifacts were too few to build reliable classifiers given the severely unbalanced class distribution. Our requirements for event tolerance and persistence, rather than merely crossing a VS threshold, likely screened out some artifact, thereby losing key discriminatory information and blunting the models. In a future study we plan to run the models with the persistence requirement lifted to answer the question. Some disparity between Blocks 1 and 2 event prevalence and VS type may have contributed to novel features in Block 2, although this would bias our results to be weaker than they could have been otherwise. Finally, our data were obtained from one institution and one monitor manufacturer. Future algorithm refinement may benefit from data obtained from more than one institution and multiple monitor manufactures to improve generalizability.

CONCLUSIONS

VS events are common in continuously monitored clinical data, and the majority of most single VS parameter artifacts are associated with patterns specific for each VS. ML is applied to repositories of high frequency vital sign physiologic monitoring to learn artifact patterns and then recognize them in unseen data. Providing the ability to differentiate VS events as real alerts or artifact in local clinical monitoring data repositories is an important breakthrough. Such ability is necessary to utilize large data repositories to build meaningful prediction models in big data.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funding: This study was funded with grant support from the United States National Institutes of Health, National Institute of Nursing Research RO1 NR13912 and the National Science Foundation NSF 1320347. The funding bodies approved the study design as submitted in the grant application proposal, but had no role in the data collection, analyses, or interpretation, manuscript preparation, or decision to submit the manuscript for publication.

REFERENCES

1. Otero A, Félix P, Barro S, Palacios F. Addressing the flaws of current critical alarms: a fuzzy constraint satisfaction approach. *Artif Intell Med.* 2009; 47:219–238. [PubMed: 19796924]
2. Takla G, Petre JH, Doyle DJ, Horibe M, Gopakumaran B. The problem of artifacts in patient monitor data during surgery: a clinical and methodological review. *Anest Analg.* 2006; 103:1196–1204.
3. Smith M. Rx for ECG monitoring artifact. *Crit Care Nurse.* 1984; 4:64. [PubMed: 6559660]
4. Murdoch TB, Ketsky AL. The inevitable application of big data in health care. *JAMA.* 2013; 309:1351. [PubMed: 23549579]
5. The Big Data Research and Development Initiative. http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf
6. Merelli I, Perez-Sanchez H, Gesing S, D'Agostino D. Managing, analysing, and integrating big data in medical bioinformatics: Open problems and future perspectives. *DioMed Research International.* 2104 Article ID 134024/ <http://dx.doi.org/10.1155/2014/134023>.
7. Peek N, Holmes JH, Sun J. Technical challenges for big data in biomedicine and health: Data sources, infrastructure and analytics. *Yearb Med Inform.* 2014; 9(1):42–47. [PubMed: 25123720]
8. Zhang B, Wang Y, Chen F. Multilabel image classification via high-order label correlation driven active learning. *IEEE Trans Image Process.* 2014; 23:1430–41. [PubMed: 24723538]
9. Zhu Y, Zhang S, Liu W, Metaxas DN. Scalable histopathological image analysis via active learning. *Med Image Comput Comput Assist Interv.* 2014; 17(Pt 3):369–76. [PubMed: 25320821]
10. Bishop, CM. *Pattern recognition and machine learning.* Springer; New York: 2006.
11. Acharya UR, Sree SV, Ribeiro R, et al. Data mining framework for fatty liver disease classification in ultrasound: A hybrid feature extraction paradigm. *Med Phys.* 2012; 39:4255–2264. [PubMed: 22830759]
12. Acharya UR, Sree SV, Muthu Rama Krishnan M, et al. Automated classification of patients with coronary artery disease using grayscale features from left ventricle echocardiographic images. *Comput Methods Programs Biomed.* 2013; 112:624–632. [PubMed: 23958645]
13. Suzuki K. *Machine Learning in Computer-Aided Diagnosis of the Thorax and Colon in CT: A Survey.* *IEICE Trans Inf Syst.* 2013; E96-D(4):772–783. [PubMed: 24174708]

