DIMENSION REDUCTION TECHNIQUES FOR CLINICAL PREDICTION MODELS ON HEALTH CARE DATA

Research aim: What is the impact of applying different types of dimension reduction (DR) techniques on the predictive performance, generalizability, and interpretability of clinical prediction models?

Problem formalization

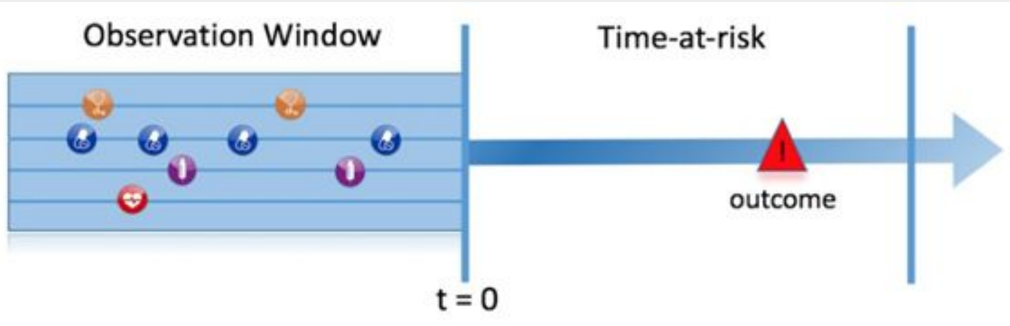
- The Electronic Health Record(EHR) data is **complex** and contains many medical codes for each diagnosis, procedure, and medication of a patient.

- The **sparsity** and **high dimensionality** of these EHRs present significant challenges in implementing widely applicable clinical prediction models.

- Current existing methods make use of techniques like **feature selection** which leads to a **loss of information**.

Many clinical events exist (high dimensionality), but patients typically only experience a few (sparsity)

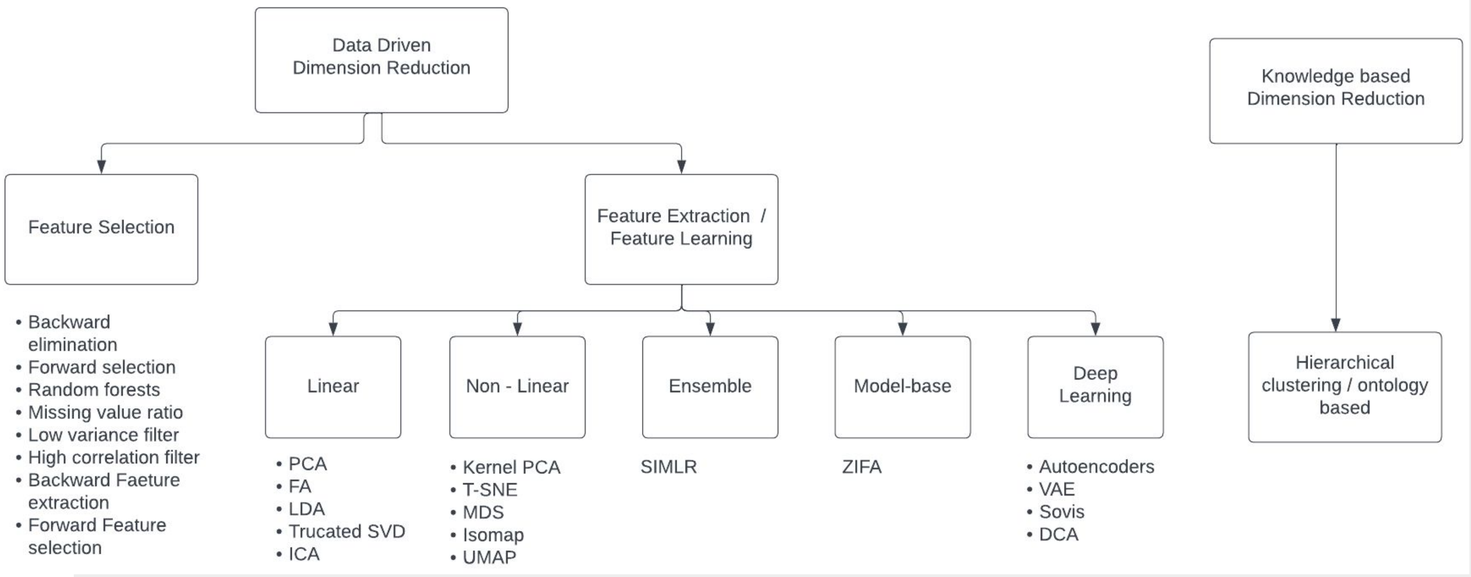
Patient Level Prediction framework



Prediction problem:
Among a population at risk, we aim to predict which patients at a defined moment in time (t=0) will experience some outcome during a time-at-risk. Prediction is done using only information about the patients in an observation window prior to that moment in time.

