pHealth 2022
B. Blobel et al. (Eds.)

doi:10.3233/SHTI220988

A Visualization and Benchmarking Simulator for Clinical Data

Marlies MORGEN[a,1], Lejla BEGIC FAZLIC[a], Arne PEINE[c], Lukas MARTIN[c], Anke SCHMEINK[b], Ahmed HALLAWA[b] and Guido DARTMANN[a]
[a] *ISS, Trier University of Applied Sciences, Trier, Germany*
[b] *Chair of Inf. Theory & Data Analytics, RWTH Aachen University, Aachen, Germany*
[c] *Department of Intensive & Intermediate Care, University Hospital Aachen, Germany*

Abstract. The availability of Big Data has increased significantly in many areas in recent years. Insights from these data sets lead to optimized processes in many industries, which is why understanding as well as gaining knowledge through analyses of these data sets is becoming increasingly relevant. In the medical field, especially in intensive care units, fast and appropriate treatment is crucial due to the usually critical condition of patients. The patient data recorded here is often very heterogeneous and the resulting database models are very complex, so that accessing and thus using this data requires technical background knowledge. We have focused on the development of a web application that is primarily aimed at clinical staff and researchers. It is an easily accessible visualization and benchmarking tool that provides a graphical interface for the MIMIC-III database. The anonymized datasets contained in MIMIC-III include general information about patients as well as characteristics such as vital signs and laboratory measurements. These datasets are of great interest because they can be used to improve digital decision support systems and clinical processes. Therefore, in addition to visualization, the application can be used by researchers to validate anomaly detection algorithms and by clinical staff to assess disease progression. For this purpose, patient data can be individualized through modifications such as increasing and decreasing vital signs and laboratory parameters so that disease progression can be simulated and subsequently analyzed according to the user's specific needs.

Keywords. Big Data, Health Care, Simulation, Time Series, Anomaly Detection

Introduction

The SARS-Cov-2 pandemic has shown that hospitals can become overloaded very quickly. In the future, digital decision support systems will be able to relieve hospital staff. A major challenge is that research team members tend to be highly specialized. Clinical staff provide the medical expertise, while programmers and database specialists are technically proficient and mathematicians perform the calculations.

This article is dedicated to this topic. The open-source software presented here is a visualization and benchmarking simulator for clinical data (ViBeSiC)[2]. It enables time-based simulation, manipulation, and analysis of patient vital signs and laboratory data.

[1] Corresponding Author.
[2] The project is available at: https://gitlab.rlp.net/m.morgen/vibesic

The data come from the Medical Information Mart for Intensive Care (MIMIC)-III database, a well-known research database in the field of critical care [1]. Such datasets are elementary for research, especially for training artificial intelligence algorithms. Such algorithms can help, for example, to support ventilation strategies [2] or to make predictions about possible courses of disease [3][4].

1. State of the Art

Most of the applications for visualizing data in health care focus on the visual presentation of disease progression or the identification of specific patient cohorts [5][6]. Some tools require experience in programming or query languages, while others require knowledge of the structure and content of the MIMIC-III database e.g. PhysioBank's Automated Teller Machine (ATM) [7] or LightWAVE [8]. Qualified use of both applications requires knowledge about the databases and the data sets. In 2019 a tool that visualizes and extracts patient-specific time series from the MIMIC-III database was presented [9]. The MIMIC database also contains non-public records that users can access if they have been given access by the PhysioNet curators. Unlike the previously mentioned tools, in this application it is possible to include these records as well. There are no filtering options for patient selection, but the visual display is more appealing and simultaneous display of multiple patients is possible. The functionalities the previously mentioned applications are limited to visualization and extraction and do not include real-time simulation, modification or analysis of patient data.

There are tools which, like our software, aim to make it easier for inexperienced users to use MIMIC-III database. One web-based application [10] extracts and visualizes data regarding to a specific patient's stay within one timeline. The selection and preparation of the data is clear and intuitive. The application has a risk prediction function for the in-hospital mortality of patients where various features such as respiratory rate, oxygen saturation and recorded events are analyzed. For previous versions of the MIMIC database, there are other tools that, in addition to visualizing patient data, include predictive models for patient mortality [11].

