

Extracting Alarm Events from the MIMIC-III Clinical Database

Jonas Chromik¹^a, Bjarne Pfitzner¹^b, Nina Ihde¹^c, Marius Michaelis¹^d, Denise Schmidt¹^e,
Sophie Anne Ines Klopfenstein²^f, Akira-Sebastian Poncette²^g, Felix Balzer²^h
and Bert Arnrich¹ⁱ

¹Hasso Plattner Institute, University of Potsdam, Germany

²Charité – Universitätsmedizin Berlin, Berlin, Germany

Keywords: Patient Monitor Alarm, Medical Alarm, Intensive Care Unit, Vital Parameter, Data Cleaning, Data Extraction.

Abstract: Lack of readily available data on ICU alarm events constitutes a major obstacle to alarm fatigue research. There are ICU databases available that aim to give a holistic picture of everything happening at the respective ICU. However, these databases do not contain data on alarm events. We utilise the vital parameters and alarm thresholds recorded in the MIMIC-III database in order to artificially extract alarm events from this database. Prior to that, we uncover, investigate, and mitigate inconsistencies we found in the data. The results of this work are an approach and an algorithm for cleaning the alarm data available in MIMIC-III and extract concrete alarm events from them. The data set generated by this algorithm is investigated in this work and can be used for further research into the problem of alarm fatigue.

1 INTRODUCTION

Alarm fatigue is the desensitisation of medical staff due to an excessive number of alarms, most of them being false or irrelevant (McCartney, 2012). This results in a lack of response to the alarm stimulus. The problem of alarm fatigue has been widely investigated, both qualitatively (Cvach, 2012) and quantitatively (Drew et al., 2014).

However, building technical solutions for alleviating alarm fatigue is hindered by a lack of readily available data on patient monitor alarms. Public medical databases usually provide no data on patient monitor alarms as is the case with eICU CRD (Pollard et al., 2018) and HiRID (Hyland et al., 2020).

The MIMIC-III database (Johnson et al., 2016) provides alarm data. However, there are only alarm

thresholds recorded in the database and not the alarm events themselves. Alarm thresholds are lower and upper boundaries for a certain vital parameter, such as the heart rate (HR). For example, for HR the low alarm threshold might be set to 60 bpm and the high alarm threshold to 120 bpm. Whenever the measured parameter, i.e. HR, drops below the low alarm threshold or exceeds the high alarm threshold, an alarm event goes off at the patient monitor and alerts the medical staff.


The objective of this work is to extract these alarm events from the MIMIC-III database by taking into account alarm thresholds and the actual parameter value. This is done for the following vital parameters:


HR: heart rate, as measured by an electrocardiogram (ECG) and expressed in beats per minute (bpm)


NBP_s: non-invasively measured systolic blood pressure, as measured by a sphygmomanometer and expressed in millimeters of mercury (mmHg)


S_pO₂: peripheral blood oxygen saturation, as measured by a pulse oximeter (usually on the patients finger) and expressed in %


Furthermore, we uncover and rectify inconsistencies in the recorded alarm thresholds such as unrealistically high or low values or instances where the high


^a <https://orcid.org/0000-0002-5709-4381>


^b <https://orcid.org/0000-0001-7824-8872>


^c <https://orcid.org/0000-0001-5776-3322>


^d <https://orcid.org/0000-0002-6437-7152>

^e <https://orcid.org/0000-0002-6299-0738>

^f <https://orcid.org/0000-0002-8470-2258>

^g <https://orcid.org/0000-0003-4627-7016>

^h <https://orcid.org/0000-0003-1575-2056>

ⁱ <https://orcid.org/0000-0001-8380-7667>

alarm threshold is below the low alarm threshold.

The rest of this work is structured as follows: In Section 2 we describe which parts of the MIMIC-III database we are using, how we address data inconsistencies through data cleaning, and finally how we extract alarm events. In Section 3 we describe the alarm event data set that is produced as a result of this work. Finally, in Section 4 we discuss our findings as well as potential applications and limitations of this data set.

2 MATERIALS & METHODS

The MIMIC-III database contains 26 tables providing a wide range of information on the events at the intensive care units (ICUs) of Beth Israel Deaconess Medical Center. For our use case, however, only the CHARTEVENTS table is of interest. This table contains, among others, measured values and alarm thresholds of the vital parameters listed in Section 1.

