Closing the Data Loop: An Integrated Open Access Analysis Platform for the MIMIC Database

Mohammad Adibuzzaman[1], Ken Musselman[1], Alistair Johnson[2], Paul Brown[3], Zachary Pitluk[3], Ananth Grama[4]

[1]Regenstrief Center for Healthcare Engineering, Purdue University, West Lafayette, USA
[2]Laboratory for Computational Physiology, Massachusetts Institute of Technology, Cambridge, USA
[3]Paradigm4, Waltham, USA
[4]Department of Computer Science, Purdue University, West Lafayette, USA

Abstract

We describe a new model for collaborative access, exploration, and analyses of the Medical Information Mart for Intensive Care - III (MIMIC III) database for translational clinical research. The proposed model addresses the significant disconnect between data collection at the point of care and translational clinical research. It addresses problems of data integration, pre-processing, normalization, analyses (along with associated compute back-end), and visualization. The proposed platform is general, and can be easily adapted to other databases. The pre-packaged analyses toolkit is easily extensible, and allows for multi-language support. The platform can be easily federated, mirrored at other locations, and supports a RESTful API for service composition and scaling.

1. Introduction

The Precision Medicine Initiative was recently launched to develop a new model of patient-focused research, to "accelerate biomedical discoveries and provide clinicians with new tools, knowledge, and therapies to select which treatments will work best for which patients" [1]. Precision Medicine can help develop best practices and to enhance safety by integrating multiple lines of evidence. The Intensive Care Unit (ICU) represents a unique data source for supporting precision medicine. ICU records are comprised of clinical treatment data, large ECG and blood pressure monitoring data sets, and associated outcomes. ICU data is often invaluable in understanding cardio physiology and the impact of medicines on heart physiology, as measured by blood pressure and ECG. Additionally, ICU records may provide enough data to support expansion of accepted endpoints and deeper insights into drug safety. One of the challenges of cardiac safety research is that there are relatively rare events, the potential for drug-drug interactions, and these events may be dependent on the physiological status of the patient.

"Cardiac safety concerns are a leading reason why pharmaceutical companies withdraw drug applications prior to approval and why approved drugs are removed from the market" [2]. The FDA is critically aware that advanced analytics are essential here, "We hope to identify patterns that will help us predict which patients are at an increased risk for cardiovascular side effects. This knowledge can guide the development of safer treatments" [2]. Moreover, the FDA understands that clinically relevant results will require large data sets, "In addition, very small increases in the QT interval appear to carry risk, so studies that assess cardiac drug effects require collection of many thousands of ECGs" [2].

Recent developments in models and methods for data sciences such as deep learning, coupled with massively parallel computing platforms are enabling significant advances in applications such as image processing, natural language processing, and computational biology. Analysis of large volumes of ECG and blood pressure data support data driven clinical decision making. However, improvements in computational throughput have not translated into increases in clinical understanding. The amount of translational research effectively utilizes only a small fraction of the terabytes of data currently being collected in clinical settings [3]. Primary obstacles to more effective utilization include poor lines of communication between data scientists and clinicians on disease prognoses, technical difficulties due to the heterogeneity and complexity of physiological data, and lack of regulatory guidelines. This latter consideration also includes the absence of research with data driven systems to assess the risk and benefit of using such systems in clinical settings. The importance of data driven systems in clinical decision making is at the core of the Precision Medicine initiative: collection of clinical data in the form of electronic health

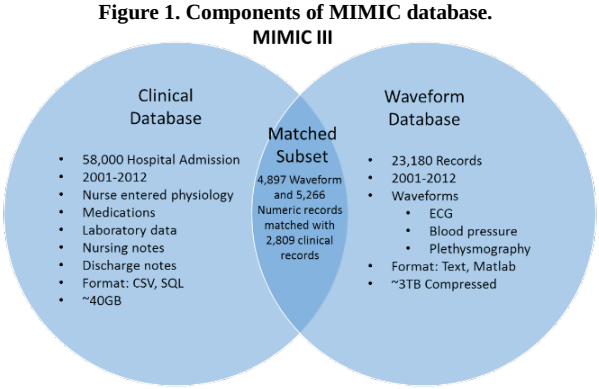
TABLE 1. RESEARCH PROJECTS WITH MIMIC DATABASE. 'C' FOR CLINICAL AND 'W' FOR WAVEFORM OR NUMERIC DATABASE.

