Accurate and reproducible prediction of ICU readmissions

Dinh-Phong NGUYEN[1, 2,✉]

[1]WIND-DSI, AP-HP, Paris, France
[2]Sorbonne Université, UPMC Univ Paris 06, Paris, France

Readmission in the intensive care unit (ICU) is associated with poor clinical outcomes and high costs. Traditional scoring methods to help clinicians deciding whether a patient is ready for discharge have failed to meet expectations, paving the way for machine learning based approaches. Freely available datasets such as MIMIC-III have served as benchmarking media to compare such tools. We used the OMOP-CDM version of MIMIC-III (MIMIC-OMOP) to train and evaluate a lightweight tree boosting method to predict readmission in ICU at different time points after discharge (3, 7 and 30 days), outperforming existing solutions with an AUROC of 0.805 for 3-days readmission.

ICU readmission | MIMIC | OMOP-CDM | gradient boosting
Correspondence: ***dinh-phong.nguyen@aphp.fr***

Introduction

Recent studies have shown that readmission in the intensive care unit (ICU) is associated with poor clinical outcomes, increased length of ICU and hospital stay, and high costs (1, 2). One of the main reasons for ICU readmission that has been identified is premature discharge (3); in fact the transfer of patients from an ICU to a general hospital ward represents a high-risk event, and thus the decisions about which patients are ready to be discharged are daily struggles for ICU clinicians (4). Other studies have shown that determining the best timing for ICU discharge is usually based on subjective intuitions and that readmission prediction tools can help physicians in this endeavor, provided their performance and ease of adoption (5, 6). As traditional scores based on logistic regression or Cox proportional hazards models such as the Stability and Workload Index for Transfer score (SWIFT) or the LACE index have failed to meet expectations (6–10), numerous prediction models using machine learning have been proposed in the recent past, several of which trained and evaluated on the Multiparameter Intelligent Monitoring in Intensive Care (MIMIC-II or MIMIC-III) open database (11–15).

MIMIC-III is a large ICU EHR database widely accessible to researchers internationally under a data use agreement, allowing clinical studies to be reproduced and benchmarked (16, 17). In order to make multicenter ressearch possible, a valuable effort has been made to convert MIMIC-III to the Observational Medical Outcomes Partnership common data model (OMOP-CDM), which provides structural and conceptual models relying on international reference terminologies (18, 19). For the sake of reproducibility and ease of subsequent implementation in other centers using the same data model, we chose to use the OMOP-CDM version of MIMIC-III (MIMIC-OMOP), for which documentation and a mapping Extract-Transform-Load (ETL) process are freely available (20).

Related works

Among the numerous works aiming to provide decision-making tools for ICU clinicians at discharge time, two in particular caught our attention in terms of performance and similarity of setting to our own.

Lin et al. (12) proposed an advanced neural network for 30-day ICU readmission prediction (LSTM-CNN based model) achieving an Area Under Curve of the Receiver Operating Characteristic (AUROC) metric of 0.791 on MIMIC-III, using chart events 48h time series, diagnostic ICD-9 codes embeddings, and demographic information of the patients. The authors claim to offer higher sensitivity (0.742) compared to existing solutions, regardless of the specificity trade-off. With a fixed specificity at 0.850 and 0.800, they achieve a sensitivity of 0.548 and 0.619 respectively with their best model. There is no mention of precision nor F1-score.

Pakbin et al. (13) trained a simpler and more interpretable gradient boosting model (XGBoost) for predicting risk of ICU bounceback and readmission at a variety of time points using MIMIC-III, achieving AUROC of 0.76 and 0.75, F1-score of 0.20 and 0.34, for 72h and 30-days ICU readmission respectively. They use chart events time series, ICD-9 codes indicators, as well as admission, demographic and length-of-stay information of the patients. There is no mention of sensitivity, specificity or precision.

