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 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 P2-418 Abstract 942

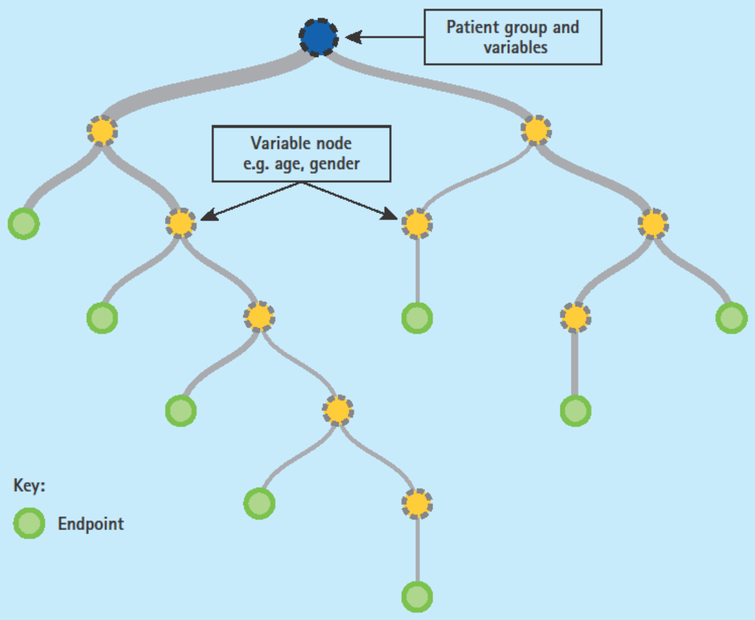
Introduction to Random Forest Classification

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

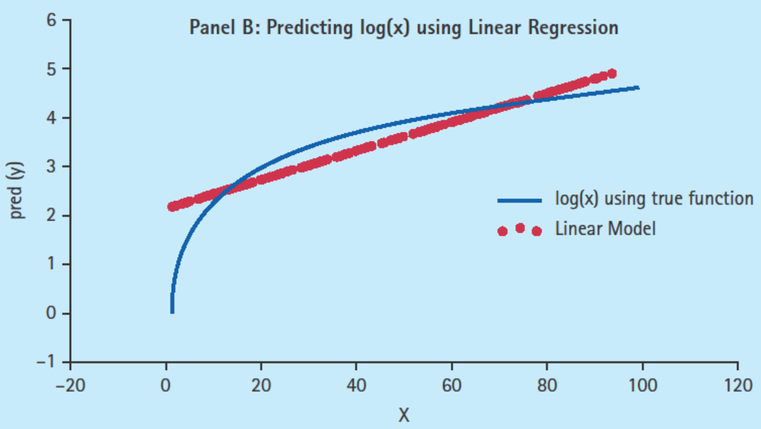
- Unrivalled in accuracy among current algorithms
- Runs efficiently on large databases
- Gives estimates of what variables are important in the classification
- Maintains accuracy when a large proportion of the data are missing
- No problem with overfitting
- Not very sensitive to outliers in the training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

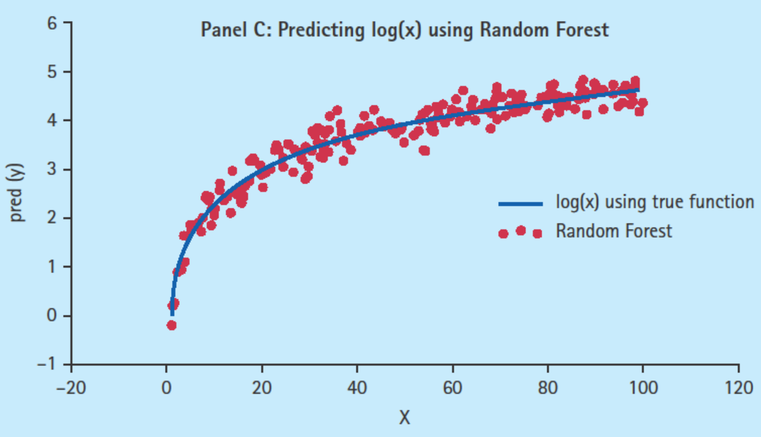
Cons

- 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

- Prediction of response to r-GH is currently based on regression modelling1. This approach generates a prediction equation which can be applied to data from an individual child. However this method can underestimate the effect of inter-dependent variables. Random forest classification (RFC) is an alternative prediction method based on decision trees that is not sensitive to the relationships between variables[2] (see **Poster P1-83, Abstract 953**).