The objective of this work is to extract alarm events by comparing the measured parameter values with the corresponding alarm thresholds. However, before the alarm events can be extracted, we have to deal with a number of data inconsistencies that we uncovered. The inconsistencies and our corresponding rectification approaches are presented in the following.

2.1 Data Cleaning

The CHARTEVENTS table contains a multitude of information, such as routine vital signs, ventilator settings, laboratory values, and mental status.¹ For extracting alarm events, however, only a small subset of this information is relevant, i.e. measurements of the three vital parameters we consider in this work (HR, NBP_s, and S_pO₂) as well as their respective high and low alarm thresholds. Hence, the first step in data cleaning is removing all irrelevant information by retaining only a specific subset of ITEMIDs which are listed in Table 1.

Invalid Value Removal. Besides removing irrelevant data items, we are also interested in validating the correctness of the relevant data items. For example, S_pO₂ can not exceed 100% and HRs above 350 bpm are rare. For the considered vital parameters, we found that their recorded values are not always within clinically valid ranges. We assume that these extreme values are either erroneous or bear a special but un-

¹<https://mimic.mit.edu/iii/mimictables/chartevents/>

Table 1: Complete list of ITEMIDs retained while filtering the CHARTEVENTS table with their respective label as recorded in the D.ITEMS table.

ITEMID	Label
220045	HR
220046	HR Alarm - High
220047	HR Alarm - Low
220179	NBP _s
223751	NBP _s Alarm - High
223752	NBP _s Alarm - Low
220277	S _p O ₂
223769	S _p O ₂ Alarm - High
223770	S _p O ₂ Alarm - Low

documented meaning. Therefore, we decided to remove values outside the clinically valid ranges listed in Table 2.

Invalid value removal is done both for the parameter measurements themselves and the threshold setting (both low and high) using the same ranges. The ranges are intentionally chosen to be conservative in order not to remove valid measurements or settings. For measurements, this removal means that the value at this point in time is missing afterwards but could be reproduced by interpolating between the neighbouring measurements. Concerning alarm thresholds, only changes of these thresholds are recorded in the MIMIC-III database. Hence, we do not have dedicated threshold information for each parameter measurement. Thus, for alarm threshold, invalid value removal means that the threshold update is lost and the previous threshold is carried over.

Table 2: Clinically valid ranges for the vital parameters considered in this work. Adapted from (Harutyunyan et al., 2019).

Parameter	Lower Limit	Upper Limit
HR	0	350
NBP _s	0	375
S _p O ₂	0	100

To show how the cleaning steps affect the original data set, we look at the distributions of thresholds and measurements for NBP_s. NBP_s serves only as an example here. The same cleaning steps were also performed for HR and S_pO₂. Figure 1 shows that the majority of threshold values and measurements are in the valid range. However, there is still a wide range of outliers with implausibly high values. These outliers are removed by the invalid value removal cleaning step. Figure 2 shows a rectified and more reasonable value distribution after the outliers are removed.

Apart from values outside clinically relevant ranges, we also found inconsistencies within the alarm thresholds. These inconsistencies are essen-

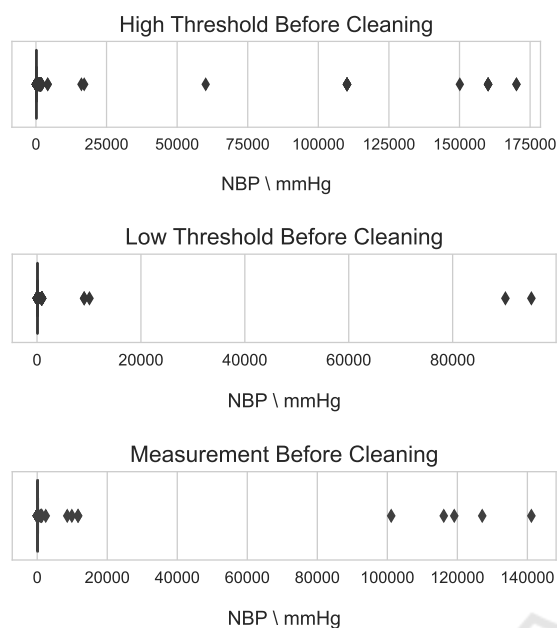


Figure 1: Boxplots showing the distribution of high alarm thresholds, low alarm thresholds and measurements before cleaning. The distribution is vastly skewed with the valid range barely visible at the far left corner and a wide range of outliers.

tially periods of time, where the high alarm threshold is below the low alarm threshold for a vital parameter. We further distinguish between exact threshold swaps and threshold overlaps which we both describe in the following.