Methods

Data and patients. MIMIC-III integrates deidentified, comprehensive clinical data of patients admitted from 2001 to 2012 at the Beth Israel Deaconess Medical Center in Boston, Massachusetts. We restricted our analyses to ICU stays of patients over 18 years old, ending up with a dataset of 55135 stays. Variables used in the model were age, gender, length of stay, provenance from a surgery ward, current count of ICU visits for the same patient (=visit rank), and three values for a number of measures and blood tests: the first entry for a given stay, the last one and the absolute difference between the two. Those measures were total glasgow coma scale (GCS), motor GCS, verbal GCS, eye movement GCS, systolic blood pressure, heart rate, respiratory rate, body temperature, oxygen saturation, oxygen inspired fraction, body weight, urine output, serum bicarbonate, serum urea, total bilirubin, serum sodium, serum potassium, serum creatinine, blood platelets,

hemoglobinemia, blood hematocrit, blood leukocytes, serum lactates, blood PH, blood glucose and the International Normalized Ratio (INR). Three variables were extracted from medical and nurses text notes: history of AIDS, metastatic cancer and/or of advanced hematologic condition (myeloma, lymphoma or leukemia). Finally, one feature counting the number of available values for all the previous variables was added. Missing data were then imputed via multivariate imputation from all other available variables (21), with a Bayesian ridge regression as the estimator (22). A short summary of the dataset's main characteristics is reported in Table1.

		Overall
Total ICU visits, n		55135
Gender, n (%)	Male	31219 (56.6)
	Female	23916 (43.4)
Length of stay (days), mean (SD)		3.2 (4.9)
Age, n (%)	18-25	1346 (2.4)
	25-45	6271 (11.4)
	45-65	18887 (34.3)
	65-85	22800 (41.4)
	85-95	3226 (5.9)
	95+	2605 (4.7)
Visit rank, mean (SD)		0.2 (0.5)
Provenance, n (%)	Other wards	27391 (49.7)
	Surgery	27744 (50.3)
Personal history of AIDS, n (%)	No	53013 (96.2)
	Yes	2122 (3.8)
Personal history of metastatic cancer, n (%)	No	50076 (90.8)
	Yes	5059 (9.2)
Personal history of advanced hematologic condition, n (%)	No	52359 (95.0)
	Yes	2776 (5.0)

Table 1. Description of the dataset's main characteristics

Outcome definition. We used a similar outcome definition to Lin et al. (12), where positive cases were regarded as the visits where the patients could benefit from a prediction of readmission before being transferred or discharged: visits where the patients were either transferred or discharged but returned to ICU, or died before a defined time limit (3, 7, or 30 days).

Model training and evaluation. Several model families were tested in a screening phase, among which linear models, support vector machines, naive bayes, decision trees and ensemble methods. XGBoost, a gradient tree boosting method that is widely used to achieve state-of-the-art results on many machine learning problems, was consistently outperforming the others on all metrics and was thus selected as the prediction model for this study (23). Schematically, gradient boosting methods work by iteratively fitting "weak" models to the residuals of the previous model, and adding the newly estimated residuals to the previous model's prediction, thus forming a "new" prediction, and so on until a stopping criterion is met. XGBoost implements this algorithm with decision trees, an additional custom regularization term in the objective function, and several computing tweaks to optimize speed and performance, such as parallel learning or sparsity awareness.

Our dataset was divided into three parts, each representing 64%, 16% and 20% of the whole respectively: the first part to train our model (the *training set*), the second to tune the model's output classification threshold and to apply the *early stopping* method (the *validation set*) and the third part to evaluate the performance of our model on unseen data (the *test set*). The splits were stratified and grouped by individual patients, meaning each set contained about the same proportion of each outcome class and that all visits of a patient were grouped in the same set.

To reduce overfitting, we used the *early stopping* method. At each training epochs, the model's performance is evaluated on the validation set by measuring the negative log-likelihood: if it hasn't improved after a fixed number of epochs (in our case 10), then the training is stopped. As for the final evaluation, we decided to report a comprehensive set of metrics, for a full overview of the model's abilities and to ease future comparability with other approaches: area under the receiver operating characteristic curve (AUROC), precision (positive predicted value), specificity (true negative rate), sensitivity (true positive rate, also known as the recall) and F1-score, the harmonic mean between precision and recall. All metrics were calculated over the test set.

The probability threshold above which the output of our model would be classified as a positive outcome has been chosen according to an iterative procedure optimizing for the highest F-β score on the validation set, with $\beta = 1.5$. For reference, the F-β score is a weighted weighted harmonic mean between precision and recall, favoring recall when $\beta > 1$, and vice versa. Hyperparameter selection was done via a stratified group 5-fold cross-validation procedure in a grid-search setting, optimizing for the following subset of parameters: the number of trees, the maximum depth for each tree, the proportion of features used for each tree and the learning rate. Calibration of the models was assessed by separating the predicted probabilities over the test set into deciles, and assessing the proportion of realized outcomes in each bin. A model is said to be well calibrated when each bin's true outcome proportion is close to the bin's predicted probabilities, resulting in a calibration curve close to the diagonal line when plotted.

