EHR DATA ANALYSIS USING MACHINE LEARNING MODELS

Cao Cong Luan Tran, Sarwan Ali, Murray Patterson
Georgia State University, Atlanta, GA 30303

MOLECULAR BASIS OF DISEASE

Background

Researchers are using EHR data and machine learning techniques to predict adverse outcomes in healthcare. However, preprocessing EHR data for ML tasks is complex due to its messy nature, high dimensionality, irregular sampling, and missing values. Open-source frameworks have been developed to simplify data preparation for hierarchically multimodal EHR data. These frameworks are generalizable across prediction times, ML algorithms, and data sets, and robust to user-defined arguments settings. They are effective in low-resource clinical settings but have limitations, such as computational inefficiency.

Purpose

The paper proposes a novel pretraining strategy for machine learning models in electronic health records (EHRs) to improve performance and interpretability. The strategy incorporates hierarchical relationships within EHR data, capturing the complexity and interdependencies within the data. This approach can speed up ML research using EHR data and provide a rapid starting point for researchers.

Methodology

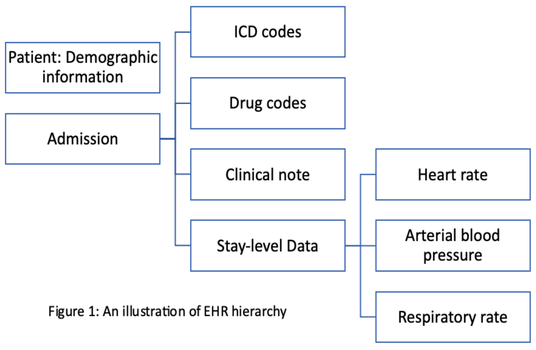

Figure 1: An illustration of EHR hierarchy

Input	`Data columns (total 4 columns):`
	`# Column Dtype`
	`0 ID int64`
	`1 t float64`
	`2 variable_name object`
	`3 variable_value object`
	`dtypes: float64(1), int64(1), object(2)`

Pre-filter: Rows with timestamps outside observation period are eliminated, and rarely occurring variables are removed, speeding up downstream analyses but potentially causing loss of useful information.	ID t variable_name variable_value 1 - Age 59 1 - Sex M 1 0.2 Heart Rate 72 1 1.7 Abx Amount 500 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 ...
Transform: Depending on the types of timestamps, processing continues. Data corresponding to a variable name are processed as "time-invariant" if the timestamp t for that variable name is null; otherwise, the data are processed as "time-dependent" in order to capture longitudinal trends and dynamics.	Time-invariant N $\tilde{\mathbf{s}}$ d Time-dependent N L $\tilde{\boldsymbol{x}}$ D
Post-filter: The data is transformed by removing redundant features, combining duplicated features into a single feature, and converting them into a matrix $S \in R^{N*d}$ and a tensor $X \in R^{N*L*D}$.	$\mathbf{s}$ N d $\boldsymbol{x}$ N L D
Output	$s_i \in \mathbb{R}^d$ for $i = 1 \ldots N$ $x_i \in \mathbb{R}^{L \times d}$

Table 1: Processing steps

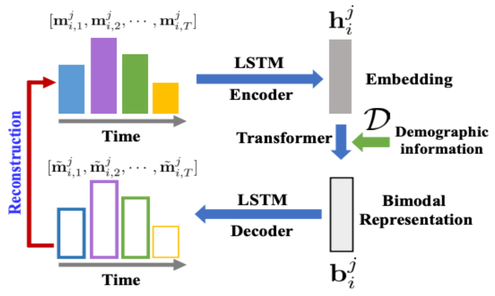

Figure 2: Stay-level self-supervised pretraining.

The study uses a bottom-to-up methodology for self-supervised learning, with two pretraining approaches for intra-modality relations and modality-level contrastive learning for inter-modality relations. A two-stage training technique is used from stay to admission levels, using the MIMIC-III dataset to forecast patient likelihood of acute respiratory failure, shock, or fatality, estimate readmission rate, and perform health risk prediction tasks.

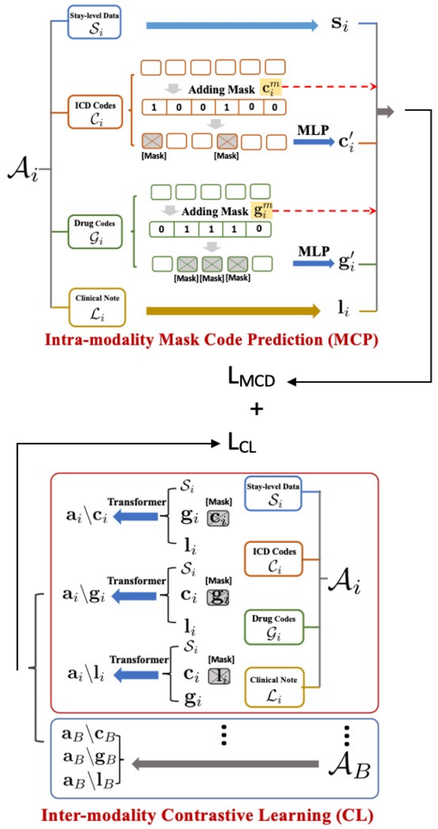

Figure 3: Admission-level self-supervised pretraining

Results

Task	In-hospital mortality, 48h		ARF,4h n=2358		ARF,12h n=2093		Shock, 4h n=2837		Shock, 12h n=2612	
Method	AUROC	AUPR	AUROC	AUPR	AUROC	AUPR	AUROC	AUPR	AUROC	AUPR
LR	0.856	0.444	0.817	0.657	0.757	0.291	0.825	0.548	0.792	0.274
RF	0.857	0.357	0.817	0.658	0.758	0.292	0.826	0.549	0.793	0.275
CNN	0.858	0.531	0.827	0.659	0.759	0.293	0.827	0.55	0.794	0.276
LSTM	0.859	0.51	0.827	0.66	0.76	0.294	0.828	0.551	0.795	0.277

Table 2: Summary of performance on MIMIC-III

Conclusion

The framework aids machine learning researchers in preprocessing EHR data, promoting faster and uniform preprocessing stages for therapeutically beneficial ML tools. However, it does not eliminate model checking. Experiments show high-dimensional feature vectors can leverage EHR structure, potentially exploiting hospital variables. Researchers may adjust filtering thresholds or use downstream feature selection approaches.

Future Work

- Investigate the impact of various hierarchical structures on the performance of pretrained models.
- Identify and address any limitations in the pipeline's ability to generalize across diverse datasets.
- Evaluate the pipeline's flexibility and effectiveness in handling different types of clinical predictions.

Acknowledgements

[1] Shengpu Tang, Parmida Davarmanesh, Yanmeng Song, Danai Koutra, Michael W Sjoding, and Jenna Wiens. Democratizing EHR analyses with FIDDLE: a flexible data-driven preprocessing pipeline for structured clinical data. Journal of the American Medical Informatics Association, 27(12):1921–1934, 10 2020. ISSN 1527-974X. doi: 10.1093/jamia/ocaa139. URL https://doi.org/10.1093/jamia/ocaa139.

[2] Xiaochen Wang, Junyu Luo, Jiaqi Wang, Ziyi Yin, Suhan Cui, Yuan Zhong, Yaqing Wang, and Fenglong Ma. Hierarchical pretraining on multimodal electronic health records. pages 2839–2852, December 2023. doi: 10.18653/v1/2023.emnlp-main.171.URL https://aclanthology.org/2023. emnlp-main.171.

We thank Molecular Biology of Disease Fellowship for the support.