2. Implementation of the Simulator

The MIMIC-III database was used for the visualization tool. This is a freely accessible database that contains information such as demographic data, vital signs and laboratory values for more than 40,000 patients treated in the ICU [1]. We added the MIMIC-III Waveform Database Matched Subset (Waveform) with records of physiological signals and vital signs [12]. We extracted a cohort of patients with matching records in MIMIC-III and Waveform. This task was complex due to issues such as missing values, non-matching timestamps and incorrect linkages. The database was enriched, for example, by the indicator of whether patients developed sepsis during their stay. For this purpose, we integrated research findings identified by Johnson et al. [13]. Furthermore, we added relevant attributes to avoid time-consuming calculations and SQL join clauses.

A major challenge in the research sector is that the skills of team members are often highly specialized. Clinical staff provide the medical expertise, while programmers are technical specialists and mathematicians perform the background calculations. Since visualizations can improve the comprehensibility of Big Data and thus support the

evaluation of algorithms, the developed patient simulator serves as a user interface for the application and validation of algorithms. For demonstration, we used the patients demo database [12].

The MIMIC-III database was created as a local PostgreSQL database and extended with Python scripts that generate and execute SQL commands to achieve the database state described above. The front end of the application is a dashboard, for which programming languages were combined due to the various components that need to communicate with each other. This approach is used to provide the functionality and to integrate different analyses (R, SQL, Python, Matlab), as well as to customize the layout (HTML, CSS) whose approaches are more extensive to configure or faster to execute in another language.

The web-based, interactive patient simulator allows time-based simulation of vital and laboratory data. It is a two-screen application, where the user can select, show and modify patient data. Within the first screen, a user chooses a cohort of patients based on criteria like age and diseases. This input is converted into an SQL query and applied to the database to extract suitable ICU stays (Figure 1).

Figure 1. Screen 1 – Patient overview with tab showing graph with vital values.

By selecting an ID, a summary of the patient's stay appears. Key data about the patient and a graphical representation of the vital signs appear. The user can switch tabs to get an overview of the laboratory parameters as well as notes recorded during the stay. In the second window, the selected ICU stay is animated as a time series simulation (Figure 2). It shows values of the heart rate, respiratory rate and arterial blood pressure (y-axis) depending on the selected point in time (x-axis). Tables underneath show all vital signs and lab values in relation to time. Furthermore, patient parameters can be manipulated through predefined modifiers. The manipulation starts from the selected point in time and continues until the end of the time series. Modifications affect and visualize a copy of the MIMIC-III database to preserve the original data, but they can be saved as new patients. To avoid mixtures of the original with the manipulated entries, a prefix is added to the original IDs and the data is stored in a second database. The storage

function allows to recall, further edit and analyze previously configured patient cases at a later time.

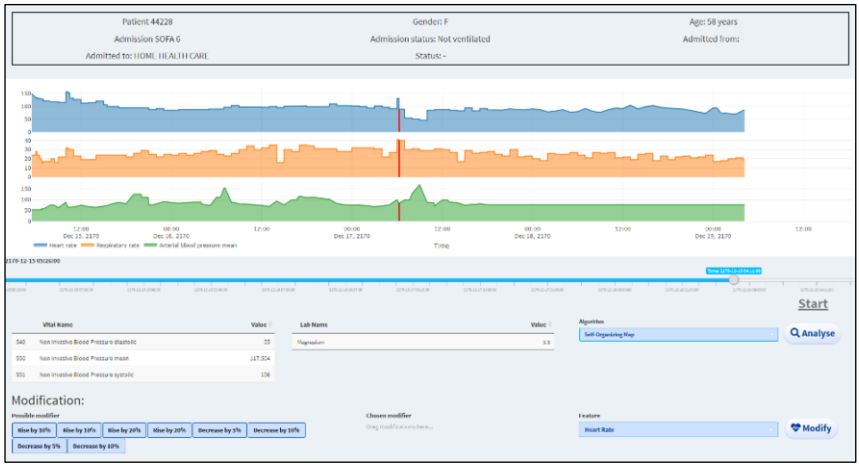

Figure 2. Screen 2 - Visualization of anomaly analysis of a modified patient data set.

3. Benchmarking

The simulation includes an anomaly detection module in addition to the visualization of the recorded patient values to indicate irregularities in the values in order to make the user aware of them. Some of the algorithms are based on statistical probabilities of low complexity to visualize outliers within a specific vital feature while others consider multiple features and work with trained networks to locate anomalies.