Objective

- To assess the predictive value of RFC in first year growth response to recombinant human growth hormone (GH) in GHD children.

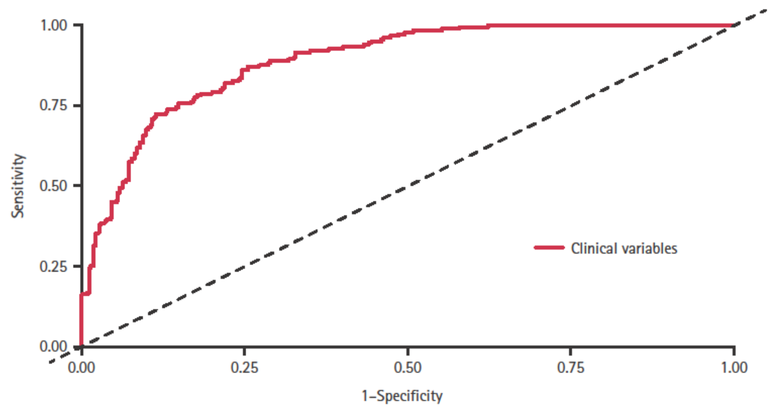
Methods

- 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).
- Single nucleotide polymorphisms (SNP) previously identified to be associated with first year growth response to GH were genotyped[3] (**Table 1**).
- Random forest classification (RFC) 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 (average daily dose by body weight [mg/kg/day])
  - distance to target height SDS (DTH)
  - mid-parental height SDS (MPH)
  - GH peak (µg/L).
- Accuracy ([true positives + true negatives]/ total population) of the RFC models was assessed and a variable importance score (VIS) calculated by permutation.
- 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 (disease vs. normal).

Results

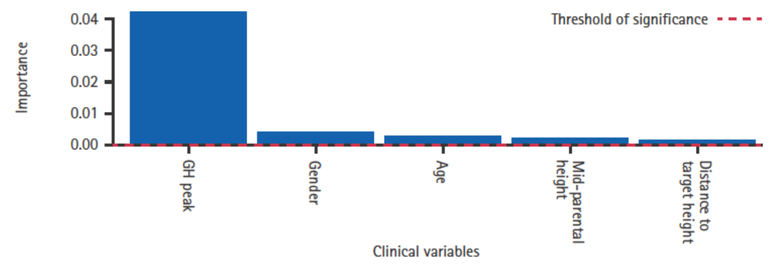
- RFC demonstrated that basal clinical variables could predict growth response (Change in height [cm]) ($p<1.1x10^{-39}$) (**Figure 1a**)
  - Accuracy 80.6 %
  - AUC 88.3%

Figure 1A. Received Operating Characteristic curve showing the use of clinical variables as predictors of first year growth response to recombinant human growth hormone categorised by median value by random forest classification (402 GHD patients)



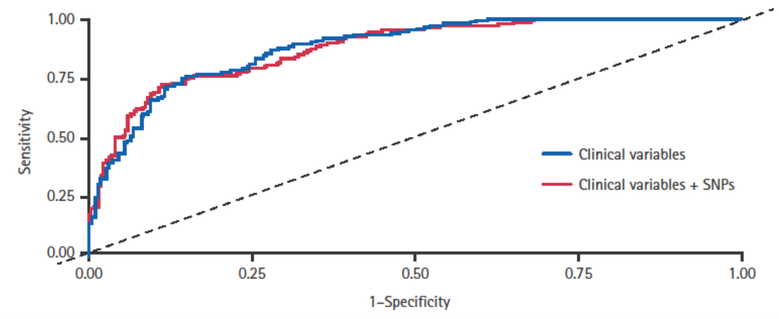
- The variables were ranked by VIS as follows (**Figure 1b**):
  - GH peak
  - gender
  - age
  - mid-parental height SDS
  - distance to target height SDS

