Genetic markers contribute to the prediction of response to GH in severe but not mild GH deficiency

A. Stevens[1], P. Murray[1], J. Wojcik[2], J. Raelson[3], E. Koledova[4], P. Chatelain[5], P. Clayton[1]

[1]Institute 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; [2]Quartz Bio, Geneva, Switzerland; [3]Genizon BioSciences, St Laurent, Quebec, Canada; [4]Merck Serono, Darmstadt, Germany; [5]Department Pediatrie, Hôpital Mère-Enfant – Université Claude Bernard, Lyon, France

European Society for Paediatric Endocrinology (ESPE), Barcelona, Spain, 1-3 October 2015; Poster P1-83 Abstract 953

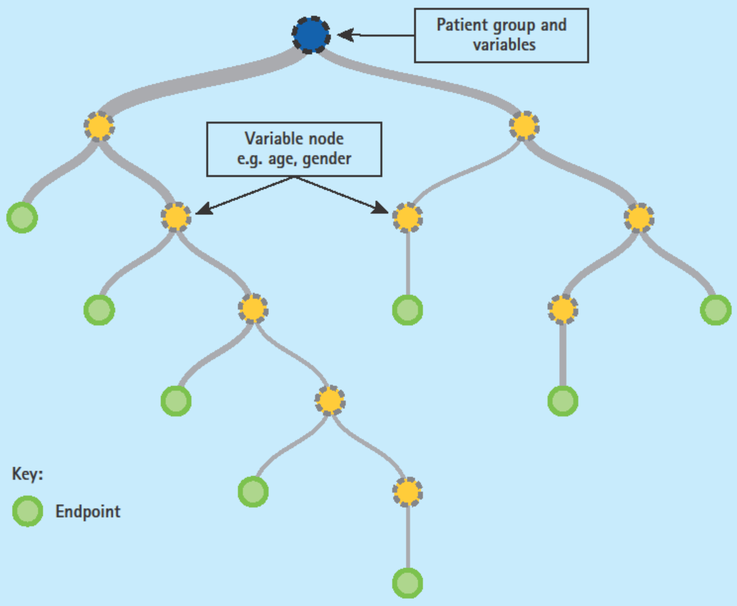
Introduction to Random Forest Classification

Overview

- Random forest[1] is a machine learning method, based on use of 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 available alternatives and their possible consequences, useful for sequential decision-making analyses



How does Random Forest Classification (RFC) work?

- A different subset of the training data are selected to train each tree
- Remaining training data are used to estimate accuracy and variable importance
- Class assignment is made by majority vote across all the trees

Features of RFC

Pro's

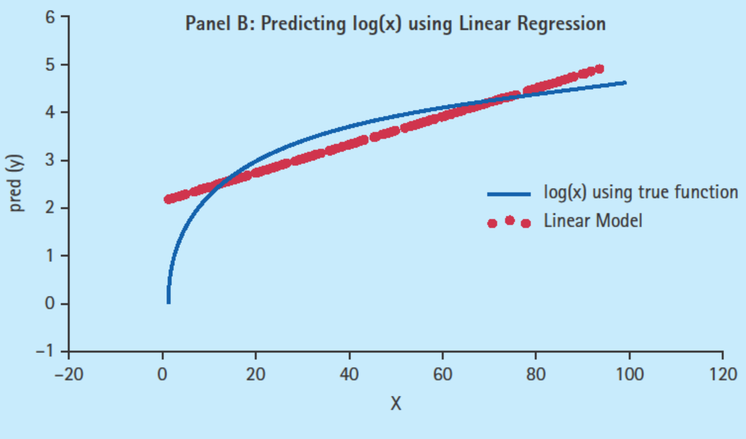
- Unrivalled in accuracy among current algorithms
- Runs efficiently on large databases
- Gives estimates of what variables are important in the classification
- Maintains accuracy when a large proportion of the data are missing
- No problem with overfitting
- Not very sensitive to outliers in the training data
- Generates computations of accuracy and variable importance
- Colinearity between variables has no effect on the analysis

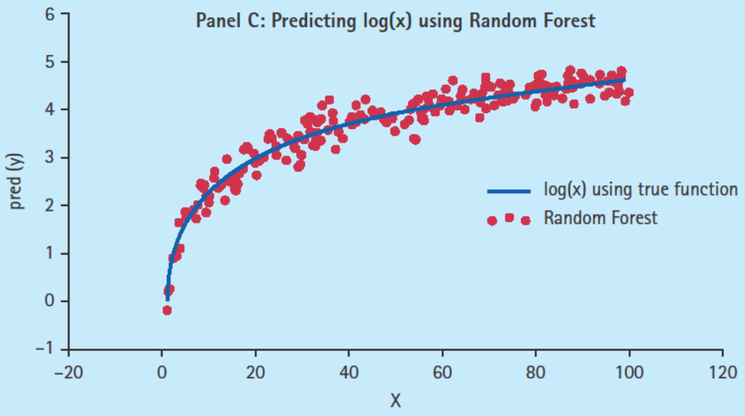
Cons

- Cannot generate a classical regression equation

Example

- If we try and build a basic linear model to predict y using x, the result is a straight line that roughly bisects the log(x) function (**Panel B**). Whereas if we use a random forest, it does a much better job of approximating the log(x) curve and we get something that looks much more like the true function (**Panel C**).