Statistical concepts include uni- and multivariate analyses. The first method uses the interquartile range (IQR) to identify outliers. This measure of dispersion is used to establish the upper and lower thresholds by which we determine when a value is either too high or too low compared to the other values [14]. When either condition is met, the corresponding value is comparatively exceptionally high or low (red line in Figure 2). This functionality can be used in combination with the modification of datasets described below, allowing the creation of disease trajectories for individual test scenarios.

Vital signs are not inherently subject to trends or seasonal variation. However, internal and external influences can cause regular fluctuations. External influences include, for example, the administration of medications that cause vital signs to respond. Internally, for example, the daily rhythm can have an influence. The resting pulse during sleep is generally lower than during waking hours. Therefore, the Moving Median Decomposition (GMD) approach was implemented [15]. A time series is decomposed into seasonal, trend, and random residual time series. The last two mentioned can be used to identify abnormal values. Moving average is often used to extract trends. Combining these analyses, which consider only one parameter at a time, allows for the inclusion of multiple characteristics. However, the disadvantage of statistical approaches is that single observations are made [16]. Each feature is analyzed separately and the anomaly can then either be an outlier of one feature or described by combining abnormal values of several features. The latter has the disadvantage that anomalies are not detected if only

one value deviates, regardless of how strong the deviation is. The Mahalanobis distance (MD) measures the relative distance of a data point to a centroid [17]. This is the point where all means form all features intersect and therefore it is considered the overall mean for the data. A data point is flagged as anomalous if it exceeds a certain quantile of the chi-squared distribution or, if it is identified by an adaptive process that looks for anomalous values in the tails of the distribution above a certain chi-squared quantile [18].

Machine learning approaches are also based on statistical concepts but can reconfigure themselves without manual effort. This means that the trained model is constantly being automatically adjusted. The hybrid algorithm was developed for sepsis prediction in MIMIC-III, but has also been applied to environmental sensors for testing purposes [19]. This model integrates several scientific fields by combining statistical methods, self-organizing maps (SOMs) [20], and linear discriminant analysis (LDAs) [21][22]. Dynamic Time Warping (DTW) finds the optimal alignment between two temporal sequences even if their speed varies using Levensthein distance. Because DTW does not include averaging techniques, Dynamic Time Warping Barycenter Averaging (DBA) was designed. The algorithm integrated here uses these two approaches [23]. The procedure involves the four steps Data Preparation, Statistical Transformation, DBA Algorithm, Classification and Validation. In addition, the probability of risk for positively classified patients is calculated using the conditional probability. The algorithm calculates the risk probability of a patient developing sepsis within a predefined period of time per time unit. There is no standard threshold value from which the risk is classified as too high, i.e., from which point a warning is necessary or useful. Due to this fact, in our case the user chooses this threshold value.

4. Discussion and Outlook

There are several applications for visualizing data from the MIMIC database, but their functionality is very limited. Our approach is a novel visualization tool that can easily extract patient time series and simulate it in real time for analysis within the simulator or for an external artificial intelligence (AI). It allows modification of the given data to artificially create anomalies and generate alerts for the user or an external AI. The current version of the simulator is based on the MIMIC-III database and all queries of the database are adapted to the structure of the MIMIC database. However, generalization is easily possible by adapting the queries and tables to another medical base according to their functionality. This application could also be used for real-time operation in hospitals. The use of the simulator is applicable to data sets other than medical. The algorithms used could detect inconsistencies ranging from simple measurement errors to the occurrence of complex events in sensory records. The functionality of anomaly detection in the time series is also extendable. Risk prediction and detection algorithms have been integrated into the simulator. These were validated not only on a group of patients but also by selecting individual patients according to the parameters of interest. This allowed young researchers (clinicians and computer science students) to work together to create different patient scenarios, analyze the data, validate the algorithm, and compare the results. Next step is to investigate and integrate further learning algorithms for anomaly detection, which can then also be executed directly in the simulator. Thus, the results of the different algorithms can be evaluated by the qualified user.

5. Acknowledgement

The results of this project were generated by funding from the BMBF under the funding codes 13GW0280C, 13GW0280D, and 13GW0280E and EIT Health grant 19549. These results are currently used in the BMEL project KI-Pilot (funding code 2820KI001) and the demonstrator is exhibited in the KI-Reallabor. We thank the BMEL for the continuation of the research especially in the area of anomaly detection and smart data platforms.

