Clinical score building from high-throughput proteomic data

Guillemette Marot, Wilfried Heyse, Vincent Vandewalle, Christophe Bauters, Florence Pinet

Clinical score building from high-throughput proteomic data

Guillemette Marot
Univ. Lille, CHU Lille METRICS & Inria MODAL

8 June, 2022

Acknowledgements : W. Heyse, V. Vandewalle, F. Pinet, C. Bauters

Overview

Precision medicine - Score building

Precision medicine :

Identifying which treatments will be effective for which patients based on genetic, environmental, and lifestyle factors.

Related questions :

- clustering in order to separate patients into different groups
- selection of clinical variables or biomarkers to predict the group

Large datasets are necessary to deal with the complexities of diseases.

Talk example : $\sim$ 5200 proteins (variables) measured simultaneously by Somascan technology, 2 cohorts with $\sim$ 240 patients

Score building

Score : tool for supporting decision-making for patient stratification (severity score, risk score), for example :

- linear combination from a discriminant analysis or a generalized linear model
- probability to belong to a class

Also called prediction clinical rule

The score, after choice of a threshold, will enable to assign a patient to a group in order to adapt his treatment.

Long-standing tradition in clinics using clinical variables to help non expert people and personalize treatments.

Many scores are unused (too many alerts for clinicians, unrealistic hypotheses, not understandable, too expensive, . . .)

Score building

The added value of omic data (here proteomic) must be clearly highlighted to convince clinicians to use a score mixing clinical and omic data.

Omic data : many variables in exploratory phasis
$\Rightarrow$ necessity to select some variables ("markers") for a use in clinics.

Selected variables must be biologically interpretable, ideally very few and easy to measure.

Variable selection is as important as prediction $\Rightarrow$ eliminates many machine learning techniques, favouring the ones assuming sparsity.

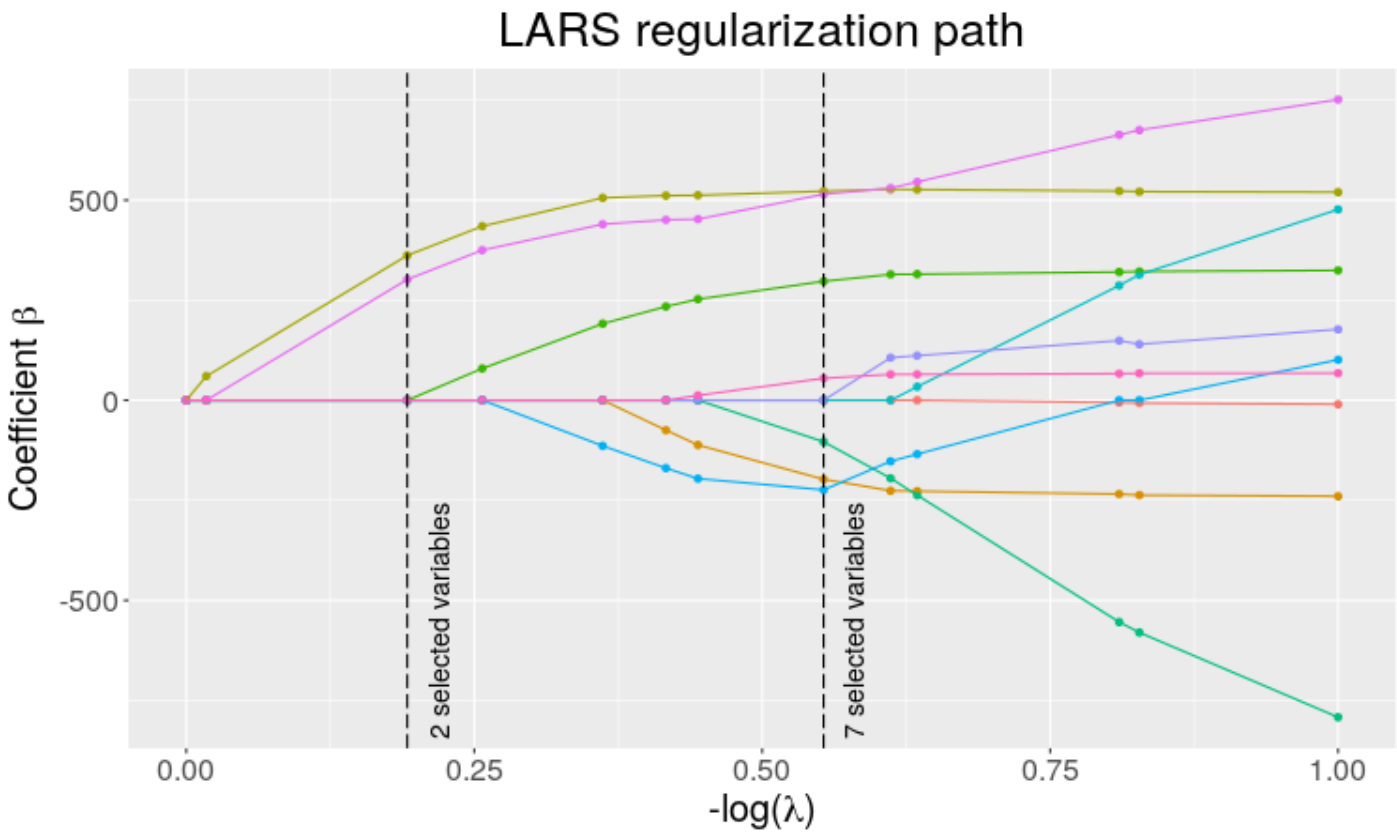
Example : Penalized regressions framework, e.g. Lasso (Tibshirani, 1996)