MANCHESTER 1824
The University of Manchester

Introduction

- Single nucleotide polymorphisms (SNPs) associated with the response to GH therapy have previously been identified in Growth Hormone Deficient (GHD) children in the PREDICT long-term follow-up (LTFU) study (NCT00699855).
- The initial exploratory random forest classification (RFC) to predict growth response, performed on the merged PREDICT LTFU Year 1 (Y1) and VALIDATION (VAL) populations, has shown that: (see **Poster P2-418, Abstract 942**)
  - Clinical covariates are a strong predictor of growth endpoints independently of SNPs
  - Adding SNPs to clinical covariates did not improve the prediction power of RFC.
  - However the SNPs alone did have predictive value implying a complex interaction between clinical covariates and genetic markers.

Objectives

- To assess the effect of GHD severity [severe (≤ 4 µg/L) vs mild (>4 µg/L) on the predictive value of genetic markers of first year growth response to recombinant human growth hormone (GH).

Methods

- We used pre-pubertal GHD children (peak GH <10µg/L) from the PREDICT LTFU study (n=113) and PREDICT validation (VAL) study (NCT01419249, n=293).
- An analysis was undertaken in all patients and in groups stratified into severe (≤4 µg/L) and mild GHD (> 4 & <10 µg/L).
- Single nucleotide polymorphisms (SNP) previously identified to be associated with first year growth response to GH (n=22) were genotyped[2] (**Table 1**).
- Random forest classification (RFC)[1] was undertaken to identify variables associated with growth response (change in height [cm]) categorised using the median value in relation to the baseline clinical variables of gender, age, GH dose, distance to target height SDS (DTH), mid-parental height SDS (MPH) and SNPs.
- As we classified patients on the basis of their peak GH, this variable was not included in RFC.
- Accuracy ([true positives + true negatives]/ total population) of the RFC models was assessed and a variable importance score (VIS) calculated.
- Area under the curve (AUC) of the Received Operating Characteristic curve is a measure of how well a parameter can distinguish between two diagnostic groups (e.g. disease vs. normal).

Results

- Growth response in the whole group and severity-stratified sub-populations can be predicted by RFC with high levels of accuracy ($1x10^{-14}$) (**Table 2**).
  - mild GHD. Accuracy 74.9%,
  - severe GHD. Accuracy 74% (**Table 2**).
- Only baseline clinical variables were important in mild GHD.
  - only GH dose and MPH (ranked by VIS) contribute to prediction (**Figure 1**).
- However, genetic and clinical markers contribute to prediction in severe GHD.
  - VIS ranked important variables as followed: DTH, SNP rs1024531 (GRB10), age, SNP rs7101 (FOS), MPH, SNP rs3213221 (IGF2), and GH dose (**Figure 1**).

Table 1. SNPs with previously identified association with growth response used in study

Disease	Response	Gene	SNP	Marker	Non-marker
GHD	Change in height (cm)	CYP19A1	rs10459592	GG	T
		FOS	rs7101	C	TT
		GRB10	rs933360	TT	C
			rs4521715	AA	G
			rs10248619	CC	T
			rs1024531	G	AA
		IGF2	rs3213221	CC	G
		INPPL1	rs2276048	G	AA
		SOS1	rs2888586	T	CC
		SOS2	rs13379306	A	CC

Table 2. GH Peak-stratified Change in Height (cm) Model data

Disease	Endpoint	GHD peak	*N*	Accuracy	AUC	*P*-value
GHD	Change in height (cm)	ALL	389	74.0%	86.8%	1.3e- 35
		mild	235	74.9%	85.0%	3.0e- 17
		severe	154	74.0%	88.4%	7.3e- 15

GHD patient numbers for change in height over one year (cm) of treatment with recombinant human growth hormone. Accuracy ([true positives + true negatives]/ total population), area under the curve (AUC) [a measure of how well a parameter can distinguish between two diagnostic groups (e.g. disease vs. normal)].

Conclusions

- **Accuracy of prediction of growth response (change in height [cm]) is similar in the whole group and GHD severity-stratified sub-populations**
- **However, the important variables differ:**
  - **In the whole group the key variable is DTH**
  - **In the mild sub-population GH dose and MPH**
  - **In the severe sub-population:**
    - **the clinical variables are GH Dose, MPH, Age and DTH**
    - **the SNPS are rs3213221 (IGF 2), rs1024531 (GRB10), rs7101 (FOS)**

References

1. Breiman, L and Cutler, A. Random Forests. Available at: http://www.stat.berkeley.edu/users/breiman/RandomForests/cc_papers.htm
2. Clayton P et al. Eur J Endocrinol 2013; 169(3): 277-89.

