

Purpose:

The use of multivariate normal tissue complication probability (NTCP) models applying logistic regression and automated variable selection has increased in recent years. The extensive data exploration in this methodology to find an optimal subset of predicting factors is often effective. However, the main risk of this approach is overfitting, resulting in lower true prediction power than initially estimated. Bootstrapping is an accepted method to reduce the risk of overfitting. The main purpose of the current study was to quantify its effectiveness for data with typical characteristics for multivariate NTCP modeling and various set sizes by measuring overfitting in simulations.

Methods:

A method was developed to generate simulated data with statistical properties similar to real clinical data sets, enabling repeated modeling and cross-validation with independent data sets. Characteristics of three clinical data sets from radiotherapy treatment of head and neck cancer patients were used to simulate data with set sizes between 50 and 1000 patients. We implemented a bootstrapping method using forward variable selection. We measured for each resulting model the estimated and true prediction power, and the selected and true optimal number of included variables.

Results:

Bootstrapping selects on average the true optimal number of variables for all different data characteristics and set sizes (mean deviation: -0.32 ± 0.20 SEM), but with considerable spread. Both the true and estimated prediction power converge asymptotically towards a maximum prediction power for large data sets, indicating that, despite the spread around the optimal number of selected variables, the bootstrapping technique is not overfitting for data sets of sufficient size. Severe overfitting (true prediction power worse than random guessing) was found in our analysis only for small data sets (95% of cases had <33 events).

Conclusions:

Bootstrapping guards multivariate NTCP modeling against overfitting for data sets of sufficient size (typically >32 events in our simulations).