Score building

Lasso example in the high dimension context :

$$\widehat{\beta}(\lambda) = \underset{\beta \in \mathbb{R}^p}{\arg\min}\{C(\beta) + \lambda \text{pen}(\beta)\}$$

with $\text{pen}(\beta) = \|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$



Overview

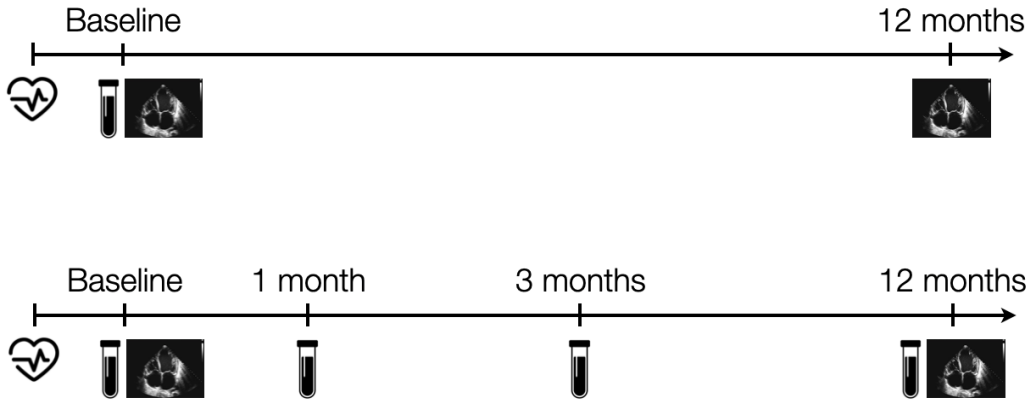
Example data : REVE-1 and REVE-2

Two cohorts designed by Pr Bauters (F Pinet's team, RID-AGE U1167) :

- Patients included after a **1st myocardial infarction**

  REVE-1 : (2002-2004)
  255 patients
  1 plasma sample

  REVE-2 : (2006-2008)
  238 patients
  4 plasma samples



- **Longterm follow-up** for heart failure and death for cardiovascular reasons

Research of proteins of interest among 5284 proteins to predict :

- LVR (left ventricular remodeling)
- heart failure

Overview

Penalized regressions

$y \in \mathbb{R}^n$: response variable (e.g. LVR, time before heart failure)
$X \in \mathcal{M}_{n,p}(\mathbb{R})$: measures of p proteins on n patients

High dimension context :

$$\widehat{\beta}(\lambda) = \underset{\beta \in \mathbb{R}^p}{\arg\min}\{C(\beta) + \lambda \text{pen}(\beta)\}$$