GET POSTER PDF

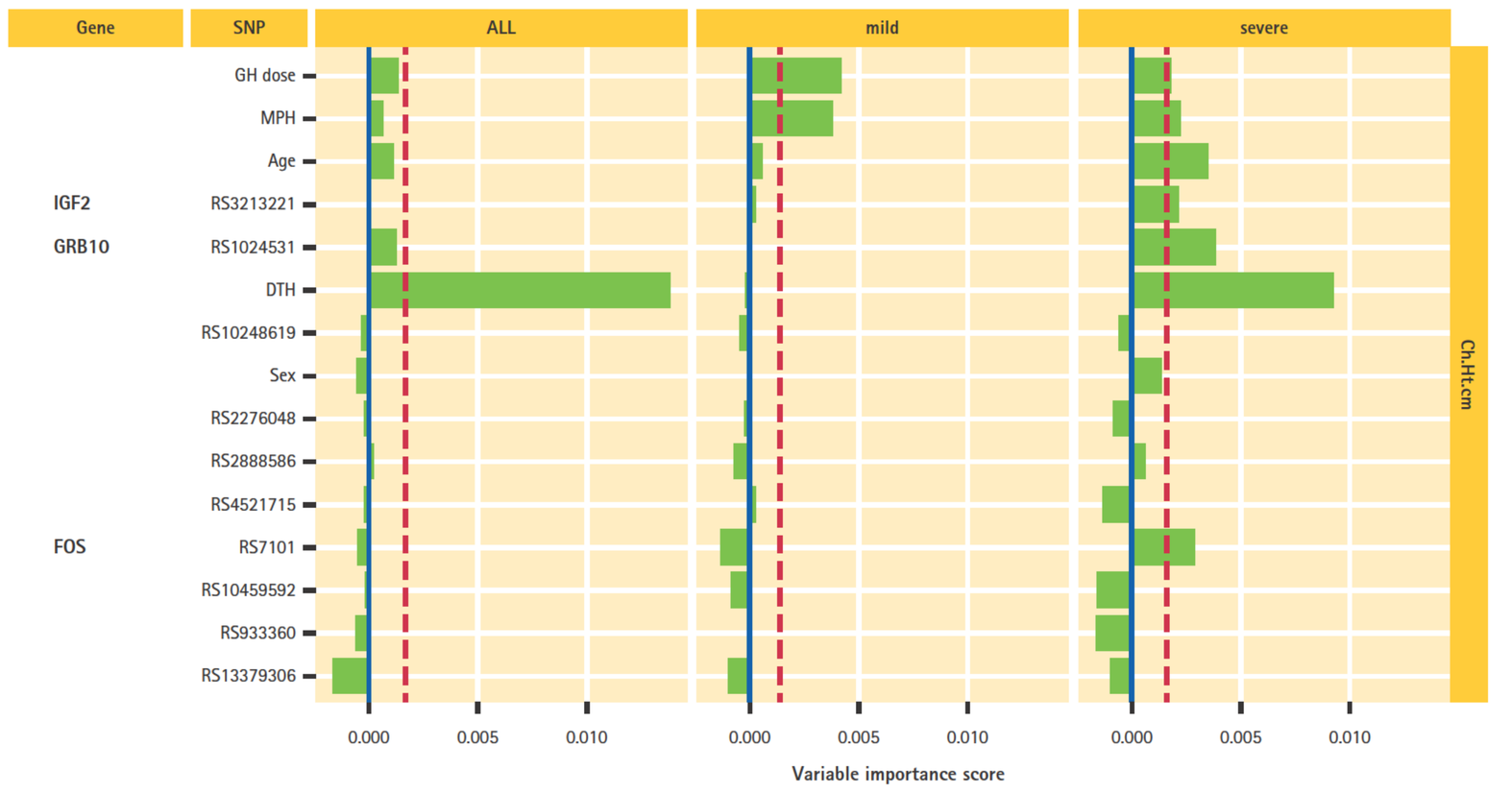
Acknowledgments

The trial was sponsored by Merck KGaA, Darmstadt, Germany. The authors would like to thank the patients and their families, investigators, co-investigators and the study teams at each of the participating centers and at Merck KGaA, Darmstadt, Germany. Medical writing assistance was provided by David Candlish, inScience Communications, Chester, UK, funded by Merck KGaA, Darmstadt, Germany.

Disclosures

AS, PCh, PCl, honoraria and research grants from Merck KGaA Darmstadt, Germany. JW is an employee of Quartz Bio, Geneva, Switzerland. JR is an employee of Genizon Biosciences, Quebec, Canada. EK is an employee of Merck KGaA, Darmstadt, Germany. PM declares no financial interest.

Figure 1. GH Peak-stratified Change in Height (cm) random forest variable importance



Variables used to predict change in height over one year of recombinant human growth hormone therapy. Variable importance score (VIS) calculated by permutation, threshold of significance shown by dotted red line. MPH = mid-parental height, DTH = Distance to Target Height, GH dose = dose of recombinant human growth hormone, RS numbers of a range of single nucleotide polymorphisms previously associated with growth response [gene names of associated SNPs shown where significant].