Current prediction question: Among hospital discharges of adult(18+) patients, who will go on to have hospital admission with 2 to 30 days?
Observation window: 365 days

Data Driven versus Knowledge Based

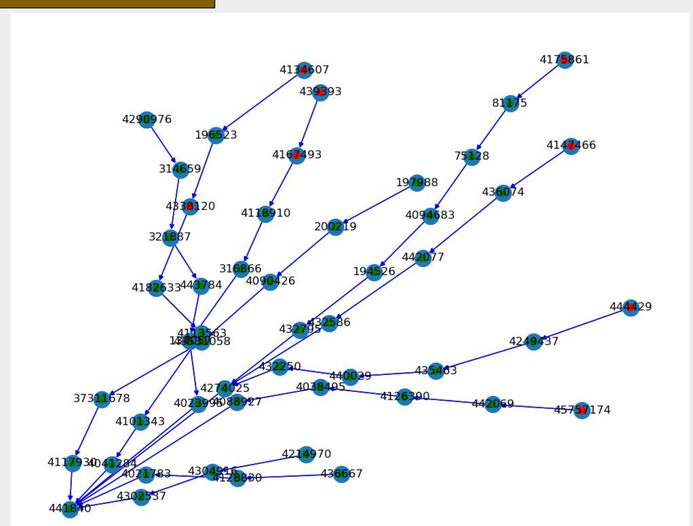


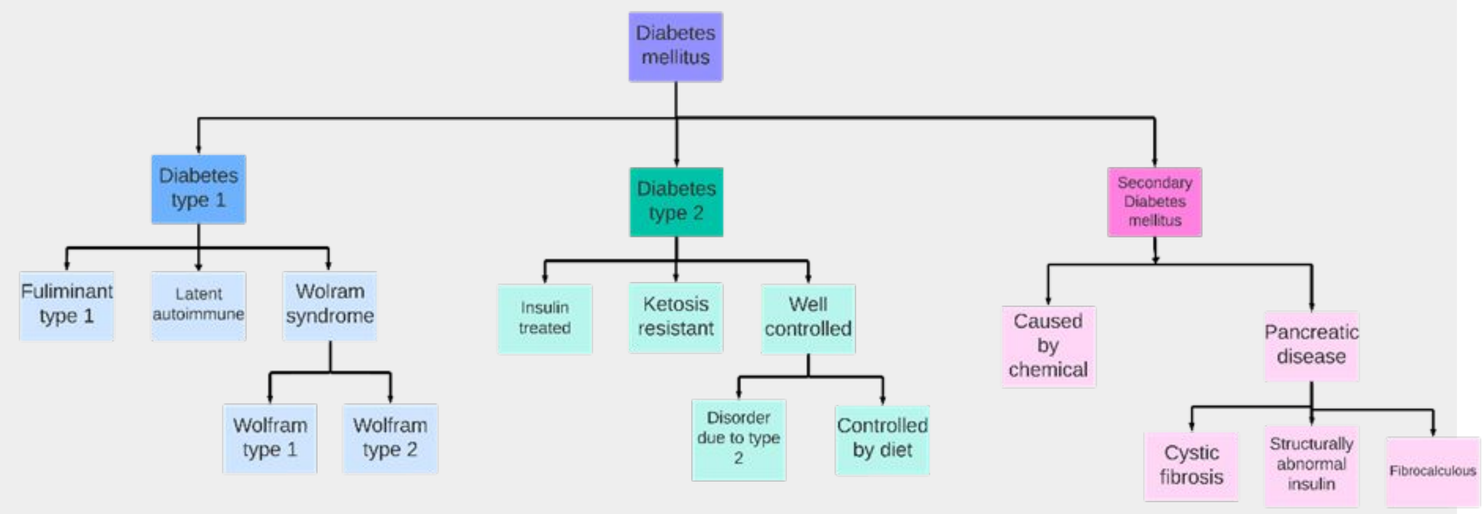
Data Driven Dimension Reduction:
A data driven algorithm learns the data structure and the characteristics of the data components, denoises the data, increases the separation between the components and projects the data onto a lower number of dimensions.

Knowledge based Dimension Reduction:
Knowledge based makes use of the hierarchical structure of the OMOP standarized vocabulary to reduce the number of dimension. The hierarchical structure can be verified by an domain expert.

Methods

Example of Hierarchical structure created from data





Example of structure used for Knowledge based Dimension Reduction

Table 4.2: Knowledge based approach (Levels based)

Input	Number of Candidate features	Model AUC train	Model AUC test	Number of Model features
L = 1	4	52.60	63.00	3
L = 2	33	51.12	55.33	4
L = 3	1187	77.22	73.97	68
L = 4	2008	76.85	76.21	85
L = 5	2206	76.03	69.31	112
L = 6	1403	69.86	58.27	77

Table 4.3: Knowledge based versus Data driven approach

Input	Number of Candidate features	Model AUC train	Model AUC test	Number of Model features
Condition Occurence	2795	77.51	75.37	85
Condition Group	2633	78,67	75.38	121
PCA	100	49.29	63.39	1
SVD	100	74.18	52.86	1
Auto Encoder	7	52.84	51.00	2

Health Data Science

Roëlle Bänffer, Aniek Markus, Tom Seinen

Erasmus MC University Medical Center Rotterdam