

Making ICU Alarms Meaningful: a comparison of traditional vs. trend-based algorithms

Roy Schoenberg MD, Daniel Z. Sands MD MPH, Charles Safran MD

Center for Clinical Computing

Beth Israel Deaconess Medical Center, Harvard Medical School, Boston MA

Abstract

Much of the work in the ICU revolves around information that is recorded by electronic devices. Such devices typically incorporate simple alarm functions that trigger when a value exceeds pre-defined limits. Depending on the parameter followed, these "boundary based" alarms tend to produce vast numbers of false alarms. Some are the result of false reading and some the result of true but clinically insignificant readings. We present a computerized module that analyzes real-time data from multiple monitoring devices using a customizable logic engine. The module was tested on 6 intensive care unit patients over 5 days, running alarm algorithms for heart rate, systolic and diastolic blood pressure as well as arterial oxygen saturation. Results show a ten-fold increase in positive predictive value of alarms from 3% using monitor alarms to 32% using the module. The module's overall sensitivity was 82%, failing to detect 18% of significant alarms as defined by the ICU staff. The results suggests that implementation of such methodology may assist in filtering false and insignificant alarms in the ICU setting.

Introduction

Patient care in the ICU revolves around multiple monitors and devices. Alarms in these devices are put to heavy use in an attempt to detect and prevent clinical deterioration.¹ Such alarms typically trigger when the current reading (a single signal) exceeds a preset boundary.² The yield of those alarms depends on their sensitivity and specificity as well as the conditioning of the ICU staff to respond to them. Such conditioning is hampered by the plethora of false alarms produced by most devices, in some cases over 90% of all alarms.³ This number includes both technically false alarms as well as alarms that are based on true reading but are considered to be clinically insignificant.⁴

It comes as no surprise, that nurses and physicians, frustrated by the flood of noise,⁵ implement their own techniques⁶ of alarm filtering. This is done either by silencing notorious devices or setting alarm limits that are unlikely to be exceeded.^{7,8} This attempt can be justified in view of the negative behavioral

conditioning that results from multiple false alarms.⁹ As of today, no study has quantified the morbidity afflicted on patients as a result of this action but it is likely that some events are missed when alarms are silenced or disregarded. The motivation to develop smarter alarm systems will probably be proportional to the cost of this excess morbidity.

Much work was done in an attempt to reduce the number of insignificant alarms.¹⁰ Such work typically targeted one of three aspects of the problem:

1. Avoiding erroneous readings.¹¹
2. Filtering meaningless readings.^{12,13} (e.g., during intervention or calibration)
3. Introducing logic to filter clinically insignificant readings.

Whereas the first two challenges are relatively straightforward, the issue of establishing medical logic has always been problematic. Ideally, one would assess the "significance" of an alarm by judging the number of adverse events it prevented. Since this number is unknowable, this judgement is very difficult to make. Many other parameters are taken into account when reacting to an alarm and the logic behind that judgement is not always reproducible. In many cases it is not the alarm itself but rather the clustering of repetitive alarms from multiple devices that attracts attention. Similar physiological readings (that trigger the same set of alarms) may have different significance depending on the patient state and nature of illness. Reaching a consensus on the "right" logic to identify a significant alarm and the definition of significance is an unachievable goal.

In this work, we tried to implement an alternative approach to the problem. Instead of focusing on device readings and their limits we tried to identify physiological trends that are detrimental to the patient. We argue that such trends can be easier to follow if we look at the data in more detail. It might be very difficult to automatically identify septic shock in a patient, but it is not difficult to detect some of the physiological trends of this state. Each one of those trends in itself is not enough to suggest septic

shock but if identified in parallel, may be sufficient to warrant an alarm. It is likely that a concomitant decrease in systolic blood pressure, narrowing of the pulse pressure and a negative trend in body temperature should be pointed out to the staff. This methodology is effective both in identifying “significant” trends and in filtering “noise” alarms. The nature of erroneous readings (“noise”) is that they are randomly distributed and average out on summation. Such readings are unlikely to produce a trend and will never be correlated with other physiological readings. If alarms are based on the identification of a combination of trends, most of those readings will be filtered.

The notion of identifying physiological trends is not new and many algorithms have been published over the years in medical and artificial intelligence literature.¹⁴ A significant drawback in the design of ICU devices was the way in which they are implemented—they are discrete units not interconnected. One of the cornerstones of trend identification is the availability of sufficient data. Employment of a meaningful trend algorithm requires that data be saved over time. If such a trend is to be intersected with others, data from all different monitors and devices has to be gathered in one place. Until recently, the only place where this occurred has been the paper flow chart that possessed no processing power. Clearly, device communication and a central database must be brought into the ICU before this methodology of alarm filtering is attempted.¹⁵ It was not until recently that PC-based computer systems targeted these goals and allowed real-time data acquisition from multiple devices.¹⁶ The availability of data in a programmable environment made implementation of the old algorithm idea possible.¹⁷

This work investigates the possible added value of a computerized module that allows alarm logic to be constructed using multiple signals, trends and formulas. The aim of such a module is to point out those alarms designated as “significant” by the ICU staff and filter out all others. In other words, alarms generated by the module should have a better positive predictive value (PPV or the likelihood that a generated alarm is significant) than the PPV of boundary-based alarms generated by ICU devices.