Apart from performance, we also wanted to understand which features were important to the model. Contributions of each feature were reported using the *TreeExplainer* from the SHAP python library, a state-of-the-art explanation framework for tree based-methods that enables the tractable com-

putation of optimal local (ie. per sample) explanations, as defined by desirable properties from game theory elements, such as Shapley values (24, 25). It is important to note that this might not necessarily mean that features with a higher importance are significantly associated with the outcome in a causal relationship, but it is nonetheless the best available way to assess the internal correlations learned by the model.

Reproducibility. All code used to produce this work, essentially written in Python, is available at http://github.com/deepphong/icu-readmission, accompanied by a step by step example Jupyter notebook and generic functions to exploit any database in OMOP format for other use cases.

Results

The evaluation results on the test set are reported in Table2 for each metric and outcome. Overall, at the classification threshold set for optimal F1.5-score, our model consistently equals or outperforms the results previously reported by Lin et al. (12) for 30-days readmission (AUROC 0.794 vs. 0.791, recall 0.796 vs. 0.742) and Pakbin et al. (13) for 3-days, 7-days and 30-days readmission (AUROC 0.805 vs. 0.76, 0.807 vs. 0.77 and 0.794 vs. 0.75 respectively; F1-score 0.481 vs. 0.22, 0.527 vs. 0.32 and 0.560 vs. 0.37 respectively). Performance for other classification thresholds are reported in Figure 1 for 3-days readmission prediction.

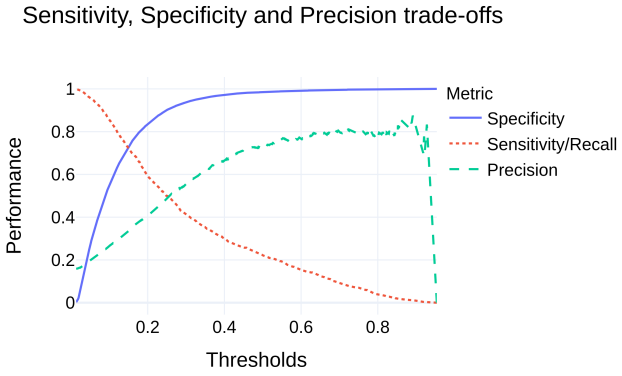

Fig. 1. Performance trade-offs for specificity, sensitivity and precision on the test set according to different classification thresholds, on 3-days readmission prediction

Calibration of the model was overall very good, as visually assessed in Figure 2, with a near perfect fit towards the extremes, meaning that the model is more frequently right the more confident it is in making its predictions. While calibration statistical tests exist, such as the Hosmer-Lemeshow test, we feel like a calibration plot gives more information and intuition.

Figure 3 shows the model's top 20 features importance for 3-days readmission prediction, sorted by the sum of Shapley value magnitudes over all samples. While the length-of-stay seems to be the most discriminative feature for model, with shorter duration associated with a higher chance of readmission, an interesting observation is the feature that comes just

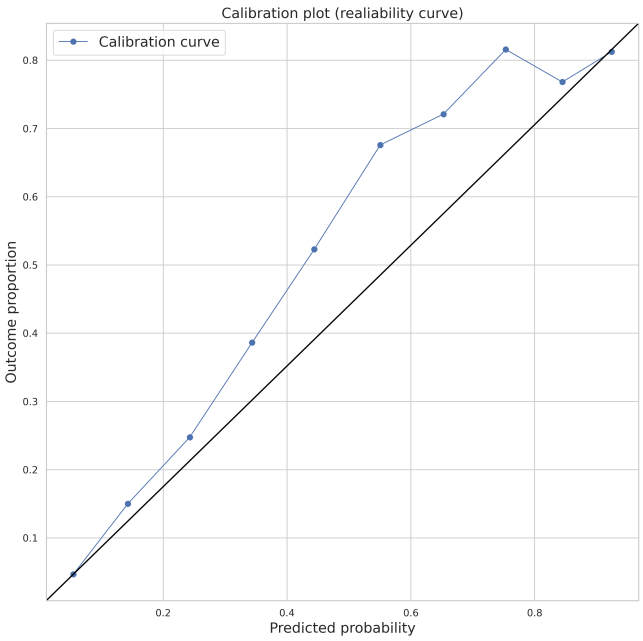

Fig. 2. Calibration plot showing the actual proportion of readmission in each decile of predicted probability of readmission for 3-days readmission predicsion

after: the number of non-missing features of an individual for a given stay. This feature can be assimilated to the number of different measurements and blood tests taken on an individual for a given stay, and is plausibly correlated with the severity of the patient's condition.