Exact Threshold Swaps. In MIMIC-III, changes in corresponding high and low thresholds are always recorded simultaneously. Every newly recorded high threshold is associated with a low threshold being recorded at the same time and vice versa. At times, these thresholds are exactly swapped, i.e., the high thresholds taking the value of the low threshold and vice versa as shown in Fig. 2a. This, however, would create an alarm with every further measurement which is why we consider this to be an erroneous recording that needs rectification. Exact threshold swaps are easily identified and corrected by swapping the high and the low threshold as shown in Fig. 2b.

Threshold Overlaps. Besides the exact alarm threshold swaps, there are also cases where high and low alarm thresholds overlap but are not exactly swapped and therefore generate an alarm too. For example, a high alarm threshold might be set unreasonably low and falls below the corresponding low alarm threshold as shown in Fig. 3a. At the same time, the low alarm threshold continues to stay at a reasonable

value. Such an overlap is usually present for short periods only. These cases cannot be corrected by swapping. Therefore, since high and low alarm thresholds are always recorded pairwise in the MIMIC-III database, both high and low alarm thresholds are removed in the respective segment where they overlap. The last clinically meaningful alarm thresholds prior to the overlapping thresholds are chosen instead, as shown in Fig. 3b.

2.2 Extracting Alarm Events

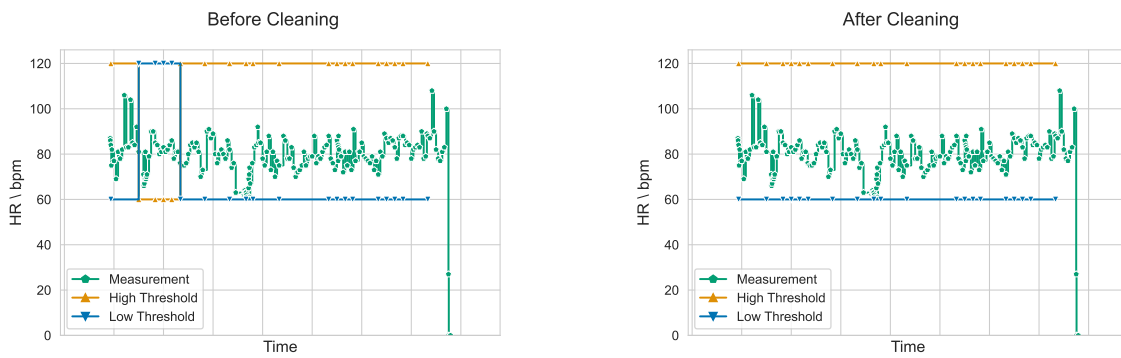
In Section 2.1 we described how we cleaned the CHARTEVENTS table from out-of-range values and inconsistencies. In this section, we show how we used the actual measurements of the vital parameters and their corresponding threshold setting in order to find actual alarm events in the data. As shown in Algorithm 1, we first isolated measurements and thresholds for each ICU stay (single stay of a single patient at the ICU) and each vital parameter. Then, we went through each high threshold setting and low threshold setting respectively and checked whether any measurement within the relevant time frame exceeded the high threshold or falls below the low thresholds. Whenever this happened, we return either a high or a low alarm event at the respective measurement's timestamp.

A shortcoming of this approach is that the number of alarms is subject to the sampling frequency of the respective vital parameter. Higher sampling frequencies produce more alarms because there are more measurements in a period of time where the vital parameter is out of range. Figure 4 shows the differences in the number of samples for the data items listed in Table 1. Clearly, HR and S_pO_2 are measured or at least recorded more often than NBP_s . This can result in an over-representation of HR and S_pO_2 alarms as compared to NBP_s alarms.