Methods

The alarm module was developed as part of iMD Soft's MetaVision paperless ICU suite.¹⁸ The MetaVision system acquires the data directly from monitors, ventilators and infusion pumps in intervals of five seconds and records it into a central SQL

database. Each reading (a signal) is identified by type, time and patient ID. Our module then queries this database in an attempt to satisfy the logic in its algorithms and to identify significant alarms.

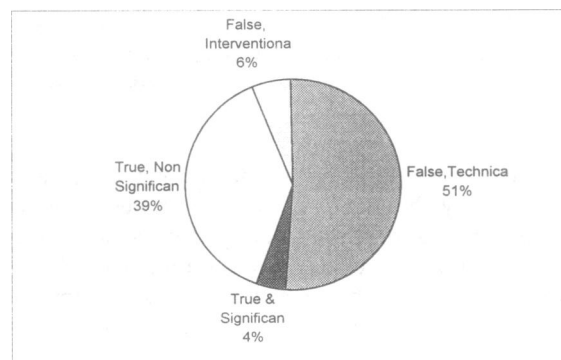
The construction of an algorithm is divided into several steps. First, the user defines the physiological trend. For example, the change in the average heart rate over one minute between the current minute and three minutes ago. Then a criterion is chosen for this physiologic change (e.g., > 15 beats per minute). The evaluation of this criterion can be true, false or unknown due to missing values. Each of these outcomes is assigned a score. The sum of scores at any point in time is compared with a threshold value to determine whether or not an alarm should be activated.

To evaluate our algorithm module's ability to detect significant alarms, we needed to compare it with a list of truly significant alarms during the same time period. For that purpose, we (authors and medical students) manually documented all device alarms and asked the ICU staff to classify them immediately as they occurred. Device alarms were thus classified by the staff as either false or true, where the latter were further classified as significant or insignificant. We then compared the PPV of alarms generated by our algorithm module to PPV of alarms generated by ICU devices in identifying the subset of clinically significant alarms.

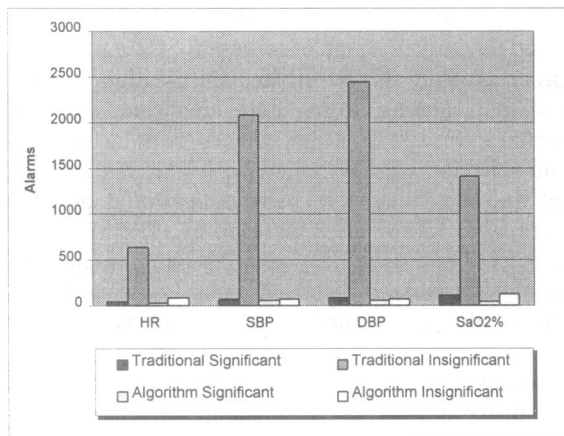
We ran our test on data generated from patients in five ICU beds over five days (120 hours).

Results

Six patients were followed during the study period for a total of 337 patient-hours. During that time, 482,453 physiologic measurements were collected and 6,872 alarms were recorded for heart rate,



systolic and diastolic blood pressure and oxygen saturation. 56.9% (3,912) of alarms were categorized by the ICU staff as false and 43.1% (2,960) as true. 4.2% (297) of all alarms were categorized as significant. The positive predictive value for traditional alarms was thus 3.8% (95% CI 2.9-4.7%). Application of our module to the same events elicited 544 alarms. Two hundred and forty two alarms (44.4%) corresponded to 81.4% of alarms that had been designated as significant by the ICU staff, while 302 (55.6%) did not correspond to documented significant alarms and were classified as insignificant. The positive predictive value of the algorithm module alarms was 35.7% (95% CI 24-46%). The overall difference in positive predictive value between the traditional alarm system and the algorithm system was found to be 31.9%.



Discussion

We evaluated the performance characteristics of an algorithm-based alarm module incorporating many different physiologic parameters compared to traditional single parameter boundary-based alarms. In trying to evaluate an alternative to current alarm methodology, we used the clinical judgment of the staff for each alarm as the gold standard. Although staff judgment is not perfect, we assumed that on a large number of alarms (the study population), mistakes in judgement are randomly distributed and would not cause a directional bias. We showed that our system substantially improved the positive predictive value of ICU alarms.

