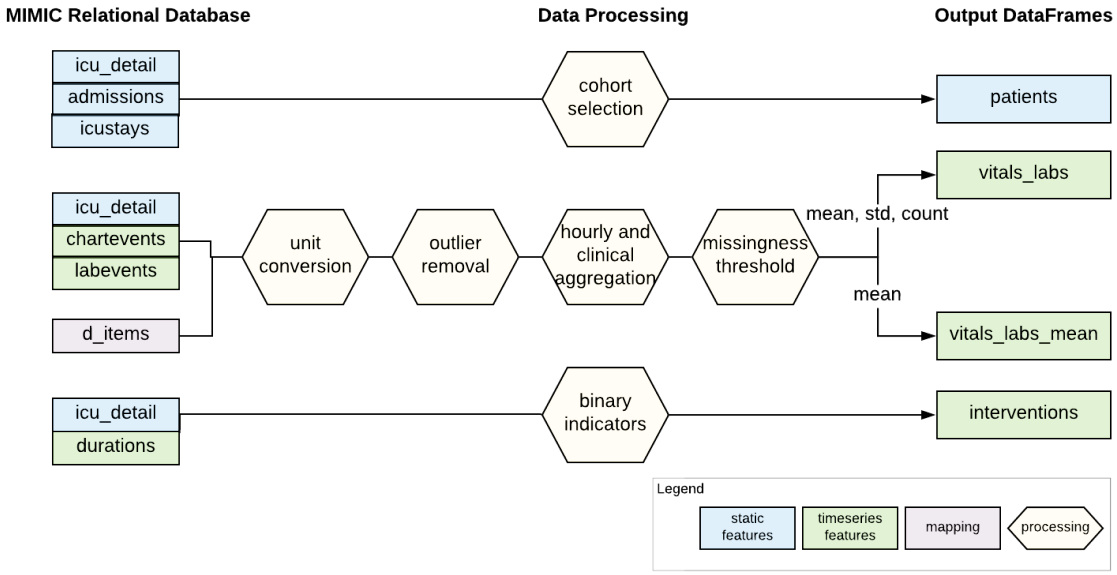
MIMIC-Extract: A Data Extraction, Preprocessing, and Representation Pipeline for MIMIC-III

Shirly Wang,[1] Matthew B. A. McDermott,[2] Geeticka Chauhan,[2] Michael C. Hughes,[3] Tristan Naumann,[4] and Marzyeh Ghassemi[1,5]
[1]University of Toronto [2]Massachusetts Institute of Technology [3]Tufts University [4]Microsoft Research [5]Vector Institute

Background.
Open-source supervised machine learning benchmarks help ensure reproducibility and encourage research efforts across many institutions to build toward common capabilities. For the application of machine learning to critical care tasks, the MIMIC-III database has been widely used but mostly via siloed bespoke pipelines not designed with reproducibility in mind. Though a few welcome benchmark efforts exist, they are limited to a small set of prediction tasks and may be vulnerable to label leakage by not enforcing a temporal gap between the last measurement time and the target event.

Methods.
MIMIC-Extract is an open-source pipeline for transforming raw MIMIC-III electronic health record (EHR) data into dataframes directly usable in common supervised machine learning workflows. We demonstrate utility on a range of tasks – per-admission mortality and length-of-stay prediction, and per-hour predictions of need for interventions with vasopressors and mechanical ventilation – all of which avoid label leakage.



Conclusion.
MIMIC-Extract addresses three primary challenges. First, it provides standardized data processing, including unit conversion, outlier detection, and aggregation of semantically equivalent features, thus reducing redundancy and missingness. Second, it provides benchmarks for clinically actionable prediction tasks given a time-series of patient history. Finally, the pipeline is extensible to different cohorts, covariates and prediction tasks.

Code and Full Paper.
Available at https://github.com/MLforHealth/MIMIC_Extract.