Citation	Research Problem	Methods	Cohort Selection Criteria	Data	Cohort Size
[4]	Mortality Prediction with acute kidney injury (AKI)	Multivariable Regression	ICD9 = AKI and ICU stay ≥3 days	C	1,400
[5]	Local customized mortality prediction, outcome is survival to hospital discharge	Logistic Regression (LR), Bayesian Network (BN), Artificial Neural Network (ANN)	ICD9 = Acute Kidney Injury (AKI) AND/OR ICD9= subarachnoid hemorrhage (SAH)	C	1,400 for AKI, 223 for SAH
[6]	Whether 'similar' dynamical patterns can be identified across a heterogeneous patient cohort	Switching Vector Autoregressive framework (SVAR)	At least 8 hours of continuous minute by minute HR and BP trend within the first 24 hour of admission	C & W	450
[7]	Whether red cell distribution width (RDW) has the potential to improve prognostic performance	Multivariable Regression	All adult patients who had RDW measurements at admission	C	17,922
[8]	Investigate discriminatory pattern in hemodynamic data	Artificial Neural Network (ANN)	Defined by clinical event of HE from 'matched Subset'	C & W	1,311

records (EHRs) [9], including detailed local databases [10], or large administrative databases as collected by Medicare and Medicaid [11]. Despite these initiatives, there remains a major disconnect between data collection and translational research [12]. Limited advances in regulatory insights stem from the challenges of conducting research with data driven systems to assess the risk and benefit of using such systems in clinical settings. Ironically, the rate limiting steps in this process correspond to assembling the data into cohorts, something that is commonplace in data analytics, irrespective of the domain [13], and being able to analyze longer time windows of signal data over hundreds to thousands of patient days. What is needed is a new data model that allows different types of data to be stored, queried, and analyzed in the database. SciDB is a computational, array native database that allows data to be stored, queried and analyzed in the database. In this work, SciDB is central to the development of a new tool to support interactive exploration of the MIMIC III dataset. MIMIC III [14] is created and maintained by the Laboratory for Computational Physiology at the Massachusetts Institute of Technology. The database contains high resolution waveform data and clinical information on patients admitted to the Intensive Care Unit (ICU) since 2001 at the Beth Israel Deaconess Medical Center. The database has three main components: clinical, numeric, and waveform data. Numeric data contains minute by minute physiologic parameters extracted from waveform data, and usually every waveform record has a simultaneously recorded numeric record (e.g. the electrocardiogram waveform record has an associated heart rate numeric record). Clinical data is collected distinctly from the patient's EHR. A 'matched subset' links the waveform/numeric records to the corresponding clinical records: about 7% of the total data [Figure 1]. This large collection of heterogeneous data is from various sources and stored in a variety of formats. Consequently, it is difficult for clinical researchers to perform a cursory overview of the data and subsequently dig deeper into the data without a sound understanding of programming languages and database architectures, without access to powerful computational capabilities. The summary in Table 1 lists prominent research initiatives that use the MIMIC database. Each research initiative must deal with the following highly technical but repetitive manual tasks:

- High level browsing and exploration of the database
- Integration of heterogeneous data sources
- Cohort selection based on clinical criteria, and
- Use of different machine learning and statistical algorithms.

Figure 1. Components of MIMIC database.



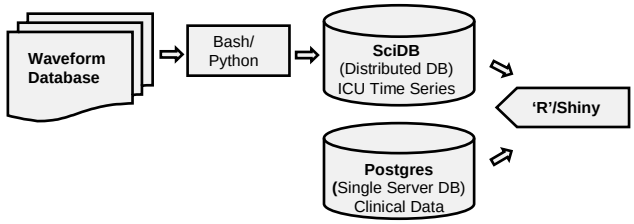
This, obviously, can be a limiting factor in performing translational research. To address this on a smaller scale, namely with one database (MIMIC III), we developed a new model that enables high level exploration and

browsing of the database, integration of heterogeneous data sources, automatic cohort selection with a minimal amount of programming.

