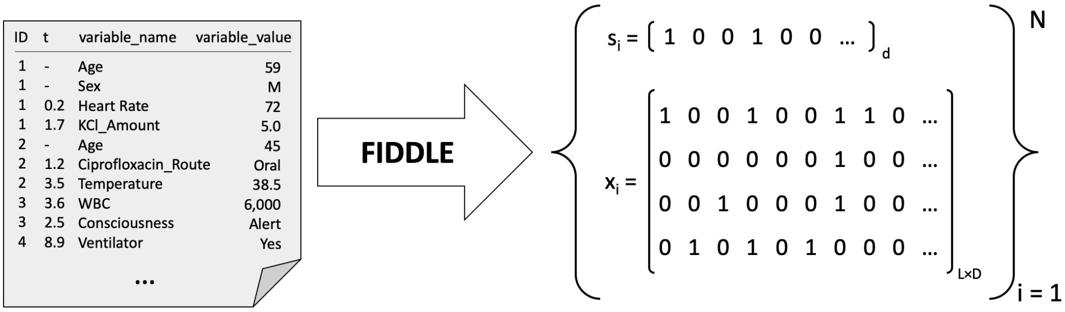
Democratizing EHR Analyses - A Comprehensive, Generalizable Pipeline for Learning from Clinical Data

Michael Sjoding MD MSc[1], Shengpu Tang[2], Parmida Davarmanesh[3], Yanmeng Song[4], Danai Koutra PhD[2], and Jenna Wiens PhD[2]

[1] Internal Medicine, UM, [2] Computer Science and Engineering, UM, [3] Math, UM, [4] Statistics, UM

Background. Widely shared electronic health record (EHR) datasets like MIMIC-III have given machine learning (ML) researchers increasing ability to work on important healthcare problems. In applying ML to EHR data, many decisions must be made *before* any ML is applied; due to the size and complexity of EHR data, such preprocessing requires substantial effort and is prone to error. As the role of ML in healthcare grows, there is an increasing need for systematic and reproducible preprocessing techniques for EHR data. We present "FIDDLE" – **F**lex**I**ble **D**ata-**D**riven pipe**L**in**E**, an open-source framework to speed up and standardize EHR data preprocessing.

Methods. FIDDLE is a pipeline that systematically transforms structured EHR data into feature vectors that can be used as input to ML algorithms (**Figure 1**). FIDDLE allows users to define the prediction window (e.g., 48hrs), the temporal granularity at which to consider the observations (e.g., hourly vs. daily), and the extent of filtering (e.g., features that occur in less than 0.1% of the population). Our proposed approach is largely data-driven, incorporates good practices from the literature, and relies on few assumptions. FIDDLE was designed to be generalizable, and it applies broadly to structured clinical data. In our experiments, we applied FIDDLE on all structured tables in MIMIC-III, including vital signs, laboratory results, medications, charted observations, fluid outputs, and microbiology. To assess the utility of generated features, we considered five prediction tasks for ICU patients involving three adverse outcomes after their ICU admission (predicting in-hospital mortality at 48h, predicting ARF at 4h and 12 h, and predicting shock at 4h and 12h). Pragmatic definitions of the outcomes were developed to enable identification using EHR data (ARF: positive pressure ventilation; shock: vasopressor administration). Study cohorts were split randomly into training, validation, and held-out test sets. On each prediction task, we trained models using the FIDDLE-generated feature set and four ML algorithms: (1) L2-regularized logistic regression, (2) random forest, (3) RNNs with LSTM cells, and (4) CNNs with 1D convolutions. We compared these models against several baselines including the national early warning score (NEWS) and the Harutyunyan et al. mortality benchmark. Performance was evaluated using the area under the receiver operating characteristics curve (AUROC) and precision-recall curve (AUPR).



ID	t	variable_name	variable_value
1	-	Age	59
1	-	Sex	M
1	0.2	Heart Rate	72
1	1.7	KCl_Amount	5.0
2	-	Age	45
2	1.2	Ciprofloxacin_Route	Oral
2	3.5	Temperature	38.5
3	3.6	WBC	6,000
3	2.5	Consciousness	Alert
4	8.9	Ventilator	Yes

...

$$\left\{ \begin{array}{l} s_i = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 0 & \ldots \end{bmatrix}_d \\ x_i = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & \ldots \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & \ldots \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & \ldots \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & \ldots \end{bmatrix}_{L\times D} \end{array} \right\}_{i=1}^{N}$$

Figure 1: The proposed preprocessing pipeline, FIDDLE, transforms structured EHR data into time-invariant and time-dependent feature vectors, s_i and x_i, that can be used as input to common ML algorithms.

Results. The five study cohorts varied in size from 8,577 to 19,342 examples, and the formatted input EHR data contained up to 46 million rows. Using default settings, FIDDLE extracted 4,143 to 7,508 features across these cohorts in approximately 30 to 150 minutes. On all tasks, FIDDLE-based models achieved good discriminative performance on the held-out test set with AUROC of 0.733-0.886. Specifically, for mortality prediction (test N=1,264, 10.4% positive), the FIDDLE-based LSTM performed significantly better than the Harutyunyan et al. benchmark (AUROC of 0.868 vs. 0.839, *p*-value < 0.001). On the ARF and shock prediction tasks, all FIDDLE-based models outperformed NEWS (*p*-value < 0.05), while no two ML models were significantly different (*p*-values ≥ 0.05).

Conclusion. FIDDLE, an open-source preprocessing pipeline, facilitates applying ML to structured EHR data and presents researchers with a quick and reasonable starting point from which they can build. FIDDLE is available online at https://gitlab.eecs.umich.edu/MLD3/FIDDLE. By accelerating and standardizing labor-intensive preprocessing, FIDDLE can help stimulate progress in building clinically useful ML tools for EHR data.