A Principled Approach to Single Predictor Cutpoint Selection to Identify Patients with a Defined Likelihood of Outcome in a Multi-Factorial Risk Prediction:
Application of Joint Density Estimation and Monte Carlo Simulation

The ROYAL MARSDEN — Life demands excellence

Mansour T A Sharabiani; Clare Peckitt; Lawrence Kim; Declan Cahill

2018 PSI CONFERENCE

Background

- Clinicians require simple rule such as cutpoint for classification of patients based on their prognosis. Determining such a cutpoint for a single predictor (primary variable) could be challenging if other predictors are accounted for.
- **Aim**: to define a cutpoint for patient's Membranous Urethral Length (MUL) - primary variable - such that for patients with MUL above the cutpoint, the probability of gaining continence 6 months after prostate surgery is > 90% taking into account other variables (in this study, patients' age)

Problem statement:

- There is a primary variable (patient's MUL), a binary outcome (continence), and one (or more) secondary variable (age).
- There is a positive correlation between the primary variable and the binary outcome
- To seek a cutpoint for the primary variable, such that the probability of the binary outcome (continence) is above a particular threshold (here 90%)

Methods

Patients:

- 190 patients with localised prostate cancer who underwent robot-assisted robotic prostatectomy (RARP) were retrospectively reviewed.
- Continence was defined as no pad or a safety pad.

Probablistic framework:

- Considering the vector of all variables (x), decomposed into the primary (x_i) and the secondary (x_{-i}) parts, $x=[x_i , x_{-i}]$, we can think of two entities:

1. $P(y=1|x)=P(y=1|x_i,x_{-i})$: A binary classification model that predicts the binary outcome, conditioned on the variable
2. $P_c(x)=\begin{cases}\alpha\, P(x) \text{ if } x_i \geq c \\ 0 \quad \text{otherwise}\end{cases}$: **Truncated Joint Density Distribution** for all variables, where α is a constant and c is cutpoint. α is:

$$\int P_c(x)dx = 1 = \alpha \int_{x_i\geq c} P(x)dx \longrightarrow \alpha = \left\{\int_{x_i\geq c} P(x)dx\right\}^{-1}$$

- Expectation of $P(y=1|x)$ with respect to $P_c(x)$ is:

$$f(c) \equiv E_{P_c(x)}[P(y=1|x)] = \int P_c(x)P(y=1|x)dx$$

$$= \left\{\int_{x_i\geq c} P(x)\, P(y{=}1|x)\, dx\right\} \Big/ \int_{x_i\geq c} P(x)dx$$

- To find the right cutpoint, we set the above to the threshold probability (continence), p_0, and solve for cutpoint, c

$$c_{optimal} = f^{-1}(P_0)$$

- All the scripts were written in R
- Bootstrap samples: 2000 samples were generated from the original data. For each bootstrap sample:

1. Conditional Density: Logistic regression generated the conditional density of response
2. Joint Density: Mean vector and covariance matrix of features (age & MUL) calculated.
   - Using mean and covariance, 100,000 Monte Carlo (MC) samples generated from a multivariate normal density.
3. Calculation of Expected Response: Predicted response probability was calculated for each MC generated above.
4. For each cutpoint selected from a grid of values:
   - Mean predicted response probability was calculated for all points where MUL was below the selected cutpoint.
   - For each bootstrap sample, the optimal cutpoint was selected as the smallest cutpoint where the mean predicted response is above the desired threshold (90% chance of gaining continence).
   - The above resulted in a distribution of cutpoints whose median, 0.025% and 97.5% percentiles were chosen to correspond to point estimate and 95% Confidence Intervals, respectively

Results & Conclusions

- Flexibilities arising from the probabilistic framework
  - *A priori assumption about expected joint density distribution (e.g., obtaining covariance matrix from a larger population) can be used*
  - *Valid prediction can be used in the absence of information related to secondary covariates*
- Application of the method to clinical the data showed with 95% confidence that above MUL cutpoint of 18.7mm, 90% of patients regain continence in 6 months cross all ages (Figure 1)

Figure 1: Probability of Incontinence (with 95%CIs) by MUL length cross all ages