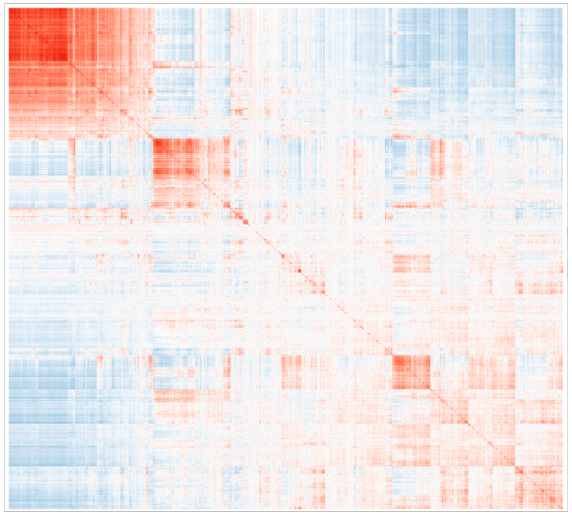
with for example

$$C(\beta) = \begin{cases} \|y - X\beta\|_2^2 = \sum_{i=1}^n (y_i - \sum_{j=1}^p x_{ij}\beta_j)^2 & \text{(linear reg.)} \\ -\sum_{i=1}^n (y_i \ln p_i + (1-y_i)\ln(1-p_i)) & \text{with } p_i = \frac{e^{x_i^T\beta}}{1+e^{x_i^T\beta}} \\ & \text{(logistic reg.)} \\ \sum_{i=1}^n \delta_i \left(\ln \left(\sum_{r \in \mathcal{R}} e^{\sum_j x_{rj}\beta_j} \right) - \sum_j x_{ij}\beta_j \right) & \\ \text{with } \delta_i \text{ indicative of event} & \\ \text{and } \mathcal{R} \text{ indices of individuals at risk} & \text{(Cox model)} \end{cases}$$

Penalized regressions

Problems when using Lasso-based regressions :

- Difficult choice of λ :
  cross-validation led us to over-fitting by selecting too many variables.
- Highly correlated data : correlation structures due to high dimension and to biological relations.



Solutions tested (PhD W. Heyse) :

- Stability selection (Meinshausen and Bühlmann, 2010)
- Use of variations of lasso : group-Lasso (Yuan et Lin, 2006), adaptive Lasso (Zou, 2006), multi-layer group-Lasso (Grimonprez et al., 2022).

Penalized regressions vs univariate competing risks analyses

Negative results when studying LVR : no validation on REVE-2, difficulties to interpret selected variables.

Possible explanations :

- LVR is a too complex mechanism to be predicted only with circulating biomarkers
- Not enough patients to build a model with enough proteins

$\Rightarrow$ Change of endpoint (study of heart failure or death) and change of statistical strategy (use of univariate analyses and correction for multiple testing)

The use of Fine-Gray competing risks model led to the selection of 50 proteins significantly associated to the occurence of heart failure on REVE-1.

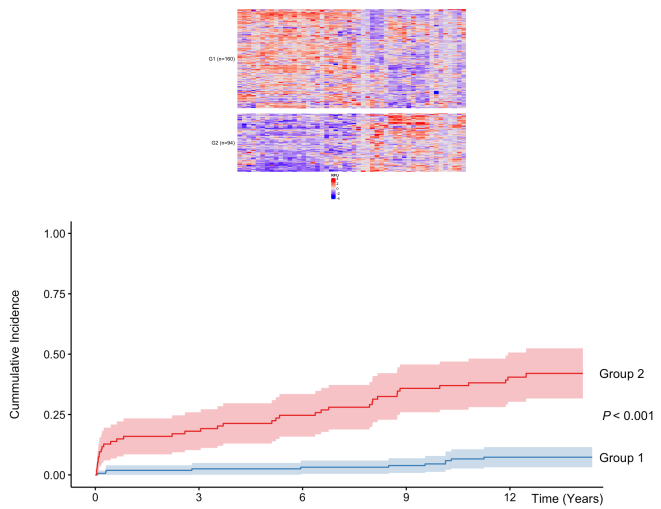
Clustering and occurence of heart failure

Based on these 50 proteins, *k*-means clustering of patients was performed in order to identify groups of patients whose centers can be used as a prognosis tool.

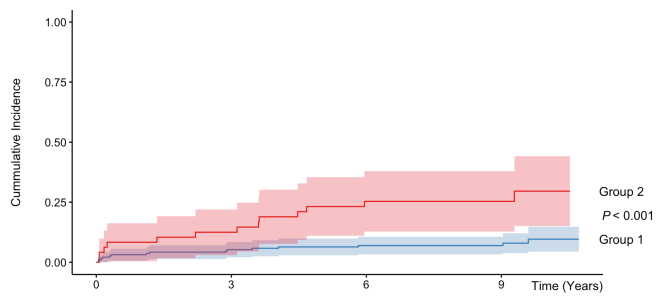
2 clusters of patients were identified on REVE-1 and applied to REVE-2 showing opposite proteomic expression profiles and distinct clinical profiles (age, diabetes, hypertension).