The main limitation of this design is that the study population of alarms is determined by one of the competing methods, i.e., all alarms included in the study (both insignificant *and* significant) are generated (and detected) by the traditional alarm system built into ICU devices. As a result, the device

alarm system will have 100% sensitivity by design. That does not limit our ability to make inferences about positive predictive value but we are unable to draw conclusions about negative predictive value.

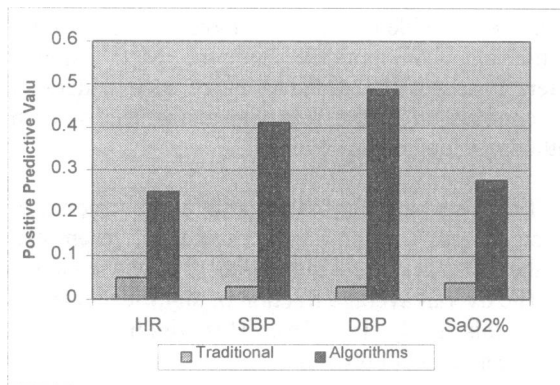
We assumed that by utilizing alarms that incorporate more data, we would be able to filter false and insignificant alarms. Rapid changes in readings that swiftly return to normal are more likely non-physiological than those which are persistent and form a trend. Slow sinusoidal fluctuations outside normal limits may be less significant than steep changes within those limits. Reciprocal changes that quickly return to their starting point may be a sign of measurement instability whereas recurrent cumulative change calls for early attention. These patterns can only be measured if the scope of data considered is wide enough. We have used the ability of the algorithm module and the MetaVision™ system to assess the added value of such widening. The comparison of the resulting alarm sets of both methods yielded the following:

1. The distribution of alarms into the true/false, significant /insignificant categories resembled the ratios described in previously published data.
2. The overall average fraction of significant alarms in this study is 4.2%. This finding is also consistent with previous work.
3. Algorithms (as used in this study) showed a minimum of 74% sensitivity to significant alarms, depending on the parameter followed, and 81.7% overall. This finding is significant in view of the data-sampling rate of once every five seconds used to feed the algorithm system. It suggests that a shift from continuous to intermittent sampling may not significantly alter alarm sensitivity and can be allowed to an extent. This observation is extremely important when system design issues like data storage size and query performance are addressed. Future work can possibly address the question of whether algorithms can identify significant alarms not found by the traditional system.
4. Algorithms filtered over 95% of false alarms.
5. Positive predictive value for algorithm-generated alarms was more than 10 times higher than the PPV of traditional alarm systems.

This data supports, at the very least, the assumption that the use of algorithms is effective in filtering false alarms. This ability can in part be attributed to the following:

1. Averaging values over times decreases the incidental effect of noise.

2. Attaching higher weight to persistent changes and trends emphasizes physiological changes that tend to have such a pattern and filters incidental changes that are random in nature.
3. Algorithms are filtering readings that are clearly erroneous in value. This capability to set multiple ranges for signal classification sets algorithms apart from traditional monitor alarms that have only two boundary settings (high and low).
4. The system's inability to collect data in intervals shorter than 5 seconds gives more credence to physiological (continuous) trends over technical reading errors (that tend to be erratic).



Although 81.7% of significant alarms were identified, the missing 18.3% must be accounted for. Possible reasons include:

Missing data -

The algorithms are accurate enough to point out all significant alarms but the sampling interval of five seconds prevented them from accomplishing this task to the full extent of their capabilities. Short events are sampled fewer times and their contribution to averages and trends is masked by the marginal readings. Shortening the sample interval acts here is a double-edged sword, increasing noise and sensitivity at the same time. If we increase sample interval to 1 second, we will in fact imitate the traditional monitors (which produce a continuous signal) and reach 100% sensitivity with considerable more noise. Finding the optimum sampling rate may in itself serve as a topic for further work.

Bad algorithms -

Changing averaging time ranges, tweaking value boundaries or changing mathematical calculation of trend slopes may each have a dramatic effect on the number of alarms produced by the algorithm.

Changing the weight contributed by each statement or the trigger value for the alarm as a whole will further modulate the behavior of the system. Adding new statements that address frequency or recurrence of events may evoke the missing alarms. It was not the purpose of this study to evaluate different algorithms but rather to investigate the potential benefit of their use. It is clear however, that "playing" with such algorithms seems attractive and may potentially be extremely rewarding.

