Random forest classification predicts response to recombinant growth hormone (r-GH) in growth hormone deficient (GHD) children using baseline clinical parameters and genetic markers

A Stevens1, P Murray2, J Wojcik3, J Raeselon4, E Kedgova1, P Chatelan1, F Clayton1

1Institute of Human Development, Faculty of Medical and Human Sciences, University of Manchester and Manchester Academic Health Science Centre, Royal Manchester Children’s Hospital, Central Manchester University Hospitals NHS Foundation Trust, Manchester, UK; 2Quantic Bios, Geneva, Switzerland; 3Brionten BioSciences, St Laurent, Quebec, Canada; 4Mennd Senoni, Qiomstadt, Germany; 5Department Pediatric, Hospital Men and Infant -- Universit Claude Bernard, Lyon, France

Introduction to Random Forest Classification

Overview

- Random forest (RF) is a machine learning method, based on an ensemble of decision trees, i.e. a forest (Panel A). This process is repeated multiple times to build an overall model.

Panel A: Schematic Decision Tree

A schematic representation of the variable selection and their positive consequences, useful for improved decision-making analysis.

 heir and variables

heir and variables

heir and variables

Methods

- We used pre-pubertal GHD children (peak GH <10 μg/L) from the PREDICT ITU study (n=113) and PREDICT validation (SVH) study (n=141,494, n=219).

- Single nucleotide polymorphisms (SNPs) previously identified to be associated with first year growth response to GH were grouped into 2 groups (Table 1).

- Random forest classification (RFC) was undertaken to identify variables associated with response (change in height %SDS) categorized using the median value in relation to the baseline clinical variables of:
  - gender
  - age
  - GH peak (average daily dose by body weight mg/kg/day)
  - distance to target height SDS (0.75)
  - mid-parental height SDS (0.6)
  - GH peak (mg/L)

- Accuracy (true positives + true negatives/ total population) of the RFC models was assessed and a variable importance score (BFS) calculated by permutation.

- AUC under the curve (AUC) of the Random Operating Characteristic curve is a measure of how well a model can distinguish between two diagnostic groups (disease vs. normal).

Results

- RFC demonstrated that basal clinical variables could predict growth response (change in height %SDS) (r=0.11, p<0.05) (Figure 1a).

- Accuracy 80.66%

- AUC 0.83%

Conclusions

- The Random forest models predict 82% of the variability in first year response in GHD with GH peak as the most significant variable.

- The set of clinical variables in this study also generated a very good predictor of growth response using RFC (AUC=0.83).

- Interestingly, two genetic markers alone are positively predictive with an accuracy of 72% (compared with 80% for clinical variables) (See Table 1 and Figure 1b).

References


Acknowledgments

The work was supported by Merck KGaA, Darmstadt, Germany. The authors wish to thank the patients and their families, investigators, co-investigators and the staff members of all participating centers in the PREDICT clinical trials.

Disclosures

The author states that he/she has no conflicts of interest to disclose.