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

A. Stevens1, P. Murray1, J. Wojcik2, J. Raebone1, E. Koledova1, P. Chatelain2, P. 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; 2Quartz Bio, Geneva, Switzerland; 3Geoniz BiSciences, St Laurent, Quebec, Canada; 4Merck Serono, Darmstadt, Germany; 5Department Pediatric, Hôpital Mère-Enfant – Université Claude Bernard, Lyon, France

Introduction

Brief history of GH and IGF Treatment

GH and IGF Treatment

Introduction to Random Forest Classification

Overview

- Random forest 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 outcomes, useful for sequential decision-making analysis.

How does Random Forest Classification (RFC) work?

- A different subset of the training data are selected in each tree branch.
- 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

Pros

- Involves accuracy among current algorithms.
- Works efficiently on large datasets.
- 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.
- Consistency between variables has no effect on the analysis.
- 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 touches the linear function (Panel B). Whereas if we use a random forest, it does a much better job of approximating the log(y) curve and we get something that looks much more like the true function (Panel C).

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

<table>
<thead>
<tr>
<th>Disease</th>
<th>Response</th>
<th>Gene</th>
<th>SNP</th>
<th>Marker</th>
<th>Minor-allele</th>
</tr>
</thead>
<tbody>
<tr>
<td>GH</td>
<td>Change in height (cm)</td>
<td>GH14A</td>
<td>rs104500538</td>
<td>GD</td>
<td>T</td>
</tr>
<tr>
<td>GH</td>
<td>Change in height (cm)</td>
<td>rs1071097</td>
<td>C</td>
<td>TT</td>
<td></td>
</tr>
<tr>
<td>GH</td>
<td>Change in height (cm)</td>
<td>rs10204114</td>
<td>T</td>
<td>TT</td>
<td></td>
</tr>
<tr>
<td>GH</td>
<td>Change in height (cm)</td>
<td>rs10204114</td>
<td>A</td>
<td>A</td>
<td></td>
</tr>
<tr>
<td>GH</td>
<td>Change in height (cm)</td>
<td>rs10204114</td>
<td>G</td>
<td>G</td>
<td></td>
</tr>
<tr>
<td>GH</td>
<td>Change in height (cm)</td>
<td>rs10204114</td>
<td>C</td>
<td>C</td>
<td></td>
</tr>
<tr>
<td>GH</td>
<td>Change in height (cm)</td>
<td>rs10204114</td>
<td>G</td>
<td>G</td>
<td></td>
</tr>
<tr>
<td>GH</td>
<td>Change in height (cm)</td>
<td>rs10204114</td>
<td>C</td>
<td>C</td>
<td></td>
</tr>
<tr>
<td>GH</td>
<td>Change in height (cm)</td>
<td>rs10204114</td>
<td>G</td>
<td>G</td>
<td></td>
</tr>
<tr>
<td>GH</td>
<td>Change in height (cm)</td>
<td>rs10204114</td>
<td>C</td>
<td>C</td>
<td></td>
</tr>
<tr>
<td>GH</td>
<td>Change in height (cm)</td>
<td>rs10204114</td>
<td>G</td>
<td>G</td>
<td></td>
</tr>
</tbody>
</table>

Table 2. DRI Peak-trasformed Change in Height (cm) Model data

<table>
<thead>
<tr>
<th>Disease</th>
<th>Endpoint</th>
<th>GH peak</th>
<th>N</th>
<th>Accuracy</th>
<th>AUC</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>GH</td>
<td>Change in height (cm)</td>
<td>ALL</td>
<td>208</td>
<td>74.9%</td>
<td>0.86</td>
<td>1.38e-11</td>
</tr>
<tr>
<td>GH</td>
<td>Change in height (cm)</td>
<td>nD</td>
<td>205</td>
<td>74.9%</td>
<td>0.86</td>
<td>1.38e-11</td>
</tr>
<tr>
<td>GH</td>
<td>Change in height (cm)</td>
<td>severe</td>
<td>154</td>
<td>74.9%</td>
<td>0.86</td>
<td>1.38e-11</td>
</tr>
</tbody>
</table>

Methods

- We used pre-pubertal GH11 dosage (peak GH: <15μg/l) from the PREDICT study (n=1153) and PREDICT trial 1 (PREDICT trial 1 study (n=1154)) to classify patients.
- An analysis was undertaken in all patients and in subgroups classified into severe (44μg/l) and mild GH (<4.6μg/l).
- Single nucleotide polymorphisms (SNPs) previously identified to be associated with first year growth response to GH (rs10204114, rs10204114, rs10204114, rs10204114) were selected for analysis (Panel B).
- Genetic interaction (based on the linear model of GH14A) was undertaken to identify variables associated with growth response (Panel B).
- The change in height (cm) target was used using the median value in relation to the baseline clinical variables of gender, age, GH dose, distance to target height (SDI), mid-parental heights (SDH) and BMI.
- As we classified patients on the basis of their peak GH, this variable was not included in the RFC model.
- Accuracy (true positives + true negatives / total population) of the RFC models was assessed and a variable importance score (VIMP) calculated.
- Area under the curve (AUC) of the Random Forest Classification curve is a measure of how well a prediction function distinguishes 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 andGH severity-stratified sub-populations.
- However, the important variables differ:
  - In the whole group the key variable is GH14A.
  - In the mild sub-population GH dose and MPN.
  - In the severe sub-population:
    - the clinical variables are GH Dose, BMI, Age and SDH.
  - The SNPs: rs10204114 (DRI), rs104500538 (GH14A, n=7601 (DRI, GH14A, n=7601 (DRI) rs10204114 (GH14A, n=7601 (DRI))

References


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 centres and at Merck KGaA, Darmstadt, Germany. Medical writing assistance was provided by David Castilla, Infusion Communications, Cheltenham, UK, funded by Merck KGaA, Darmstadt, Germany.

Disclosures

No financial or other relationships exist: Research grants and research support from Merck KGaA, Darmstadt, Germany. A. Stevens is an employee of Pfizer UK, London, UK. P. Murray is an employee of Merck KGaA, Darmstadt, Germany. P. Clayton is an employee of Merck KGaA, Darmstadt, Germany. The abstract does not discuss any patent.