Performance trade-offs, calibration plots and features importance plots for the other evaluated outcomes are available in the appendix.

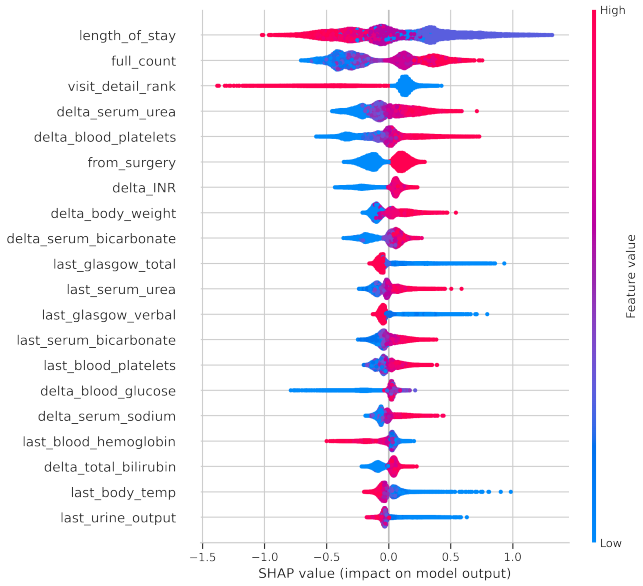

Fig. 3. *TreeExplainer* features importance (top 20) on the test set, ranked by the sum of Shapley value magnitudes over all samples for the model prediction. Each dot in the visualization represents one prediction. The color is related to the real data point; if the actual value in the dataset was high, the dot is colored in red; blue indicates the actual value is low. Features prefixed with "last_" are the last available measurements before discharge and features prefixed with "delta_" are the difference between the first and last available measurement of the stay

	3-days readmission (CIR=15.8%)	7-days readmission (CIR=19.7%)	30-days readmission (CIR=27.3%)
AUROC	0.805	0.807	0.794
Specificity	0.812	0.757	0.604
Recall/sensitivity	0.633	0.725	0.796
Precision	0.388	0.418	0.432
F1-score	0.481	0.527	0.560

Table 2. Performance metrics over the test set for the three different outcomes, at the classification threshold for optimal F1.5-score. CIR = class imbalance ratio (proportion of positive cases among the population)

Conclusion and discussion

In this work, we proposed a model based on a tree boosting method to predict ICU readmission at 3, 7 and 30 days using data of from the patients visit available at discharge on the freely available MIMIC-III database. Our solution is open-source and has the advantage of having been conceived with the OMOP-CDM standard, allowing for easier external validation and implementation. While this work improves on existing solutions for ICU readmission prediction, several points still need to be addressed.

The prediction model was trained and evaluated on MIMIC-OMOP, and as of this time, no external validation has been conducted yet. To facilitate this process, all code needed to reproduce the results has been open-sourced. Although efforts have been put into its ease-of-transfer on any other electronic health records (EHR) database using the OMOP-CDM standard, some non neglectible amount of work is always needed to adapt the code to new data. The feasibility of external validation is currently being evaluated on the largest French EHR database. Further work also needs to be done to integrate and evaluate such prediction models in clinical practice. The choice of developing on common health data standards such as OMOP-CDM and/or HL7's Fast Health Interoperability Resources (FHIR) (26) is a step forward in this direction, as more and more applications are being made compatible with those.

Although there are numerous similar studies claiming state-of-the-art performance for machine learning models on various tasks, most only report one or two metrics, mainly AUROC. We argue that it is only by reporting a fully comprehensive set of metrics that models can be made comparable and reproducible. The strengths and weaknesses of a model often rely on the performance trade-offs; depending on the use-case, one would want to favor one metric over the others (eg. sensitivity for non-invasive cancer screening), and would be able to assess the model's ability to do so with reports such as Figure 1.

Apart from our model's intrinsic performance, another interesting finding was the importance of including a variable accounting for the available measures among the ones selected for the model. This shows that the missing measures were missing not at random (MNAR), and the rationale behind this seems to be that patients with a poorer prognosis usually have more tests and measurements done to them. This feature could possibly be indirectly correlated with the care providers' overall feeling of the patient's current state, and we postulate that existing models could be improved by adding similar information.