2. Related work

Systems for integrating and exploring disparate data sets are the subject of significant prior efforts. PhysioNet hosts MIMIC [15] and has developed different software systems for accessing and visualizing the database on different platforms under the umbrella of the Physionet toolkit. The problem of data set selection according to required criteria and the analysis of disparate data are extensively investigated in computer science and database technologies [16]. Recent advances in sensor technologies and the ubiquity of handheld devices are further pushing researchers to address this problem. One research initiative in this area is the Intel Science and Technology (ISTC) Big Data Working Group (BigDaWG) [17]. This group is building the polystore architecture with the MIMIC database as a test bed for their system. While this initiative is promising, it sweeps aside the immediate needs of the clinical researchers interested in investigating large data sets such as MIMIC. IBM has also launched Watson Analytics which is not open source. Two other examples include the tranSMART and i2b2 [18] [19]. They are building open source software platform for clinical research that would support data exploration, complex analysis and convenient access. However, they currently does not support disparate database systems and distributed computing necessary for big data analysis on data sets such as MIMIC.

Figure 2. Data flow architecture of the system.

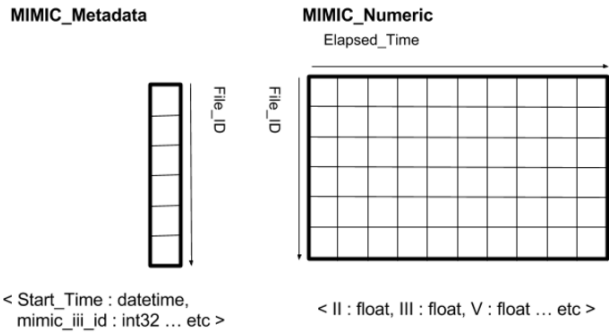


3. Our approach

Regenstrief Center for Healthcare Engineering (RCHE) at Purdue University works on population health management and prediction models for clinical intervention and data analysis for policy making in healthcare. In order to accelerate translational clinical research, the center has built a software tool for the MIMIC for,

- High level exploration and visualization of the database
- Clinical and Waveform database integration
- Open analysis with making the source code available
- Complex analysis across the database with distributed computing architecture

Figure 3. Waveform Database design in SciDB with two 2D Arrays.



The software architecture and data flow diagram is shown in Figure 2. Waveform Database (WFDB) toolbox in UNIX is used to convert the waveform data to CSV format and Bash and Python scripts are used to load the data in SciDB. R and Shiny is used for the integration, visualization and complex analysis. Figure 3 shows the database design using two 2d arrays in SciDB. For this work, only the matched subset with numeric data were used.

4. Use cases

The software architecture can be used for very simple as well as complex analyses based on the MIMIC database. Two example cases are described. The source code is available in the following links:

https://github.com/adibzaman/Shiny-MIMIC-UsecaseOne
https://github.com/adibzaman/Shiny-MIMIC-UsecaseTwo

4.1 Use case one

In June 2016, the FDA issued a safety announcement about serious heart problems with the antidiarrheal medicine Loperamide (Imodium) [20]. This warning was based on a sample of 48 patients. With a simple query of the MIMIC database, one can access 2,309 prescriptions of the drug Loperamide. Consequently, it would be of interest to know the demographic information as well as the status of the vital signs (including heart rate) for these patients. This query can be done from the following link without any programming:

http://mimic.catalyzecare.org:3838/sample-apps/usecaseone/

4.2 Use case two

Prediction of hypotensive episodes is of much interest in the clinical scenarios of an Intensive Care Unit [8] as timely detection for such events can be the difference between life and death. To develop a machine learning algorithm for the prediction of hypotensive episodes, it is necessary to define a hypotensive episode from blood pressure data (e.g., hypotension can be defined as 90

percent of the mean arterial blood pressure measurements during a 30 minute period being between 10 and 60 mmHg) and then search for those events in the waveform database, and build an algorithm by integrating necessary clinical information concurrent to these events from the clinical database. This can now be achieved by just one click of the following URL:

http://mimic.catalyzecare.org:3838/sample-apps/usecasetwo/

5. Conclusion

The combination of SciDB and PostgreSQL RDBMS software architectures can help to accelerate translational clinical research with MIMIC database. Integrated datasets will drive collaboration between data scientist, clinicians and software engineers. Using this approach, in the future, it should be possible to reproduce an entire research outcome from a single mouse click, allowing other researchers to test their own hypothesis by changing the parameters and cohort selection for their research and thus bring the open source movement to data driven translational clinical research [21]. This platform concept could change the way that we think about data intensive clinical research, the project life cycle and regulatory approval process, reducing the time from idea to translation and bringing together all the different stakeholders in a transparent data driven platform.

