A Web-Based Survey for Expert Review of Monitor Alarms

Benedikt Baumgartner[1], Kolja Rödel[1], Ulrich Schreiber[2], Alois Knoll[1]

[1] Technische Universität München, Munich, Germany
[2] German Heart Center Munich, Munich, Germany

Abstract

Bedside monitors trigger alarms if the state of a patient needs attention. However, many studies have shown that there is a high rate of false or clinically irrelevant alarms. Despite technological and methodological advances false alarms persist in today's monitoring systems. There is still a lack of annotated, standardized and publicly available alarm databases that would foster promising results of studies that use data mining methods, knowledge-based systems or alike for alarm rate reduction. We present a web-based survey conducted at the German Heart Center Munich in which physicians had to rate given monitoring events. Our results unveiled significant differences in the ratings within the group of participants (deviations of 21.16%) but also in comparison to an established alarm set of a second group of reviewers (31.28%). Manual alarm ratings depend on subjective assessments and are therefore error-prone. Our web-interface gives access to reviewers on a large scale. This allows substantial analysis and can help to validate existing alarm databases as well as to establish new ones.

1. Introduction

High rates of false or clinically irrelevant alarms triggered by bedside monitors are a persistent problem in intensive care. False alarm rates of up to 94% have been reported [1]. Multiple studies in the last decades document the negative impact of false alarms on both patients and staff which affects quality of care and patient safety. E.g. Gabor et al. [2] reported sleep disorders leading to a slowed recovery. Imhoff and Kuhls [3] give a comprehensive overview on the alarm situation in critical care and review methods from statistics and computer science for alarm rate reduction. They criticize that still many alarms are simple threshold alarms generating false positives without clinical meaning. The rate of false alarms has basically not changed over the last 20 years despite technological advances. However, approaches in the field of machine learning, artificial intelligence and knowledge-based methods have shown promising results [4–6]. A common problem in these studies is the limited size of training and test data. Developing and evaluating methods on a small dataset can lead to good results which are not representative though. In fact, a representative and standardized database of monitoring alarms is essential to develop, test and compare methods for alarm validation and reduction. Therefore Aboukhalil et al. [7] introduced a standard set of critical ECG alarms. They reviewed 5386 critical alarms from 447 patients in the MIMIC II database [8]. Each alarm was labeled as true, false or ambiguous by two experienced annotators. Mismatched alarms, where the reviewers disagreed, were labeled again by a third experienced user. In the follwing we refer to their dataset as the MIT set. In the second part of their work they tested an algorithm that uses morphological and timing information from the arterial blood pressure (ABP) curve to suppress critical ECG alarms. They reported alarm-specific suppression rates between 33.0% and 93.5% and showed the potential of data fusion for false alarm reduction. In a recent study we evaluated several data mining methods on a subset of their alarm set and were able to increase those suppression rates (75.24% - 99.23%) [9]. The results in both studies strongly depend on the training and test sets. Obviously alarms of the MIT set are unequally distributed (see Figure 1). Asystole alarms naturally occur less frequently than

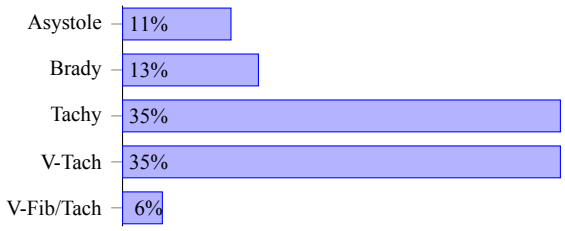

Figure 1. Distribution of alarm types in the MIT set (5386 alarms in total).

tachycardia alarms, for example. Therefore an even larger annotated database is desirable. However, the task of reviewing the alarms is time-consuming, error-prone and relies on the knowledge of few domain experts. Comparing several studies collecting and reviewing monitoring data also reveals different experience levels of annotators rang-

ing from medical students to experienced physicians [10]. To overcome those deficiencies we introduce a web-based survey for alarm annotation and show results of a preliminary study with 30 participants. The web interface allows access to a large number of reviewers and the validation of reviewed alarms.