Figure 1B. Variable importance of clinical variables as predictors of first year growth response to recombinant human growth hormone identified by random forest classification (402 GHD patients)



- The addition of genetic data did not improve prediction ($p< 2.8 x10^{-38}$) (**Figure 2**)
  - Accuracy 80.7%
  - AUC 88.2%

Figure 2. Received Operating Characteristic curve showing the use of clinical variables (blue line) and clinical variables along with single nucleotide polymorphisms (SNPs) (red line) as predictors of first year growth response to recombinant human growth hormone by random forest classification (389 GHD patients)



- However SNPs alone could act as weaker but distinct predictors of growth response ($p< 1.9 x10^{-13}$) (**Figure 3A & 3B**)
  - Accuracy 65.4 %
  - AUC 71.6%
  - The SNPs with predictive value were:
    - rs1024531 (GRB10)
    - rs7101 (FOS).

Figure 3A. Received Operating Characteristic curve showing the use of clinical variables (red line) and just single nucleotide polymorphisms (SNPs) (blue line) as predictors of first year growth response to recombinant human growth hormone categorised by median value by random forest classification (393 GHD patients)

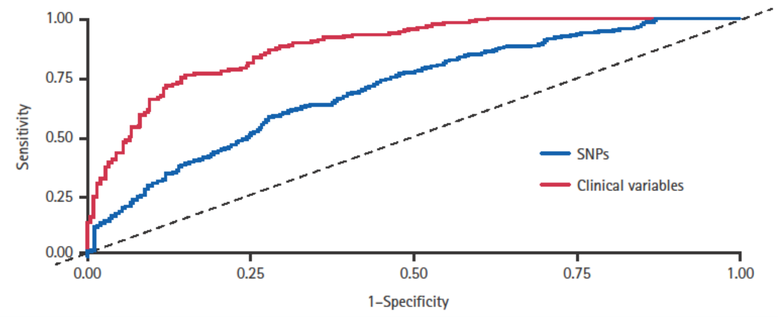


Figure 3B. Variable importance of single nucleotide polymorphisms (SNPs) as predictors of first year growth response to recombinant human growth hormone categorised by median value identified by random forest classification (393 GHD patients)

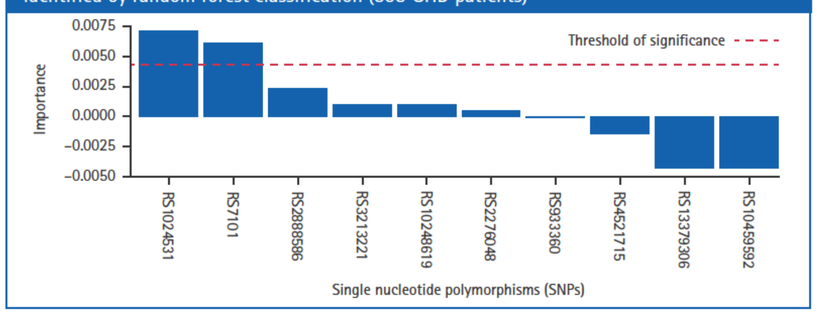


Table 1. SNPs with previously identified association with growth response used in study

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

Conclusions

- The Ranke regression model[1] predicts 65% 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 generates a very good predictor of growth response using RFC (AUC~90%).
- Interestingly, two genetic markers alone are positively predictive with an accuracy of 72% (compared with 88% for clinical variables (rs1024531 [GRB 10] and rs7101 [FOS]).

References

1. Ranke MB et al JCEM 1999; 84:1174-83.
2. Breiman, L and Cutler, A. Random Forests. Available at: http://www.stat.berkeley.edu/users/breiman/RandomForests/cc_papers.htm
3. 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.

ESPE 2015 418--P2
GH and IGF Treatment
Adam Stevens
DOI: 10.3252/pso.eu.54espe.2015
Poster presented at: ESPE 2015 BARCELONA
PosterSessionOnline.com