Acknowledgment

We express our gratitude to Drs. Leo Celi and Roger Mark from the Laboratory of Computational Physiology at the MIT for their support. We also wish to express our appreciation to the Regenstrief Foundation for their funding support of this work and Rosen Center for Advanced Computing (RCAC) at Purdue University for the technical support.

References

[1] WH. (2015). Available: https://www.whitehouse.gov/the-press-office/2015/01/30/fact-sheet-president-obama-s-precision-medicine-initiative

[2] FDA. (2012). Regulatory Science in Action. Available: www.fda.gov/downloads/drugs/drugsafety/ucm300948.pdf

[3] O. Badawi, T. Brennan, L. A. Celi, M. Feng, M. Ghassemi, A. Ippolito, et al., "Making big data useful for health care: a summary of the inaugural mit critical data conference," JMIR medical informatics, vol. 2, p. e22, 2014.

[4] L. Celi, R. Tang, M. Villarroel, G. Davidzon, W. Lester, and H. Chueh, "A clinical database-driven approach to decision support: Predicting mortality among patients with acute kidney injury," Journal of healthcare engineering, vol. 2, pp. 97-110, 2011.

[5] L. A. Celi, S. Galvin, G. Davidzon, J. Lee, D. Scott, and R. Mark, "A database-driven decision support system: customized mortality prediction," Journal of personalized medicine, vol. 2, pp. 138-148, 2012.

[6] L.-w. H. Lehman, R. P. Adams, L. Mayaud, G. B. Moody, A. Malhotra, R. G. Mark, et al., "A physiological time series dynamics-based approach to patient monitoring and outcome prediction," Biomedical and Health Informatics, IEEE Journal of, vol. 19, pp. 1068-1076, 2015.

[7] S. Hunziker, L. A. Celi, J. Lee, and M. D. Howell, "Red cell distribution width improves the simplified acute physiology score for risk prediction in unselected critically ill patients," Crit Care, vol. 16, p. R89, 2012.

[8] J. Lee and R. G. Mark, "An investigation of patterns in hemodynamic data indicative of impending hypotension in intensive care," Biomedical engineering online, vol. 9, p. 62, 2010.

[9] CMS. (March 14, 2016). Available: https://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/index.html

[10] M. Saeed, M. Villarroel, A. T. Reisner, G. Clifford, L.-W. Lehman, G. Moody, et al., "Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): a public-access intensive care unit database," Critical care medicine, vol. 39, p. 952, 2011.

[11] ResDAC. (March 14, 2016). Available: http://www.resdac.org/cms-data

[12] L. Anthony Celi, R. G. Mark, D. J. Stone, and R. A. Montgomery, ""Big data" in the intensive care unit. Closing the data loop," American journal of respiratory and critical care medicine, vol. 187, pp. 1157-1160, 2013.

[13] T. N. Y. Times. (2014). Available: http://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html?_r=1

[14] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, et al., "MIMIC-III, a freely accessible critical care database," Scientific data, vol. 3, 2016.

[15] PhysioNet. (2016). Available: www.physionet.org

[16] M. Stonebraker and U. Çetintemel, "" One size fits all": an idea whose time has come and gone," in Data Engineering, 2005. ICDE 2005. Proceedings. 21st International Conference on, 2005, pp. 2-11.

[17] A. Elmore, J. Duggan, M. Stonebraker, M. Balazinska, U. Cetintemel, V. Gadepally, et al., "A demonstration of the BigDAWG polystore system," Proceedings of the VLDB Endowment, vol. 8, pp. 1908-1911, 2015.

[18] TranSMART. (March 14, 2016). Available: http://transmartfoundation.org/

[19] S. N. Murphy and A. Wilcox, "Mission and sustainability of informatics for integrating biology and the bedside (i2b2)," eGEMs (Generating Evidence & Methods to improve patient outcomes), vol. 2, p. 7, 2014.

[20] FDA. (2016). FDA Safety Announcement. Available: http://www.fda.gov/Drugs/DrugSafety/ucm504617.htm

[21] M. Eisenstein, "Big data: The power of petabytes," Nature, vol. 527, pp. S2-S4, 2015.