3 RESULTS

In Section 2 we describe methods and algorithms we used to clean the CHARTEVENTS table and extract concrete alarm events from it. This results in a data set containing all patient monitor alarms as per the MIMIC-III database in its respective observation period. In this section, we show some descriptive statistics that are made possible by the extracted data set of alarm events.

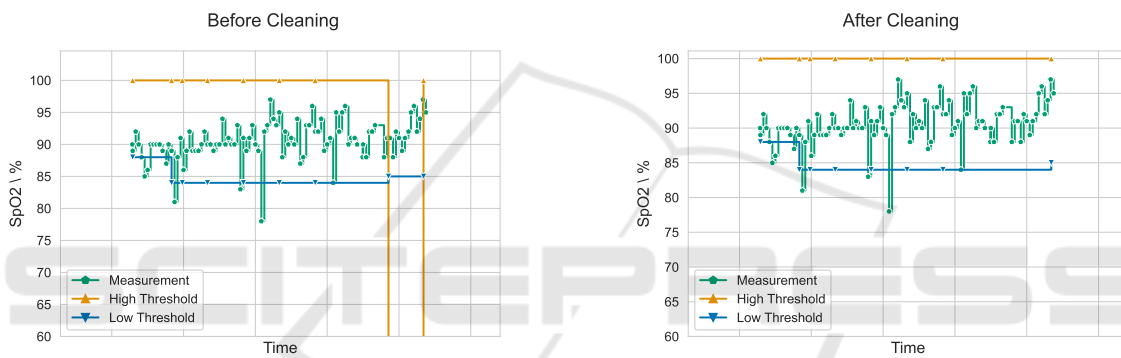
Parameters and Alarm Types. The data set gives insight into the relative counts of alarms produced



(a) Exactly swapped low and high thresholds before correction. Every measurement in the time period where the thresholds are swapped would theoretically produce an alarm.

(b) A data cleaning step removes the exact threshold swap thus rectifying the alarm threshold. No alarm events will be recognised in the respective time period.

Figure 2: Example for an exact threshold swap correction.



(a) In this case, the thresholds overlap without being exactly swapped. Here, the unreasonable low value for the high threshold would result in all measurements in the respective period of time triggering a high threshold alarm.

(b) Threshold overlap was corrected by removing the responsible alarm threshold settings. After correction, no high alarms are triggered in the respective period of time.

Figure 3: Example for threshold overlap correction.

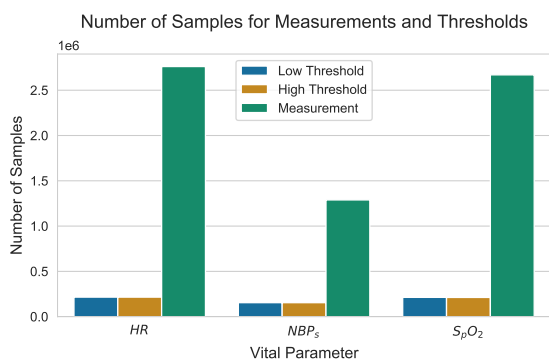


Figure 4: Comparison of the number of samples for measurements and thresholds for HR, NBP_s, and SpO₂. The number of measurements is much higher than the number of thresholds in all cases and there are distinct differences in the number of samples for the different vital parameter measurements.

by the different vital parameters. Comparing alarm counts among vital parameters might yield skewed results due to the differences in sampling frequencies, as already discussed in Section 2.2. However, comparing the counts of high and low threshold alarms for a single vital parameter yields interesting results. Figure 5 shows such a comparison. For HR and NBP_s, violations of the high threshold seem to occur more often than violations of the low threshold. However, for SpO₂ violations of the low threshold are a lot more common than violations of the high threshold. This is to be expected since a high blood oxygen saturation is rarely a problem while too low blood oxygen saturation is a harmful condition (Silverthorn, 2018).