14. Halford JJ, Schalkoff RJ, Zhou J, et al. Standardized database development for EEG epileptiform transient detection: EEGnet scoring system and machine learning analysis. *J Neurosci Methods*. 2013; 212:308–316. [PubMed: 23174094]
15. Kim S, Hamilton R, Pineles S, Bergsneider M, Hu X. Noninvasive Intracranial Hypertension Detection Utilizing Semisupervised Learning. *Biomedical Engineering, IEEE Trans Biomed Eng*. 2013; 60:1126–1133.
16. Zweigenbaum P, Lavergne T, Grabar N, Hamon T, Rosset S, Grouin C. Combining an expert-Based Medical entity Recognizer to a Machine-Learning system: Methods and a case study. *Biomed Inform Insights*. 2013; 6(Suppl 1):51–62. [PubMed: 24052691]
17. Kruppa J, Liu Y, Biau G, et al. Probability estimation with machine learning methods for dichotomous and multicategory outcome: Theory. *Biom J*. 2014; 56:534–563. [PubMed: 24478134]
18. Mohri, M.; Rostamizadeh, A.; Talwalkar, A. *Foundations of machine learning*. MIT press; 2012.
19. Saeed M, Villarroel M, Reisner AT, et al. Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): a public-access intensive care unit database. *Crit Care Med*. 2011; 39:952–960. [PubMed: 21283005]
20. Li Q, Clifford GD. Signal quality and data fusion for false alarm reduction in the intensive care unit. *J Electrocardiol*. 2012; 45:596–603. [PubMed: 22960167]
21. Clifford GC, Scott DJ, Villarroel M. User Guide and Documentation for the MIMIC II Database. MIMIC-II database version 2.6. Rev: 291 Last Changed Date: 2012-02-24 15:53:51 -0500 (Fri, 24 Feb 2012).
22. Aboukhalil A, Nielsen L, Saeed M, Mark RG, Clifford GD. Reducing false alarm rates for critical arrhythmias using the arterial blood pressure waveform. *J Biomed Inform*. 2008; 41:442–451. [PubMed: 18440873]
23. Lehman LH, Adams RP, Mayaud L, Moody GB, Malhotra A, Mark RG, Nemati S. A physiological time series dynamics-based approach to patient monitoring and outcome prediction. *IEEE J Biomed Health Inform*. 2014; (99):1. PP. doi:10.1109/JBHI.2014.2330827.
24. Heldt T, Kashif FM, Sulemanji M, O'Leary HM, du Plessis AJ, Verghese GC. Continuous quantitative monitoring of cerebral oxygen metabolism in neonates by ventilator-gated analysis of NIRS recordings. *Acta Neurochir Suppl*. 2012; 114:177–180. [PubMed: 22327688]
25. Hug C, Clifford GD, Reisner AT. Clinician blood pressure documentation of stable intensive care patients: an intelligent archiving agent has a higher association with future hypotension. *Crit Care Med*. 201; 39:1006–1014. [PubMed: 21336136]
26. Boumbarov, O.; Velchev, Y.; Sokolov, S. ECG personal identification in subspaces using radial basis neural networks. *Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications; 2009 IDAACS 2009 IEEE International Workshop on IEEE*; 2009. p. 446-451.
27. Paul JS, Reddy MR, Kumar VJ. A transform domain SVD filter for suppression of muscle noise artifacts in exercise ECG's. *IEEE Trans Biomed Eng*. 2000; 47:654–663. [PubMed: 10851809]
28. Marque C, Bisch C, Dantas R, Elayoubi S, Brosse V, Perot C. Adaptive filtering for ECG rejection from surface EMG recordings. *J Electromyog Kinesiol*. 2005; 15:310–315.
29. Lu G, Brittain J-S, Holland P, et al. Removing ECG noise from surface EMG signals using adaptive filtering. *Neurosci Lett*. 2009; 462:14–19. [PubMed: 19559751]
30. Ko BH, Lee T, Choi C, Kim YH, Park G, Kang K, Bae SK, Shin K. Motion artifact in the electrocardiogram using adaptive filtering on behalf of half cell potential monitoring. *Conf Proc IEEE Eng Med Biol Soc*. 2102; 2012:1590–3. doi: 10.1109/EMBC.2012.6346248. [PubMed: 23366209]
31. Hamilton PS, Curley M, Aimi R. Effect of adaptive motion-artifact reduction on QRS detection. *Biomed Instrum Technol*. 2000; 34:197–202. [PubMed: 10868261]
32. Thakral, A.; Wallace, J.; Tomlin, D.; Seth, N.; Thakor, NV. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2001*. Springer; 2001. Surgical motion adaptive robotic technology (SMART): Taking the motion out of physiological motion; p. 317-325.
33. Chong JW, Dao DK, Salehizadeh SM, McManus DD, Darling CE, Chon KH, Mendelson Y. Photoplethysmograph signal reconstruction based on a novel hybrid motion artifact detection-