Declarations

Ethics approval and consent to participate. Not applicable.

Consent for publication. Not applicable.

Competing interests. The authors declare that they have no competing interests.

Funding. No funding were perceived for this study.

Availability of data and material. All code used to produce this work, essentially written in Python, is available at http://github.com/deepphong/icu-readmission, accompanied by a step by step example Jupyter notebook and generic functions to exploit any database in OMOP format for other use cases.

Acknowledgements. This work has been made possible thanks to the support of the WIND-DSI department of AP-HP hospitals and the DIM department of Bicetre hospital, especially Christel DANIEL and Marie FRANK. Special thanks to Julien DUBIEL from WIND-DSI for his constant help with logistics.

Bibliography

1. Sydney Brown, Sarah Ratcliffe, Jeremy Kahn, and Scott Halpern. The Epidemiology of Intensive Care Unit Readmissions in the United States. *American Journal of Respiratory and Critical Care Medicine*, 185(9):955–964, May 2012. ISSN 1073-449X. doi: 10.1164/rccm.201109-1720OC.
2. Carolina R. Ponzoni, Thiago D. Corrêa, Roberto R. Filho, Ary Serpa Neto, Murillo S. C. Assunção, Andreia Pardini, and Guilherme P. P. Schettino. Readmission to the Intensive Care Unit: Incidence, Risk Factors, Resource Use, and Outcomes. A Retrospective Cohort Study. *Annals of the American Thoracic Society*, 14(8):1312–1319, May 2017. ISSN 2329-6933. doi: 10.1513/AnnalsATS.201611-851OC.
3. Uchenna R. Ofoma, Yue Dong, Ognjen Gajic, and Brian W. Pickering. A qualitative exploration of the discharge process and factors predisposing to readmissions to the intensive care unit. *BMC Health Services Research*, 18(1):6, January 2018. ISSN 1472-6963. doi: 10.1186/s12913-017-2821-z.
4. Daniel Niven, Jaime Bastos, and Henry Stelfox. Critical Care Transition Programs and the Risk of Readmission or Death After Discharge From an ICU: A Systematic Review and Meta-Analysis*. *Critical Care Medicine*, 42(1):179–187, January 2014. ISSN 0090-3493. doi: 10.1097/CCM.0b013e3182a272c0.
5. Claudia-Paula Heidegger, Miriam M. Treggiari, Jacques-André Romand, and and the Swiss ICU Network. A nationwide survey of intensive care unit discharge practices. *Intensive Care Medicine*, 31(12):1676–1682, Dec 2005. ISSN 1432-1238. doi: 10.1007/s00134-005-2831-x.
6. Uchenna R. Ofoma, Subhash Chandra, Rahul Kashyap, Vitaly Herasevich, Adil Ahmed, Ognjen Gajic, Brian W. Pickering, and Christopher J. Farmer. Findings from the implementation of a validated readmission predictive tool in the discharge workflow of a medical intensive care unit. *Annals of the American Thoracic Society*, 11(5):737–743, June 2014. ISSN 2325-6621. doi: 10.1513/AnnalsATS.201312-436OC.