2. Methods

2.1. Data source

We included annotated alarms from the MIT set as well as critical ECG alarms from the freely available MIMIC II database that have not been reviewed yet. In accordance with Aboukhalil et al. [7] we chose 5 alarm types: Asystole, Bradycardia (Brady), Tachycardia (Tachy), Ventricular Tachycardia (V-Tach) and Ventricular Fibrillation/Tachycardia (V-Fib/Tach). We only included alarms with an ECG and ABP signal present at the time of the alarm and considered a range of 15 s before and 5 s after the alarm event. If available, we added the pulmonary arterial pressure (PAP) curve and numerical values for SpO2 and central venous pressure (CVP). Furthermore the sex and age of the patient was retrieved from the database. We supplemented the information with the current heart rate, calculated from the ECG signal, and systolic, diastolic and mean values for ABP and PAP in order to provide participants with a familiar monitor screen.

2.2. Survey design

A web interface based on Java Server Pages was designed to present alarm events to the participants and to collect their annotations. We asked each participant to review 100 alarm events. Based on test runs the duration of the survey was approximately 20 min. 20 alarms were chosen from the MIT set and therefore already had labels, 35 events were identical for all participants. Another 35 events were unique for each user while 5 alarms were shown twice during the survey in order to check the consistency of each participant's ratings. The design of the page was inspired by common patient monitors showing ECG and blood pressure curves and the above mentioned numerical data. The layout is shown in Figure 2. Each user received a personal login. If participants logged out before finishing their reviews, the current event was saved and users could continue the survey when they logged in again. A status bar shows the current process of the survey. Reviewers had four possibilities for alarm categorization: (1) no action from medical staff is necessary (false alarm), (2) medical staff intervention is necessary within the next 15 min, (3) immediate intervention is needed and (4) no comment.

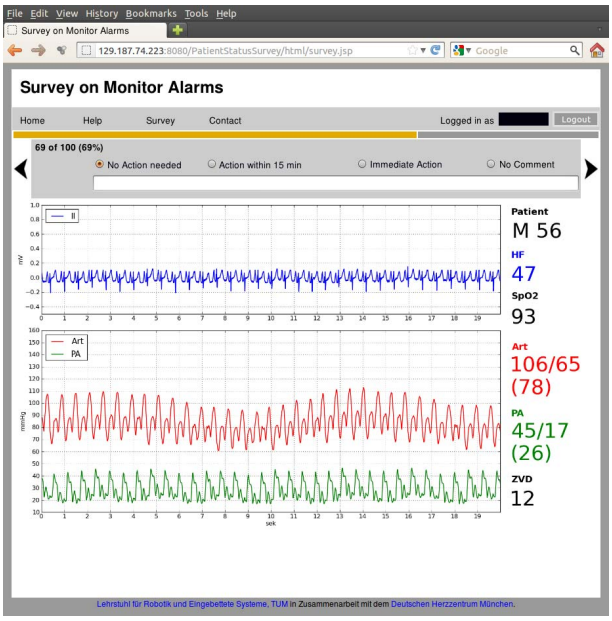

Figure 2. Screenshot of the online survey showing an alarm event for review.

3. Results

3.1. Participants

We asked 42 physicians from the German Heart Center Munich (GHC) to review monitoring alarms. So far 30 people participated. 15 users labeled all 100 monitor situations that were presented to each of them. For analysis on an individual level (section 3.4) we included 23 experts that reviewed more than 97 alarm events each. There was one student reviewer with more than 2 years of practice in the intensive care unit (ICU). All other participants were physicians with at least 2 and up to 30 years of experience in intensive care monitoring. Before they started reviewing the monitoring alarms their experience level (low, medium, high) was recorded. Within this group 5 reviewers judged their experience as low, 8 had medium and 10 users had a high experience level.

3.2. Inter-institutional level

Since a subset of the MIT standard was relabeled by the participants we were able to compare the two groups of reviewers on an intra-institutional level. Alarms labeled as "Action required within 15min" and "Immediate Action required" were considered as true alarms in the comparison. Reviewers from GHC tended to label more alarms as true (395) compared to the MIT (327). Overall, the two groups differed on 31.28% of the common alarm situations.

Table 1 shows the deviations per alarm type. Noticeable are the differences for Asystole and V-Fib/Tach alarms. 18 V-Fib/Tach events, from a total of 32, originally labeled as false (F) by the MIT were labeled as true (T) by the GHC. Label consistency between the two groups was highest for Tachy alarms (18.32% deviation).