- reduction approach. Part I: Motion and noise artifact detection. *Ann Biomed Eng.* Nov; 2014 42(11):2238–50. doi: 10.1007/s10439-014-1080-y. Epub 2014 Aug 5. [PubMed: 25092422]
34. Tsien CL. Event discovery in medical time-series data. *Proc AMIA Symp.* 2000:858–62. [PubMed: 11080006]
 35. Hu X, Sapo M, Nenov V, et al. Predictive combinations of monitor alarms preceding in-hospital code blue events. *J Biomed Inform.* 2012; 45:913–921. [PubMed: 22465785]
 36. Cao H, Norris P, Ozdas A, Jenkins J, Morris JA. A simple non-physiological artifact filter for invasive arterial blood pressure monitoring: a study of 1852 trauma ICU patients. *Conf Proc IEEE Eng Med Biol Soc.* 2006; 1:1417–1420. [PubMed: 17946044]
 37. Görges M, Winton P, Koval V, et al. An evaluation of an expert system for detecting critical events during anesthesia in a human patient simulator: A prospective randomized controlled study. *Anesth Analg.* 2013; 117:380–391. [PubMed: 23780423]
 38. Güiza F, Depreitere B, Piper I, Van den Berghe G, Meyfroidt G. Novel methods to predict increased intracranial pressure during intensive care and long-term neurologic outcome after traumatic brain injury: Development and validation in a multicenter dataset. *Crit Care Med.* 2013; 41:554–64. [PubMed: 23263587]
 39. Siebig S, Kuhls S, Imhoff M, Langgartner J, Reng M, Scholmerich J, Gather U, Wrede CE. Collection of annotated data in a clinical validation study for alarm algorithms in intensive care—a methodologic framework. *J Crit Care.* 2010; 25:128–135. [PubMed: 19327311]
 40. Bonafide CP, Sander M, Graham CS, We rich Paine CM, Rock W, Rich A, Roberts KE, Fortino M, Nadkarni VM, Lin R, Keren R. Video methods for evaluating physiologic monitor alarms and alarm responses. *Biomed Instrum Technol.* 2014; 48:220–30. [PubMed: 24847936]
 41. Siebig S, Kuhls S, Imhoff M, Gather U, Schölmerich J, Wrede CE. Intensive care unit alarms—How many do we need? *Crit Care Med.* 2010; 38:451–456. [PubMed: 20016379]
 42. Kleinberg S, Elhadad N. Lessons learned in replicating data-driven experiments in multiple medical systems and patient populations. *AMIA Annu Symp Proc.* Nov 16.2013 2013:786–95. eCollection 2013. [PubMed: 24551375]
 43. Goldstein B, McNames J, McDonald BA, et al. Physiologic data acquisition system and database for the study of disease dynamics in the intensive care unit. *Crit Care Med.* 2003; 31:433–441. [PubMed: 12576948]
 44. Silaganesan A, Manley G, Huang MC. Informatics for neurocritical care: challenges and opportunity. *Neurocrit Care.* 2014; 20:132–141.
 45. Della MV, Maddalena E, Mizzaro S, Machin P, Beltrami CA. Preliminary results from a crowdsourcing experiment in immunohistochemistry. *Diagn Pathol.* Dec.2014 9(1):1069. doi: 10.1186/1746-1596-9-S1-S6.
 46. Ranard BL, Ha YP, Meisel ZF, Asch DA, Hill SS, Becker LB, Seymour AK, Merchant RM. Crowdsourcing--harnessing the masses to advance health and medicine, a systematic review. *J Gen Intern Med.* 2014; 29:187–203. [PubMed: 23843021]