7. Marc Kastrup, Robert Powollik, Felix Balzer, Susanne Röber, Robert Ahlborn, Vera von Dossow-Hanfstingl, Klaus D. Wernecke, and Claudia D. Spies. Predictive ability of the stability and workload index for transfer score to predict unplanned readmissions after ICU discharge. *Critical Care Medicine*, 41(7):1608–1615, July 2013. ISSN 1530-0293. doi: 10.1097/CCM.0b013e31828a217b.
8. Regis Goulart Rosa, Cintia Roehrig, Roselaine Pinheiro de Oliveira, Juçara Gasparetto Maccari, Ana Carolina Peçanha Antônio, Priscylla de Souza Castro, Felippe Leopoldo Dexheimer Neto, Patrícia de Campos Balzano, and Cassiano Teixeira. Comparison of Unplanned Intensive Care Unit Readmission Scores: A Prospective Cohort Study. *PloS One*, 10(11):e0143127, 2015. ISSN 1932-6203. doi: 10.1371/journal.pone.0143127.
9. Devan Kansagara, Honora Englander, Amanda Salanitro, David Kagen, Cecelia Theobald, Michele Freeman, and Sunil Kripalani. Risk Prediction Models for Hospital Readmission: A Systematic Review. *JAMA*, 306(15):1688–1698, October 2011. ISSN 0098-7484. doi: 10.1001/jama.2011.1515.
10. Carl van Walraven, Irfan A. Dhalla, Chaim Bell, Edward Etchells, Ian G. Stiell, Kelly Zarnke, Peter C. Austin, and Alan J. Forster. Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community. *CMAJ : Canadian Medical Association Journal*, 182(6):551–557, April 2010. ISSN 0820-3946. doi: 10.1503/cmaj.091117.
11. Ye Xue, Diego Klabjan, and Yuan Luo. Predicting ICU readmission using grouped physiological and medication trends. *Artificial Intelligence in Medicine*, 95:27–37, April 2019. ISSN 1873-2860. doi: 10.1016/j.artmed.2018.08.004.
12. Yu-Wei Lin, Yuqian Zhou, Faraz Faghri, Michael J. Shaw, and Roy H. Campbell. Analysis and prediction of unplanned intensive care unit readmission using recurrent neural networks with long short-term memory. *PLOS ONE*, 14(7):e0218942, July 2019. ISSN 1932-6203. doi: 10.1371/journal.pone.0218942.
13. A. Pakbin, P. Rafi, N. Hurley, W. Schulz, M. Harlan Krumholz, and J. Bobak Mortazavi. Prediction of ICU Readmissions Using Data at Patient Discharge. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 4932–4935, July 2018. doi: 10.1109/EMBC.2018.8513181.
14. J. Venugopalan, N. Chanani, K. Maher, and M. D. Wang. Combination of static and temporal data analysis to predict mortality and readmission in the intensive care. In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 2570–2573, July 2017. doi: 10.1109/EMBC.2017.8037382.
15. A. S. Fialho, F. Cismondi, S. M. Vieira, S. R. Reti, J. M. C. Sousa, and S. N. Finkelstein. Data mining using clinical physiology at discharge to predict ICU readmissions. *Expert Systems with Applications*, 39(18):13158–13165, December 2012. ISSN 0957-4174. doi: 10.1016/j.eswa.2012.05.086.
16. Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3, May 2016. ISSN 2052-4463. doi: 10.1038/sdata.2016.35.
17. Hrayr Harutyunyan, Hrant Khachatrian, David C. Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific Data*, 6(1):96, 2019. ISSN 2052-4463. doi: 10.1038/s41597-019-0103-9.
18. George Hripcsak, Jon D. Duke, Nigam H. Shah, Christian G. Reich, Vojtech Huser, Martijn J. Schuemie, Marc A. Suchard, Rae Woong Park, Ian Chi Kei Wong, Peter R. Rijnbeek, Johan van der Lei, Nicole Pratt, G. Niklas Norén, Yu-Chuan Li, Paul E. Stang, David Madigan, and Patrick B. Ryan. Observational health data sciences and informatics (ohdsi): Opportunities for observational researchers. *Studies in health technology and informatics*, 216:574–578, 2015. ISSN 1879-8365.
19. J. Marc Overhage, Patrick B. Ryan, Christian G. Reich, Abraham G. Hartzema, and Paul E. Stang. Validation of a common data model for active safety surveillance research. *Journal of the American Medical Informatics Association : JAMIA*, 19(1):54–60, 2012. ISSN 1527-974X. doi: 10.1136/amiajnl-2011-000376.
20. Nicolas Paris and Adrien Parrot. MIMIC in the OMOP Common Data Model. *medRxiv*, page 2020.08.14.20175141, August 2020. doi: 10.1101/2020.08.14.20175141. Publisher: Cold Spring Harbor Laboratory Press.
21. Stef van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software, Articles*, 45(3):1–67, 2011. ISSN 1548-7660. doi: 10.18637/jss.v045.i03.
22. J.L. Schafer. *Analysis of Incomplete Multivariate Data*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. CRC Press, 1997. ISBN 9781439821862.
23. Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. *CoRR*, abs/1603.02754, 2016.
24. Scott M Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017.
25. Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1):56–67, January 2020. ISSN 2522-5839. doi: 10.1038/s42256-019-0138-9. Number: 1 Publisher: Nature Publishing Group.
26. D. Bender and K. Sartipi. Hl7 fhir: An agile and restful approach to healthcare information exchange. In *Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems*, pages 326–331, June 2013. doi: 10.1109/CBMS.2013.6627810.