Big-RT: Big Data Analysis to Identify Combinatorial Predictors of Radiotherapy Toxicity for Personalised Treatment in Prostate Cancer Patients

Sebastian Pölsterl, Anna Wilkins, Sarah Guillford, James Campbell, Carmen Rodriguez-Gonzalez, Sheng Yu, Veronica Garcia-Perez, Claire Griffin, Judith Bliss, John Yarnold, Uwe Oelfke, David Dearnaley, Emma Hall, Bissan Al-Lazikani, Navita Somaiya

Background

• ~ 7% of cancer patients will receive radiotherapy (RT).
• Up to 20% experience long-term RT-induced toxicity.
• Problem: no stratification by risk of side effects, which limits the ability to personalise treatment.
• Challenge: The pathogenesis of RT-induced toxicity is complex and multi-factorial, yet most prior work focused on isolated data types and standard statistical techniques, with limited success.

Aim

To identify multi-parametric predictors of late RT-induced rectal bleeding by developing novel machine learning techniques that fully integrate clinical, dosimetric, and genetic data from the CHHiP prostate RT-fractionation trial (CRUK/06/016)\(^1,2\).

Data

• 3,212 patients recruited \(\rightarrow\) 721 have all data types.
• 88 patients excluded if rectal bleeding information is missing for \(>3\) time points and/or had haemorrhoids and/or rectal bleeding prior to radiotherapy.
• 633 patients eligible for analysis.
• Rectal bleeding defined as having grade \(\geq 2^1\) at \(\geq 12\) months.

Methods

Dosimetric

Rectal vol Dose Volume Receiving RT dose (Histogram 80 bins)

Clinical (12 variables)

Age, hypertension, pelvic surgery, diabetes, IBD, previous TURP, risk group, Gleason score, pre and post hormone PSA, dose

Genetic

300K–500K genetic variants from RAPPER & PRACTICAL consortia:

• ~9 million after imputation
• ~100-500 pre-selected for analysis with highest correlation to rectal bleeding according to univariate Bayesian tests.\(^3\)

Integrative Modelling

• Evaluated >1300 models including Multilayer Perceptron, in-house hybrid functional-scalar\(^2\), Naïve Bayes, Gradient Boosting, and Support Vector Machine.
• Introduced Bayesian optimisation\(^4\) to select the final model.
• Performed 10-fold cross validation.

Results

Fig 1. Top: Median (AUROC) performance scores for models using clinical, genetic and dosimetric factors. Black lines indicate the range of scores for our 10-fold cross validation. Bottom: Results using subsets of the data and Naïve Bayes, our best performing model. The strongest predictions are achieved by jointly analysing all data types.

Fig 2. Strength of association to late RT-induced rectal bleeding for our top 10 predictors for the Naïve Bayes model using all data types. Genetic features shown were selected in at least 9 out of 10 folds. Higher risk of rectal bleeding was associated with an increase in rectal volume receiving 66-76Gy, whilst lower risk was linked to genetic variants in a region of chromosome 8.

Conclusions

• Identified novel combinatorial markers predictive of RT-induced rectal bleeding.
• Markers must be validated on independent dataset.
• Techniques developed here could be applied to other tumour types treated with radiotherapy.

References