Fig 1. Timeframe for the experiment. All data points within the full cardiorespiratory instability (CRI) epoch were considered, from the time the vital sign first crossed threshold, until values fell back below threshold

Fig 2.
Two part process for expert rule derivation followed by model building

<p>A. BLOOD PRESSURE ARTIFACT RULES</p> <p>BP Artifact Rule 1—Immediate decrease in pulse pressure (PP= SBP-DBP) to 30 mm of Hg or less --DBP increases with PP drop while SBP remains the same or high --This artifact is from the patient but is not correctly measured --This type of artifact might represent patient agitation and as such might be an independent measure of instability</p>
<p>BP Artifact Rule 2—Immediate step increase in systolic blood pressure (SBP), mean and diastolic blood pressure (DBP) --Rise in DBP values close to SBP, PP minimal to nonexistent --This type of artifact may also represent patient movement or tensing arm</p>
<p>BP Artifact Rule 3—If measurement is repeated, signal returns to normal patterns --The initial blood pressure (BP) on startup may be inaccurate, if a repeat reading is done in the first 30 minutes, then we ignore the first reading --This type of artifact may also represent patient movement or tensing arm</p>
<p>B. RESPIRATORY RATE ARTIFACT RULES</p>	
<p>RR artifact Rule 1—The presence of a RR signal in the absence of a HR signal is an artifact --No ECG signal is the root cause of the erroneous pleth signal</p>
<p>RR artifact Rule 2—Poor signal capture of the impedance—a step increase and decrease—wide changes, cyclical, oscillatory-like --As such, the value of the RR is incalculable --Instantaneous oscillatory step increase up and down of > 30 breaths a minute --This artifact can give us some information—it might indicate that the patient is agitated or that there is sweat under the electrode --Cause may be patient movement</p>
<p>C. HEART RATE ARTIFACT RULES</p> <p>HR Artifact Rule 1—A step increase or decrease in HR to a stable new state and a step decrease back to the same state with no change in other variables --This signal usually generates a new plateau, causing a square wave pattern --Cause may be patient movement leading to change in lead sensitivity doubling the R wave count.</p>
<p>D. SPO2 ARTIFACT RULES</p>

Fig 3.

Business rules developed and used to describe artifact

Fig 4.
Example of a visual rule for SpO₂, and feature derivation and application

Fig 5. Example of the ranking of numeric features (as listed in Table 2) for SpO₂ alerts in Block 1, according to the associated Vital Sign signals from which the features were derived. Features are derived from various aspects of the oxygen saturation by pulse oximetry (SpO₂), pulse pressure (PP), systolic blood pressure (SysBP), respiratory rate (RR) and heart rate (HR). Metric used is the Area Under the Receiver Operating Curve (AUC)

Fig 6.

Mean Area Under the Receiver Operating Characteristic Curve (AUC) scores from 10-fold cross-validation (CV) experiments on the array of algorithms for three categories of VS events. Training and validation (10-fold CV) on the Block 1 (dashed lines) data is displayed, along with the mean AUC scores when testing the algorithms on the Block 2 test set data (solid lines). The best performing algorithms vary across the vital sign type: Random Forest for respiratory rate (RR), Logistic Regression for blood pressure (BP), Naïve Bayes for peripheral oxygen saturation by pulse oximetry (SpO₂) (Results and standard deviations also in Table 4 electronic supplement)

Fig 7.

The Receiver Operating Characteristic (ROC) diagrams showing the performance of the best model learned from Block 1 data (Random Forest for RR, Logistic Regression for BP and Naïve Bayes for SpO₂) when tested on Block 2 data for the three vital sign event categories with respect to the True Positive Rate (TPR) relative to the False Positive Rate (FPR; 7A) and the True Negative Rate (TNR) relative to the False Negative Rate (FNR, 7B).