We also want to emphasise the effect of the cleaning steps we performed on the alarm counts. Figure 6 shows the numerical reduction of alarms for each alarm type we considered. The alarms that are

Data: MIMIC-III CHARTEVENTS

Result: List of Alarm Events

```

foreach ICUSTAY do
  foreach Parameter do
    msmts := measurements for Parameter and ICUSTAY;
    highs := high threshold settings for Parameter and ICUSTAY;
    lows := low threshold settings for Parameter and ICUSTAY;
    foreach high in highs do
      foreach msmt in msmts do
        if time(high) <= time(msmt) < time(high+1) then
          if value(msmt) > value(high) then
            Return a high alarm event at msmt;
          end
        end
      end
    end
    foreach low in lows do
      foreach msmt in msmts do
        if time(low) <= time(msmt) < time(low+1) then
          if value(msmt) < value(low) then
            Return a low alarm event at msmt;
          end
        end
      end
    end
  end
end
  
```

Algorithm 1: Algorithm for extracting alarm events from measurements and thresholds.

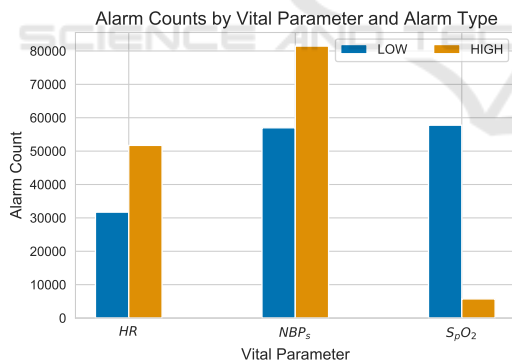


Figure 5: Comparison of alarm counts by vital parameter (i.e. HR, NBP_s, and SpO₂) and alarm type (i.e. whether a high or a low threshold was violated).

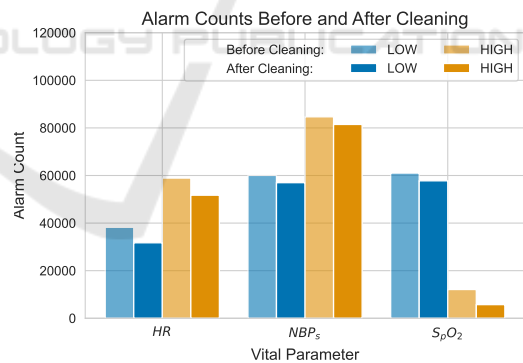


Figure 6: Alarm counts for low and high alarms regarding HR, NBP_s, and SpO₂. Extracted from an uncleaned and from a cleaned data set, respectively. This figure shows that there is a reduction in alarm count due to cleaning for each type of alarm.

not present as a result of the cleaning steps are supposedly false alarms. Hence, this shows that the cleaning step actually improves the quality of the generated data set, since the alarms not included after cleaning are supposedly false alarms.

Alarm Distribution among ICU Stays. The generated data set shows that the distribution of alarm events among the ICU stays seems to follow a Pareto

distribution. The majority of patients produce only a low number of alarms with the interquartile range (IQR) spanning from 3 to 16 alarms per ICU stay. However, there are few patients that are responsible for an excessively high number of alarms as can be seen in Fig. 7. We considered the 1% of ICU stays with the highest number alarms to be outliers and hence to not show them in the plot in an attempt to

show the distribution of the remaining 99% per cent more clearly.

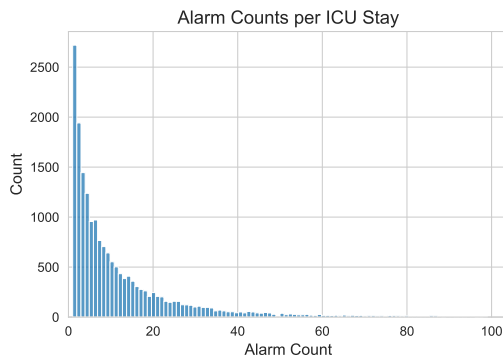


Figure 7: The distribution of alarm counts (only 99% shown) among the ICU stays follows a Pareto distribution with few patients generating a large number of alarms and many patients generating only few alarms.