Alarm Type	MIT(T)/ GHC(T/F)	MIT(F)/ GHC(T/F)	Total	Deviation
Asystole	8/(8/0)	12/(9/3)	20	45%
Brady	45/(35/10)	6/(3/3)	51	25.50%
Tachy	164/(148/16)	27/(19/8)	191	18.32%
V-Fib/Tach	7/(6/1)	25/(18/7)	32	59.38%
V-Tach	103/(88/15)	89/(61/28)	192	39.58%
Total	**327/(285/42)**	**159/(110/49)**	**486**	**31.28%**

Table 1. Deviations in the Assessments of GHC and MIT per alarm type.

3.3. Intra-institutional level

On an intra-institutional level we compared ratings within the group of the GHC. As stated above, 35 events were identical for all reviewers. Table 2 shows the label distribution of the top and bottom five alarm events ordered by the confidence level of the rating. The confidence level displays the percentage of the label that was selected by the majority of the users. E.g. 23 participants labeled a V-Tach alarm as true (Table 2, rank 1), yielding a confidence level of 100% while a majority of 13 reviewers out of 25 considered another V-Tach alarm as false, yielding a confidence level of 52% (Table 2, rank 34). Deviations in the number of total ratings for an event are due to the fact that not all participants completed the survey. Average confidence level of the alarm ratings was 78.84%.

3.4. Individual level

During the survey 5 events were presented twice to each participant. This allowed a validation of the given ratings on an individual level and verification that participants did not select labels randomly. In the analysis we only included participants that labeled all 5 pairs, which made up a total of 23 reviewers. Figure 3 displays the consistency of the individual ratings. Both plots show the number of physicians with respect to the quantity of equally labeled alarm events. Plot (a) differentiates between all four label categories (see 2.2). 4 participants labeled all 5 pairs consistently while 3 users gave identical ratings on 2 pairs. Average consistency was 73.91%. In plot (b), the categories of (2) "action within 15min" and (3) "immediate action" were combined as they both indicate a true alarm. We could observe an average consistency rate of 86.09% with identical ratings of all pairs by 10 participants.

The results on the individual level were also analyzed with respect to the experience level of each user. It was

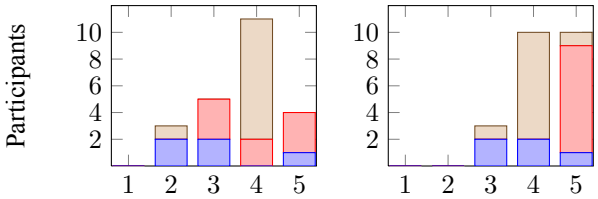

Figure 3. Consistency of individual ratings in terms of equally labeled pairs and with respect to reviewers' experience level (low , medium , high). (a) considers 4 label categories while (b) combines categories (2) and (3).

observed that 4 out of 5 participants with experience level "low" gave equal labels on 3 or less pairs (Figure 3(a)). 9 out of 10 experienced users reviewed 4 pairs equally while there was no clear trend for users with medium experience. This changed when alarm categories (2) and (3) were combined (Figure 3(b)). Here all participants with medium experience reviewed the complete set of 5 pairs consistently. The distribution of experienced users did not change significantly. However, all participants rated at least 3 pairs equally now.

4. Discussion and conclusion

Our results show that experienced physicians are more consistent in their ratings compared to beginners. Participants with medium experience were able to differentiate between true and false alarms, but there was an uncertainty whether the patient needs immediate assistance or only within 15 min. However, also experienced users were not able to give equal reviews on all identical pairs. A lack of concentration during the survey can be assumed which might be due to the duration of the survey or missing motivation of participants. Therefore it is essential to have alarm situations reviewed by more than just one expert. If we look at the intra-institutional level, we observe strong deviations for specific alarm situations. The bottom part of Table 2 illustrates that there are alarm situations which are difficult to label independent of the users' experience level. Surprisingly none of the participants labeled those situations as unknown. This shows that manual labeling depends on subjective estimates and ratings are debatable even in a large group of annotators. For a final decision on arguable monitor events additional data such as patient history or medication might be needed. On an inter-institutional level we found a rather high deviation of almost one third (31.28%). A possible reason for that is the limited scope of the monitor that was presented in the survey (15 s before and 5 s after the event). In contrast, experts working on the MIT set had the possibility to

