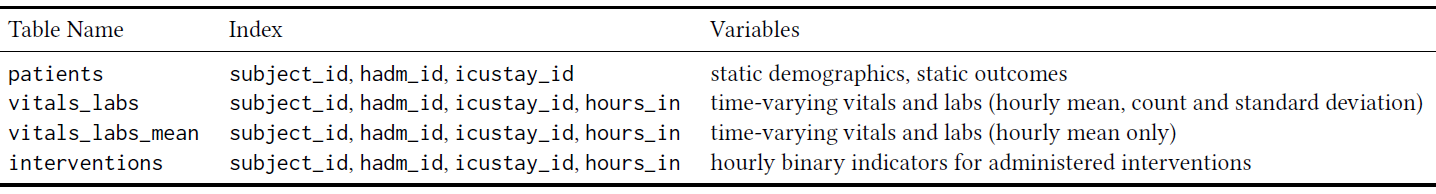
**MIMIC-Extract: A Data Extraction, Preprocessing, and Representation Pipeline for MIMIC-III**

Transforms MIMIC-III database into data structures directly usable in common time-series prediction pipelines.   
  
First, it transforms raw vital sign and laboratory measurements into usable hourly time series, performing essential steps such as unit conversion, outlier handling, and aggregation of semantically similar features to reduce missingness and improve robustness.   
Second, it extracts and makes prediction of clinically-relevant targets possible.   
Finally, the pipeline emphasizes reproducibility and is extensible to enable future research questions.  
  
Compared to the literature, they advance in 3 main contributions:   
1-Robust Representation: Converting to standardized units, detect and correct outliers, reduce missingness, and aggregate semantically similar raw features.   
2-Clinical meaningful interventions and outcomes.  
3-Focus on usability, reproducibility, and extensibility.   
   
First, a cohort is created that meets the selection criteria.   
Next, labs and vitals for the selected cohort are extracted and stored.  
Only labs and vitals that are missing less frequently (than a pre-defined threshold) are extracted and outlier values are filtered.



**Data Pipeline:**  
1-Cohort Selection: Criteria: Age>=15. First visit only, Hours of stay>=12, Days of stay<10.  
2-Variable Selection:  
- Static\_variables: 10 var’s, includes Age  
- Time\_varying\_vitals\_and\_labs: By default, it extracts 104 clinically aggregated time-series variables related to vital signs and laboratory test results. The prescripted drugs are excluded due to the unclear quality of the prescription signals. This occurs due to the lack of information about the prescriptions a patient actually took which may differ from those that were ordered for him.   
3-Unit Conversion and Outlier Detection: Use standardized units over all measurements, filter outliers over upper and lower bounds, and approximate non-outliers, yet physiologically invalid values to the nearest valid values.  
4-Hourly Aggregation: to organize less-frequent lab tests that are done every few hours (resulting into a sparse timeline compared to other signals).   
5-Semantic Grouping of Raw Features: group semantically equivalent ItemIDs together into more robust “clinical aggregate” features. Aggregate representations reduce data missingness and the presence of duplicate measures.  
6-Time-Varying Treatment Labels: Binary indicators of treatments (mechanical or liquids) on hourly basis.

**Extendibility of Data Pipeline:**Keywords:   
Functions in MIMIC-Extract use keywords to control admission cohort and time-varying features selection. Overwriting default values for the following keywords allows researchers to modify default extraction such as : min\_age, min\_duration&max\_duration, group\_by\_level2 (raw Vs clinically aggregated data), min\_percent (existing vs missing).  
Configurable Resource Files:  
The extraction code relies on information in associated resource files [that you can edit] for variable grouping and extraction (itemid\_to\_variable\_map.csv) and outlier correction (variable\_ranges.csv).  
SQL Queries:   
Modify the code or add SQL queries in the extraction code to include additional static variables, vitals and labs measurements and treatment labels in the output tables.  
Additional Dataframes:  
Extend the pipeline to output additional groups of variables (such as prescriptions) which exist in MUMIC-III but are not extracted by default here.