### Context

**Chronic kidney disease (CKD)**
- Renal function decrease: decrease of glomerular filtration rate (GFR)
- Often detected quite late & irreversible complications
- 1.7 to 2.5 million patients
- Heavy and invasive treatment (diagnosis, graft)

**Objective**
Identification of the metabolomic profiles predictive of CKD stages through an exhaustive local data mining approach with rule generation.

### Materials and Methods

**Dataset**: 10 subjects
- Features: urine $^1$H NMR spectra, Target: CKD stages, 2 modalities
- Features: signal acquisition and pre-processing

**HyperCube**
- High dimension 32X points
- Complex mixture of metabolites
- Overlaps of some peaks
- eGFR ≥ 60 ml/min/1.73 m²
  - Target: CKD Stages
  - low to mild CKD
  - 24 subjects
- eGFR < 60 ml/min/1.73 m²
  - moderate to established CKD
  - 86 subjects

**Supervised analysis**
1. **Dimension reduction**
   - Equidistant binning of 0.04 ppm (AUC)
   - Discretization of the features with a 10 bin quantization

2. **Feature selection**
   - Data mining
     - Pearson's chi-squared test ($\chi^2$)
     - Statistical test of goodness of fit and independence of the observed distribution of 2 discrete variables with respect to a theoretical distribution.
     - $\chi^2 = \sum \frac{(O - E)^2}{E}$
     - $O$ and $E$ are respectively the observed and expected frequencies.
     - $r$ and $c$ are resp. nb of rows and columns in the contingency table.

   - Supervised Mutual Information (nMI)
     - Measures of all types of features' mutual dependence.
     - $MI = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$

   - Rule mining
     - 10 rules: (Feature condition)
     - (Target modality)
     - (Range of values)

### Results

**Feature selection**

**Data mining**
- Supervised feature selection.
- Features (blue points) according to:
  - $\chi^2$ test: $-\log(p$-value)$) \geq X^2$
  - $nMI \geq \min[nMI(x,y)] = 0.14$
- Number of features selected = 36

**Rule mining**
- Relevant generated 10 rules were based on 10 out of the 36 previously selected features.

**Predictive model**
- Classification score: 0.79
  - (p-value: 0.001)
- The 10 features selected with our method (except the bucket 2.48) were in the first 16 features obtained with the Orthogonal Partial Least Square (O-PLS) model [2].
- Our method provided information on the distribution of the target modalities with respect to range of spectral values (buckets' AUC). 
- Metabolomic profiles of the CKD stages: 
  - GFR$^x_1$ => [citrate]$^x_1$ (metabolic disorders) ; 
  - dimethyl sulfone$^x_2$ (clearance) ; 
  - trigonelline$^x_3$ (protection against oxidative stress and apoptosis) ...

### Conclusion

- Our local data mining approach with rule generation combined with logistic regression supports the discriminant metabolites obtained in a previous study [2].
- It also provides information on the distribution of the CKD stages with respect to range of spectral values.

### Perspectives

- Include multi-source dataset
- Increase the number of subjects
- Refine the GFR classes
- Make a predictive model on subgroups defined by rules (i.e. local predictive model)

### References