Respiratory rate (RR, solid line) was the easiest to classify, peripheral oxygenation by pulse oximetry (SpO₂, dashed line) was the most difficult to discriminate. Performance on blood pressure (BP, dotted line) is in between. Positive refers to true alerts, and negative to artifact.

Table 1

Summary of the step-down unit (SDU) patient, monitoring, and event data for the sample.

	Block 1 N=308	Block 2 N=326	p-value	Total
Patients				
Total	308	326		634
% male (N, %)	180 (58.4%)	191 (58.6%)	p=0.970	371 (58.5%)
Age (mean years±SD)	57.42 ± 20.2	57.92 ± 19.7	p=0.749	57.68 ± 19.95
Race (N, %)				
White	221 (71.8%)	240 (73.6%)	p=0.868	461 (72.7%)
Black	43 (14%)	42 (12.9%)		85 (13.4%)
Other	44 (14.3%)	44 (13.5%)		88 (13.9%)
Charlson Deyo Index (mean±SD)	0.94 ± 1.45	1.23 ± 1.59	p=0.244	1.09 ± 1.53
SDU length of stay (mean days ± SD)	4.70 ± 4.21	1.97 ± 0.84	p<0.001	3.31 ± 3.31
Hospital length of stay (mean days ± SD)	9.17 ± 9.8	8.93 ± 16.67	p=0.827	9.05 ± 13.76
Monitoring				
Total monitoring hours	22,588	19,046	p=0.759	41,635
Mean hours per patient	80	79		80
Median hours per patient	60	51		55
Events without persistence requirement				
Total events	271,288	362,849	p<0.001	634,137
Total events by subtype				
HR	11,952 (4%)	8,429 (2%)		20,381 (3%)
RR	101,589 (38%)	54,100 (15%)		155,689 (25%)
SpO ₂	61,242 (23%)	111,106 (31%)		172,348 (27%)
Systolic BP	54,679 (20%)	92,416 (25%)		147,095 (23%)
Diastolic BP	41,826 (15%)	96,798 (27%)		138,624 (22%)
Events with persistence requirement				
(Tolerance 40 s, length 240 s, duty cycle 80%)				
Total events	812	1,521		2,333

	Block 1 N=308	Block 2 N=326	p-value	Total
Total events by subtype			p=0.001	
HR	95 (11%)	55 (3%)		150 (6%)
RR	362 (44%)	640 (42%)		1,002 (43%)
SpO ₂	259 (32%)	648 (43%)		907 (39%)
BP	96 (12%)	178 (12%)		274 (12%)
Event Annotation by Experts				
Total Events	812	1,521		2,333
Total Events by subtype			p=0.311	1,399 (60%) 934 (40%)
Real alerts	493 (60%)	906 (60%)		
Artifact	319 (40%)	615 (40%)		
Real alerts total	493	906		1,399
Real alerts by subtype			p<0.001	137 (10%) 693 (50%) 425 (30%) 144 (10%)
HR	88 (18%)	49 (5%)		
RR	226 (46%)	467 (52%)		
SpO ₂	131 (27%)	294 (32%)		
BP	48 (9%)	96 (11%)		
Artifact total	319	615		934
Artifact by subtype			p<0.001	13 (1%) 309 (33%) 482 (52%) 130 (14%)
HR	7 (2%)	6 (1%)		
RR	136 (43%)	173 (28%)		
SpO ₂	128 (40%)	354 (58%)		
BP	48 (15%)	82 (13%)		

Table 2

List of feature categories, the aspects of the vital signs signals those features were meant to capture, and 44 feature names and descriptions that were derived in Step 1C of Figure 1.