Missing alarms were insignificant -

This assumption questions one of the rules used in designing this work, that the staff's classification is the gold standard. Although controversial, it is reasonable to assume a certain degree of error in the classification process. In order to examine this assumption, we could have returned to the staff and ask them to validate their decisions in retrospect. The results of such validation would be fascinating but beyond the limited scope of this study. Here too, such a process could also result in identification of new significant alarms, also not identified by the algorithm. In addition, we would then be able to avoid selection bias by reviewing all alarms and not only the ones that were missed. It would also be intriguing to follow the logic used by the staff in the process of their re-validation of the alarms and the possible similarities between such logic and the structure of the algorithms used in this study.

The ability to repeatedly examine vast amount of data from different sources is the unique quality of the computer system. Exploiting this ability to identify physiological processes does not require additional work from the ICU staff and is not dangerous as it does not replace another trusted system. These qualities characterize it as a decision support rather than a decision-making system.

Many attempts to build such generic systems (i.e. Medical Logic Modules) have not gained momentum mainly due to lack of supportive data acquisition infra-structure. The effort to create a data-fed system translates in many cases to high financial costs and an overall unfeasibility of the project. For such systems to be developed and deployed in today's reality, they must first prove to be capable of saving money, no less than to be clinically useful or statistically sound. Alarm systems are one of the few cases where such direct correlation can be made relatively easily. In a world where large investments are made to integrate clinical data repositories and to provide better consolidation of patient data, there may be a place for the first steps of "smart" systems that utilize real time data to prevent patient morbidity.

References

1. Patient data management systems in anaesthesia: an emerging technology. Weiss YG; Cotev S; Drenger B; Katzenelson R, *Can J Anaesth*, 1995 Oct, 42:10, 914-21
2. Pulse oximetry in ventilated preterm newborns: reliability of detection of hyperoxaemia and hypoxaemia, and feasibility of Alarm settings. Paky F; Koeck CM ; *Acta Paediatr*, 1995 Jun, 84:6, 613-6
3. Poor prognosis for existing monitors in the intensive care unit. Tsien CL; Fackler JC ; *Crit Care Med*, 1997 Apr, 25:4, 614-9
4. Auditory Alarms during anesthesia monitoring with an integrated monitoring system. Block FE Jr; Schaaf C ; *Int J Clin Monit Comput*, 1996 May, 13:2, 81-4
5. A CQI approach to the investigation of noise levels within the intensive care unit environment. Stephens C; Daffurn K; Middleton S ; *Aust Crit Care*, 1995 Mar, 8:1, 20-3, 26
6. A study of the incorrect use of ventilator disconnection Alarms. Campbell RM; Sheikh A; Crosse MM ; *Anaesthesia*, 1996 Apr, 51:4, 369-70
7. Clinicians' opinions on Alarm limits and urgency of therapeutic responses. Koski EM; M?kivirta A; Sukuvaara T; Kari A ; *Int J Clin Monit Comput*, 1995 May, 12:2, 85-8
8. Response times to visual and auditory alarms during anesthesia. Morris RW; Montano SR; *Anesth Intensive Care*, 1996 Dec, 24:6, 682-4
9. Human probability matching behaviour in response to Alarms of varying reliability. Bliss JP; Gilson RD; Deaton JE ; *Ergonomics*, 1995 Nov, 38:11, 2300-12
10. Are there too many Alarms in the intensive care unit? An overview of the problems. Meredith C; Edworthy J ; *J Adv Nurs*, 1995 Jan, 21:1, 15-20
11. Clinical evaluation of a prototype motion artifact resistant pulse oximeter in the recovery room. Dumas C; Wahr JA; Tremper KK ; *Anesth Analg*, 1996 Aug, 83:2, 269-72
12. Verification & validation algorithms for data used in critical care decision support systems. Carlson D; Wallace CJ; East TD; Morris AH ; *Proc Annu Symp Comput Appl Med Care*, 1995: 188-92
13. Symbiosis of nurse and machine through fuzzy logic: improved specificity of a neonatal pulse oximeter Alarm. Bosque EM ; *ANS Adv Nurs Sci*, 1995 Dec, 18:2, 67-75
14. Building intelligent Alarm systems by combining mathematical models and inductive machine learning techniques Part 2-- sensitivity analysis. Muller B; Hasman A; Blom JA ; *Int J Biomed Comput*, 1996 Aug, 42:3, 165-79
15. Use of large databases for resolving critical care problems. Belzberg H et al ; *New Horiz*, 1996 Nov, 4:4, 532-40
16. Computers in critical care. East TD; Wallace CJ; Morris AH; Gardner RM; Westenskow DR ; *Crit Care Nurs Clin North Am*, 1995 Jun, 7:2, 203-17
17. The benefits and challenges of an electronic medical record: much more than a "word-processed" patient chart. Sujansky WV; *West J Med*, 1998 Sep, 169:3, 176-83
18. MetaVision™ - The paperless ICU system. IMD-Soft Inc. www.imd-soft.com