Clusters were used as a new variable to predict occurence of heart failure.

Cluster effect was significant on both cohorts even when adjusted on clinical variables.



REVE-1



REVE-2

Overview

Longitudinal data analysis

Objective :

Select proteins whose evolution is linked to the occurence of the outcome "heart failure or cardiovascular death"

Statistical framework : construction of a predictive model for a variable Y with a set of variables $X_1, \ldots, X_p$ measured over n individuals at t times.

Notations :

- $Y \in \mathbb{R}^n$: time before the occurrence of a heart failure or cardiovascular death (survival analysis)
- $X(t) \in \mathcal{M}_{n,p}(\mathbb{R})$: Proteomic variables ($p \simeq 5000$, $n \simeq 240$, $t = 4$)

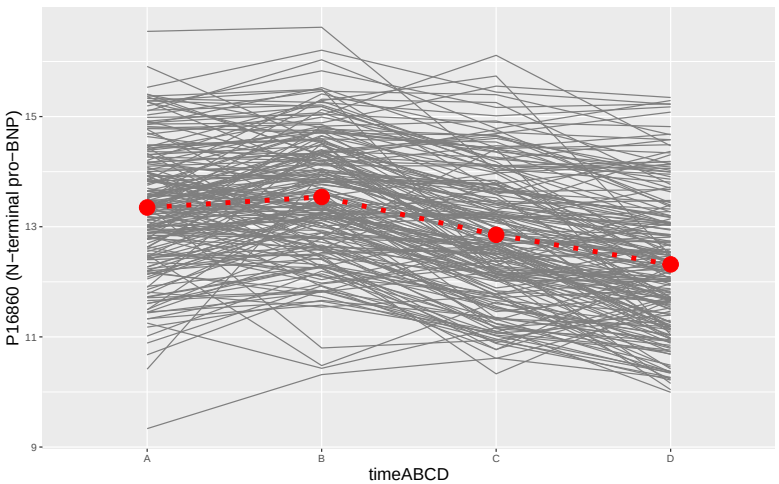
Need to jointly model temporal evolution and survival
$\Rightarrow$ Use of joint models (shared random effects models) (Wulfsohn and Tsiatis, 1997)

Shared random effects models (SREM)

$$\begin{cases} x_{ij}(t_l) = x^*_{ij}(t_l) + \epsilon_{ij}(t_l) \\ x^*_{ij}(t_l) = \beta_{jl} + b_{ijl} \\ \lambda_{ij}(t_l|b_{ij}) = \lambda_0(t_l)e^{x^*_{ij}(t_l)\gamma_j} \end{cases}$$

with :

- $x_{ij}(t_l)$ measurement for patient i of protein j at time t_l ($l \in \{1, 2, 3, 4\}$)
- $\epsilon_{ij}(t_l) \sim \mathcal{N}(0, \sigma_{ij})$
- β_{jl} fixed effect
- b_{jl} random effect
- $\lambda_0(t)$ baseline risk
- γ_j effect of protein j (survival)



Estimation within the maximum likelihood framework by considering the joint likelihood from the two submodels.

Regularized latent class model (rJLCM)

Around 480 proteins found to be of interest following univariate analyses. Need to be more selective.

$\Rightarrow$ Use of regularized latent class model for joint analysis of high-dimensional longitudinal biomarkers and a time to-event outcome (Sun et al., Biometrics 2019)

Work in progress :

- analysis of selected proteins when obtaining 2 clusters of patients for which survival and longitudinal evolution of proteins are different.
- Extend the model to incorporate biological information such as GO categories.

Overview

Conclusion - perspectives

- High dimension problem ($n \ll p$) of omic data offers challenges for statisticians
- Univariate analyses followed by multiple testing enable to model competing risks : better solution than penalized regressions when using the baseline timepoint.
- Aggregation of proteins through clustering helps summarize the data and avoid over-fitting.
- Joint models are appropriate to jointly analyse the repeated measures and the survival. Statistical reasearch is needed to improve the scalability to a high number of biomarkers.