Feature Category	Aspect of VS Signals Features Meant to Capture	Feature Name	Feature Description
Basic Statistics	Basic profile of the event	Mean	mean
		Sd	standard deviation
		Cv	coefficient of variance (sd/mean)
		Mad	median of absolute deviation from median [median (abs(x-abs(x)))]
		N	number of data points within alert period
		Min	min value
		Max	max value
		Median	median
		Range	max-min
		range_ratio	range/median
Data Density	Sparseness of signals	time_length	time elapsed during alert period
		data_den	number of valid data points/length of alert period
		data_den_trail	trailing data density at the beginning of alert, i.e. for 15 min before the start of alert [# data points/(timestamp for last data point - timestamp for first data point)]
		data_den_trail2	trailing data density at the beginning of alert i.e. for 15 min before the start of alert [# data points/(15*60)]
		delta_t	time elapsed since the end of last chunk of data before alert start
Spectrum Features	Frequency and power	max_gap	max gap between data points (also indicator of data density)
		spec_ratio	power of high frequency/lower frequency
		max_spec	max power frequency
Change Features	Change of VS in current event period from immediately before event	delta_mean	difference of mean (current alert period vs. 4 mins before start of alert)
		delta_sd	difference of sd (current alert period vs. 4 mins before start of alert)
		MW_stat	statistics from wilcoxon test to compare distribution within alert

Feature Category	Aspect of VS Signals Features Meant to Capture	Feature Name	Feature Description
Variance features	Smoothness of signals		period and those from 4 mins before
		MW_pvalue	pvalue from wilcoxon test to compare distribution within alert period and those from 4 mins before
		KS_stat	statistics from Kolmogorov-Smirnov test to compare distribution within alert period and those from 4 mins before
		KS_pvalue	pvalue from Kolmogorov-Smirnov test to compare distribution within alert period and those from 4 mins before
		t_stat	statistics from t test to compare mean within alert period and those from 4 mins before
		t_pvalue	pvalue from t test to compare mean within alert period and those from 4 mins before
		F_stat	statistics from F test to compare sd within alert period and those from 4 mins before
		F_pvalue	pvalue from F test to compare sd within alert period and those from 4 mins before
		Slope	linear slope during alert period
		Rslope	robust slope during alert period
		slope1	slope for the period before change point
		slope2	slope for the period after change point
		num_breakpoint	number of break points detected
		diff1_max	max first order difference
		diff1_min	min first order difference
		max_grad	max of slope for a series of moving window during alert period
		min_grad	min of slope for a series of moving window during alert period
		quad_rsq	rsquared from a quadratic fitting
		quad_coef1	first order coefficient of quadratic fitting
		quad_coef2	2nd order coefficient of quadratic fitting
		quad_resvar	variance of residuals from quadratic fitting
		osi_up	number of uptick (subject to certain threshold in both time and variable) within alert period
		osi_down	number of down tick (subject to certain threshold in both time and variable) within alert period

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Feature Category	Aspect of VS Signals Features Meant to Capture	Feature Name	Feature Description
		osi_ratio	and variable) within alert period
			ratio of osi_up(osi_up+osi_down)

Summary of the data Block 1 vital sign events by leading vital sign signal and the pruned list of their features used in the machine learning algorithms based on an Area Under the Curve (AUC) cut-point.

Table 3

Vital Sign	Number of events	Percentage artifact	AUC cutoff point for pruning	Features used	AUC score in feature ranking
RR	362	38.4%	<0.85	HR_data_den HR_N HR_data_den_trail HR_data_den_trail_2 RR_median	0.96 0.95 0.89 0.89 0.87
BP	96	37.5%	< 0.7	DiaBP_max PP_mean PP_median MeanBP_range DiaBP_mean	0.74 0.72 0.72 0.72 0.70
SpO ₂	259	14.2%	< 0.7	SpO ₂ _median SpO ₂ _sd SpO ₂ _cv SpO ₂ _range_ratio SpO ₂ _max SpO ₂ _min SpO ₂ _range SpO ₂ _quar_resvar SpO ₂ _min SpO ₂ _mean SpO ₂ _quad_rsqr	0.81 0.80 0.80 0.80 0.79 0.79 0.75 0.74 0.71 0.71
HR			There were too few HR artifacts for model building		

Abbreviations: RR, respiratory rate; HR, heart rate; BP, blood pressure; SpO₂, oxygen saturation by pulse oximetry.