Rank	Type	False Alarm	True Alarm	Unknown	Total	Confidence Level
1	V-TACH	0	23	0	23	100%
2	V-FIB/TACH	0	24	0	24	100%
3	BRADY	0	25	0	25	100%
4	BRADY	0	25	0	25	100%
5	TACHY	0	25	0	25	100%
⋮	⋮	⋮	⋮	⋮	⋮	⋮
31	V-TACH	15	11	0	26	57.69%
32	BRADY	13	11	0	24	54.17%
33	V-TACH	12	13	0	25	52%
34	V-TACH	13	12	0	25	52%
35	V-TACH	12	12	0	24	50%

Table 2. Intra-Institutional Deviations of the GHC. Shown are label distributions for the top and bottom five alarm events ordered by confidence level with respect to the total number of reviewers.

scroll forward or backward. The rationale for this limitation was that developed methods for alarm rate reduction would also have no possibility for taking into account future monitoring values. It is noticeable that participants from the GHC tended to confirm monitor alarms in contrast to the MIT where more alarms were assumed to be false. People working at the GHC might be more cautious about specific alarm types (e.g. Asystole) due to the patients, recovering from heart surgery, they work with on a daily basis.

We presented a web-based survey for labeling triggered monitor alarms. Annotated alarm databases are needed in order to develop methods for alarm rate reduction. The gold standard is the knowledge of the physician taking into account the monitored data as well as other patient specific data. But results show that reviews of alarm events should be done by more than just a few experts in order to obtain trustworthy labels. The framework allows access to a large number of reviewers from all over the world and can help to validate existing databases as well as to establish new ones.

Acknowledgements

We would like to thank the many participants of the survey. We especially thank Dr. Stefan Eichhorn and Matthias Kornek for their suggestions regarding the survey design. This work was supported by the TUM Graduate School of Information Science in Health (GSISH).

References

[1] Görges M, Markewitz BA, Westenskow DR. Improving alarm performance in the medical intensive care unit using delays and clinical context. Anesthesia Analgesia 2009; 108(5):1546–1552.

[2] Gabor JY, Cooper AB, Crombach SA, Lee B, Kadikar N, Bettger HE, Hanly PJ. Contribution of the intensive care unit environment to sleep disruption in mechanically ventilated patients and healthy subjects. American Journal of Respiratory and Critical Care Medicine 2003;167(5):708–715.

[3] Imhoff M, Kuhls S. Alarm algorithms in critical care monitoring. Anesth Analg 2006;102(5):1525–1537.

[4] Koski E, Sukuvaara T, Mkivirta A, Kari A. A knowledge-based alarm system for monitoring cardiac operated patients-assessment of clinical performance. International Journal of Clinical Monitoring and Computing 1994; 11(2):79–83.

[5] Oberli C, Urzua J, Saez C, Guarini M, Cipriano A, Garayar B, Lema G, Canessa R, Sacco C, Irarrazaval M. An expert system for monitor alarm integration. Journal of Clinical Monitoring and Computing 1999;15(1):29–35.

[6] Tsien CL. Event discovery in medical time series data. In In American Medical Informatics Association (AMIA) Symposium. 2000; .

[7] Aboukhalil A, Nielsen L, Saeed M, Mark RG, Clifford GD. Reducing false alarm rates for critical arrhythmias using the arterial blood pressure waveform. Journal of Biomedical Informatics June 2008;41(3):442–451.

[8] Saeed M, Lieu C, Raber G, Mark RG. Mimic ii: a massive temporal icu patient database to support research in intelligent patient monitoring. In Computing in Cardiology, number 29. 2002; 641–644.

[9] Baumgartner B, Rödel K, Knoll A. A data mining approach to reduce the false alarm rate of patient monitors. In International Conference of the IEEE Engineering in Medicine and Biology Society. 2012; Under review.

[10] Siebig S, Kuhls S, Imhoff M, Langgartner J, Reng M, Schlmerich J, Gather U, Wrede CE. Collection of annotated data in a clinical validation study for alarm algorithms in intensive care - a methodologic framework. Journal of Critical Care March 2010;25(1):128–135.

Address for correspondence:

Benedikt Baumgartner
Technische Universität München
Department of Informatics
Boltzmannstr. 3
85748 Garching b. München
bb@tum.de