Differences between Alarm Threshold and Measurement. Patient monitor alarms do not differentiate between strong and slight threshold violations. An alarm goes off whenever the measurement exceeds a high threshold or drops below a low threshold. For the patient monitor, it does not matter whether the difference between measurement and threshold is high or low. However, in clinical practice, the difference is relevant since a parameter slightly out of range is far less critical than a parameter that has by far left a physiologically healthy range. Therefore, we investigated the difference between measurement and threshold. Figure 8 shows this difference by the example of the S_pO_2 low threshold. Most of the alarms are caused by only a slight drop of the measurement below the threshold by a few per cent. On the other hand, large drops of the S_pO_2 parameter are rare. The same pattern of many low differences between measurement and threshold and few large differences are also to be found when looking at the other parameters, i.e. HR and NBP_s .

4 DISCUSSION

The analyses we have shown in Section 3 – although interesting – are only examples for the potential use cases of the data set that is created by the approach presented in this paper. Nevertheless, these results are relevant findings that can guide further research into alarm fatigue.

Structural Findings. One finding is that extensive post-processing in terms of cleaning and alarm ex-

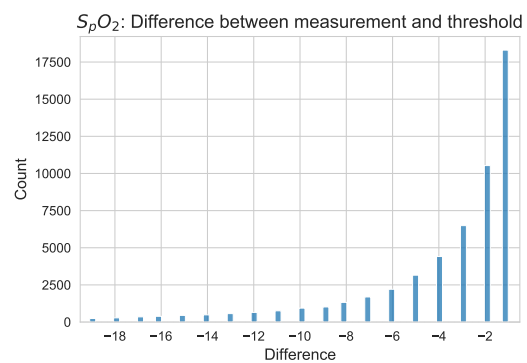


Figure 8: Histogram of the differences between alarm threshold and actual measurement for the S_pO_2 low threshold. The majority of alarms are triggered by a slight drop of the measurement below the threshold.

tracting is necessary to make sense from the alarm data in MIMIC-III. This calls for guidelines prescribing on how to appropriately provide alarm data. Vital parameters, alarm thresholds, and alarm events both in terms of threshold alarm and in terms of other alarms such as arrhythmia alarms need to be taken into account. Furthermore, data inconsistencies as uncovered in Section 2.1 need to be avoided. This can either be avoided on a device level by designing the interface of the patient monitor in a way that inconsistent thresholds are impossible to set. Or, a post-processing step is required to rectify or remove these inconsistencies.

Contentual Findings. Apart from findings related to the structure and consistency of the data, we also want to discuss the findings related to the content of the generated data set. Figure 7 shows that the majority of patients generate only a low number of alarms while few patients generate a large number of alarms. In order to alleviate alarm fatigue, it would be sensible to conduct further research into what causes these patients to generate far greater numbers of alarms. Further, we showed in Fig. 8 that the majority of alarms are caused by minimal threshold violations. This finding can be used to guide further research. For example, patient monitors could take the difference between measurement and threshold into account in order to adapt the volume of the alarm, as shown in (Greer et al., 2018). Another option would be to suppress or delay alarms caused by slight threshold violations in order to help the medical staff focus on more severe emergencies as (Schmid et al., 2013) and (Winters et al., 2018) find that alarm delays are an effective tool to reduce false alarms at the ICU.

Two design decisions are noteworthy in our approach to data cleaning. First, in the invalid value removal step, we remove measurements and thresholds

if and only if their values are outside the corresponding valid range. One result of this is that threshold updates might be partially removed, i.e. a high threshold update being removed while the corresponding low threshold is retained or vice versa. This is noteworthy because thresholds update originally occur only pairwise in the MIMIC-III database. We decided to remove only the invalid part of the threshold update in order to retain as much valid information as possible.

Second, when removing threshold overlaps, we decided to always remove both parts (high and low) of the threshold update because it is not always obvious whether one threshold part remains in a sensible range while the other part deviates or whether both parts deviate. This can not be determined without making strong assumptions about the nature of threshold updates. Hence, we decided to always remove both parts thus reverting the effective threshold to the last reasonable threshold update.

Limitations and Threats to Validity. The alarm event data set we generated from the MIMIC-III database provides some interesting insights into the problem of alarm fatigue in medicine. However, there are some limitations and threats to validity attached to our approach. The data quality of the generated alarm events data set is – apart from the cleaning steps we performed – limited by the data quality of the data set it is generated from. For example, the sampling frequencies for the data in the MIMIC-III database manifest an upper limit for the sampling frequencies in the alarm events data set. Furthermore, all changes in sampling frequency, missing data, etc. are also carried over into the alarm events data set. For example, higher sampling frequencies in the vital parameter measurements will result in a higher number of alarms. Since the sampling frequencies vary among vital parameters, as Fig. 4 shows, some alarm types (e.g. HR) might be over-represented. This has to be kept in mind when working with the data set.