References

[1] Johnson AE, Pollard TJ, Shen L, Lehmann L-WH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data. 2016 May 24;3:160035.
[2] Peine A, Hallawa A, Bickenbach J, Dartmann G, Fazlic L, Schmeink A, et al. Development and validation of a reinforcement learning algorithm to dynamically optimize mechanical ventilation in critical care. NPJ Digit Med 2021 Feb 19;4(1):32.
[3] Hossain ME, Khan A, Moni MA, Uddin S. Use of Electronic Health Data for Disease Prediction: A Comprehensive Literature Review. IEEE/ACM Trans Comput Biol Bioinform, 2021; 18(2):745-758.
[4] Frazer J, Notin P, Dias M, Gomez A, Min JK, Brock K, et al. Disease variant prediction with deep generative models of evolutionary data. Nature 2021; 599: 91–95.
[5] Lee J, Ribey E, Wallace JR, A web-based data visualization tool for the MIMIC-II database. BMC Med Inform Decis Mak 2016 Feb 4;16:15.
[6] Harris DR, Henderson DW. i2b2t2: Unlocking visualization for clinical research. AMIA Jt Summits Transl Sci Proc 2016 Jul 20;2016:98-104.
[7] MIT Laboratory for Computational Physiology, PhysioBank ATM, PhysioNet. https://archive.physionet.org/cgi-bin/atm/ATM.
[8] Moody G. LightWAVE: Waveform and annotation viewing and editing in a web browser. Comput Cardiol (2010) 2013 Sep;40:17-20.
[9] Festag S, Spreckelsen C. A visualisation and extraction tool for time series in the MIMIC III database. Stud Health Technol Inform. 2019; 267: 134–141.
[10] Levy-Lambert D, Gong JJ, Naumann T, Pollard TJ, Guttag JV. Visualizing patient timelines in the intensive care unit. CoRR abs/1806.00397, 2018.
[11] Chen R, Kumar V, Fitch N, Jagadish J, Zhang L, Dunn W, et al. explICU: A web-based visualization and predictive modeling toolkit for mortality in intensive care patients. Annu Int Conf IEEE Eng Med Biol Soc 2015;2015:6830-3..
[12] Moody B, Craig M, Johnson A, Kyaw T, Moody G, Saeed M, et al. The MIMIC-III Waveform Database Matched Subset. physionet.org, 2017.
[13] Johnson AEW, Aboab J, Raffa JD, Pollard TJ, Deliberato RO, Cheli LA, et al. A comparative analysis of sepsis identification methods in an electronic database. Crit Care Med. 2018 Apr;46(4):494-499.
[14] Tukey JW. Exploratory data analysis, Addison-Wesley, 1977.
[15] Giovanis E. Moving Median with Trend and Seasonality. SSRN Electronic Journal March 2007. doi:10.2139/ssrn.969012, 2007.
[16] Kang H. The prevention and handling of the missing data, Korean J Anesthesiol. 64(5):402-6, 2013.
[17] Mahalanobis C. On the generalized distance in statistics. Proceedings of the National Institute of Sciences (Calcutta) 1936, 49–55.
[18] Filzmoser P, Garrett RG, Reimann C. Multivariate outlier detection in exploration geochemistry. Comput. Geosci. 2005, 579–587.
[19] Fazlic L, Hallawa A, Schmeink A, Lipp R, Martin L, Peine A, et al. A Novel Hybrid Methodology for Anomaly Detection in Time Series. Int. Journal of Computational Intelligence Systems 2022; 15, 50.
[20] Kohonen T. The self-organizing map, Proceedings of the IEEE, 1990, 1464–1480.
[21] Fisher RA. The use of multiple measurements in taxonomic problems. Annals of Eugenics 1936; 7(7): 179–188.
[22] Rao CR. The utilization of multiple measurements in problems of biological classification, Journal of the Royal Statistical Society - Series B 10(2), 1948, 159–203.
[23] Fazlic LB, Hallawa A, Dziubany M, Morgen M, Schneider J, Schacht M, et al. A machine learning approach for the classification of disease risks in time series. 9th Mediterranean Conference on Embedded Computing, MECO 2020, Budva, Montenegro, June 8-11, 2020, IEEE, 2020, pp. 1–5.