Future Work. We already discussed the implications for alarm fatigue research of this work's findings as well as its limitations. Further work needs to be done in order to validate the finding from the MIMIC-III database. Especially, more extensive ICU databases are needed covering not only vital parameters, input and output events, laboratory findings, and hospital logistics but also providing data on ICU alarms.

Until such a database is created, the data set generated in this work can be used for a variety of purposes, some of them are demonstrated in Section 3. Among

others, this data set enables quantitative analyses on alarm events, alarm forecasting, and alarm threshold recommendation which are to be covered in future research.

5 CONCLUSION

The contribution of the paper is an approach and algorithm to generate alarm events from the MIMIC-III database. Publishing the generated data set itself would have been more convenient for researchers interested in data on alarm events. However, by publishing only the algorithm we ensure compliance with the data protection guidelines of the MIMIC-III database. Everyone with access to the MIMIC-III database can apply the algorithm to the database and thus create the alarm events data set themselves. The algorithms for data cleaning and alarm extraction are published on GitHub, see <https://github.com/HPI-CH/mimic-alarms>.

ACKNOWLEDGEMENTS

This work was partially carried out within the INALO project. INALO is a cooperation project between AICURA medical GmbH, Charité – Universitätsmedizin Berlin, idalab GmbH, and Hasso Plattner Institute. INALO is funded by the German Federal Ministry of Education and Research under grant 16SV8559.

REFERENCES

- Cvach, M. (2012). Monitor alarm fatigue: An integrative review. *Biomedical Instrumentation & Technology*, 46(4):268–277.
- Drew, B. J., Harris, P., Zègre-Hemsey, J. K., Mammone, T., Schindler, D., Salas-Boni, R., Bai, Y., Tinoco, A., Ding, Q., and Hu, X. (2014). Insights into the problem of alarm fatigue with physiologic monitor devices: A comprehensive observational study of consecutive intensive care unit patients. *PloS one*, 9(10):e110274.
- Greer, J. M., Burdick, K. J., Chowdhury, A. R., and Schlesinger, J. J. (2018). Dynamic alarm systems for hospitals (dash). *Ergonomics in Design*, 26(4):14–19.
- Harutyunyan, H., Khachatryan, H., Kale, D. C., Ver Steeg, G., and Galstyan, A. (2019). Multitask learning and benchmarking with clinical time series data. *Scientific Data*, 6(1):1–18.
- Hyland, S. L., Faltys, M., Hüser, M., Lyu, X., Gumbsch, T., Esteban, C., Bock, C., Horn, M., Moor, M., Rieck, B.,

- et al. (2020). Early prediction of circulatory failure in the intensive care unit using machine learning. *Nature Medicine*, 26(3):364–373.
- Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L.-W. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., and Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1):1–9.
- McCartney, P. R. (2012). Clinical alarm management. *MCN: The American Journal of Maternal/Child Nursing*, 37(3):202.
- Pollard, T. J., Johnson, A. E., Raffa, J. D., Celi, L. A., Mark, R. G., and Badawi, O. (2018). The *eICU* collaborative research database, a freely available multi-center database for critical care research. *Scientific Data*, 5(1):1–13.
- Schmid, F., Goepfert, M. S., and Reuter, D. A. (2013). Patient monitoring alarms in the *ICU* and in the operating room. *Annual Update in Intensive Care and Emergency Medicine 2013*, pages 359–371.
- Silverthorn, D. U. (2018). *Human Physiology: An Integrated Approach*. Pearson, 8th edition.
- Winters, B. D., Cvach, M. M., Bonafide, C. P., Hu, X., Konkani, A., O'Connor, M. F., Rothschild, J. M., Selby, N. M., Pelter, M. M., McLean, B., and Kane-Gill, S. (2018). Technological distractions (part 2): A summary of approaches to manage clinical alarms with intent to reduce alarm fatigue. *Critical Care Medicine*, 46(1):130–137.

