Proposal of a clinical response score and predictors of clinical response to 2 years of GH replacement therapy in adult GH deficiency

Harald J Schneider1, Michael Buchfelder2, Henri Wallaschofski3, Anton Luger4, Gudmundur Johannsson5, Peter H Kann6 and Anders Mattsson7

1Medizinische Klinik und Poliklinik IV, Klinikum der Universität München, Ziemssenstr. 1, 80336 Munich, Germany, 2Klinik für Neurochirurgie, Universität Erlangen; Erlangen, Germany, 3Institut für Klinische Chemie und Labormedizin, Universität Greifswald; Greifswald, Germany, 4Universitätsklinik für Innere Medizin III, Medizinische Universität Wien, Vienna, Austria, 5Department of Endocrinology, Sahlgrenska Academy, University of Gothenburg and Sahlgrenska University Hospital, Gothenburg, Sweden, 6Klinik für Innere Medizin, Universität Marburg, Marburg, Germany and 7Pfizer Endocrine Care; Sollentuna, Sweden

Correspondence should be addressed to HJ Schneider
Email harald.schneider@med.uni-muenchen.de

Abstract

Objective: There is no single clinical marker to reliably assess the clinical response to growth hormone replacement therapy (GHRT) in adults with growth hormone deficiency (GHD). The objective of this study was to propose a clinical response score to GHRT in adult GHD and to establish clinical factors that predict clinical response.

Design: This was a prospective observational cohort study from the international KIMS database (Pfizer International Metabolic Database).

Methods: We included 3612 adult patients with GHD for proposing the response score and 844 patients for assessing predictors of response. We propose a clinical response score based on changes in total cholesterol, waist circumference and QoL-AGHDA quality of life measurements after 2 years of GHRT. A score point was added for each quintile of change in each variable, resulting in a sum score ranging from 3 to 15. For clinical response at 2 years, we analysed predictors at baseline and after 6 months using logistic regression analyses.

Results: In a baseline prediction model, IGF1, QoL-AGHDA, total cholesterol and waist circumference predicted response, with worse baseline parameters being associated with a favourable response (AUC 0.736). In a combined baseline and 6-month prediction model, baseline QoL-AGHDA, total cholesterol and waist circumference, and 6-month change in waist circumference were significant predictors of response (AUC 0.815).

Conclusions: A simple clinical response score might be helpful in evaluating the success of GHRT. The baseline prediction model may aid in the decision to initiate GHRT and the combined prediction model may be helpful in the decision to continue GHRT.

Introduction

Growth hormone replacement therapy (GHRT) exerts several effects in adult patients with growth hormone deficiency (GHD), including improvement in lipids, bone density, quality of life and body composition. This has been shown both in randomized placebo-controlled trials (1) and in large-scale observational studies (2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12).

GHRT is an expensive therapy and has a potential risk of side effects. Therefore, it is important to identify those patients who will benefit most from GHRT. In children, growth velocity is a good and simple clinical marker that allows monitoring response to GHRT. Moreover, reliable prediction models for final height have been established in children, based on growth response to GHRT (13, 14, 15).
However, in adults, response to GHRT cannot be measured reliably with one single clinical marker. This makes it difficult to quantify clinical response and, thus, identify potential predictors of clinical response in adult GHD. Generally, the dose of GH is titrated by IGF1 levels. However, it is not known if improvement in IGF1 automatically translates into a relevant improvement of the clinical condition of GHD patients.

Therefore, we aimed to propose a generalized criterion of clinical response to GHRT in adult patients with GHD using values from simple measurements that are performed during standard follow-up. In addition, we aimed to identify factors at baseline and during the early course of treatment that predict such a clinical response to GHRT.

We analyzed data from the KIMS database for this study.

**Subjects and methods**

**Study cohort**

KIMS (Pfizer International Metabolic Database), a long-term safety and outcome study of GH replacement therapy in adults, represents a large pharmaco–epidemiological survey initiated in 1994 (16). All patients gave written informed consent for their participation in KIMS. We included all GH replaced subjects from KIMS worldwide with data at baseline and 2 years on QoL-AGHDA, waist circumference and total cholesterol, except patients who had been GH-replaced before KIMS start (only treatment-naive patients were included), patients diagnosed with Cushing’s syndrome, acromegaly and craniopharyngioma and patients from the UK. Patients with Cushing’s syndrome, acromegaly and craniopharyngioma were excluded because of the specific metabolic profiles of these diseases (6, 8). Patients from the UK were not included because QoL-AGHDA score is a criterion for GH subscription in UK, which might have caused potential bias. The data freeze was from August 20, 2012.

**Methods**

In KIMS, a large number of clinical parameters are documented. Insulin-like growth factor 1 (IGF1) was measured centrally and lipids were measured centrally or locally using standard procedures.

An analysis of 1926 observations with parallel central and local measurements of cholesterol showed a median and mean difference of 0.00 and 0.03 mmol/l. We considered this difference clinically not relevant and continued using central or local laboratory measurements for cholesterol with central measurement as preference when parallel measurements were reported for a single patient.

Until November 2002, serum IGF1 was determined by RIA after acid–ethanol precipitation of IGF-binding proteins (Nichols Institute Diagnostics, San Juan Capistrano, CA, USA). Thereafter, a chemiluminescence immunoassay (Nichols Advantage System, Bad Vilbel, Germany) was introduced. Long-term reproducibility, measured during > 1 year, showed a coefficient of variation < 9% in the concentration range of 130–850 mg/l. The assay detection limit was 30 mg/l (16). For each assay, age- and gender-specific reference parameter estimates were used to calculate IGF1 SDS.

**Statistical analyses**

For the purpose of a simple and easy to-use clinical response tool it seemed most promising to choose variables that respond well to GH replacement, are routinely assessed during follow-up and are not, or only weakly, correlated (for example, when choosing between total cholesterol and LDL cholesterol, only one should be selected, since both their baseline values and their response to GH are expected to be similar).

Variables that fulfilled these criteria were quality of life (measured by QoL-AGHDA), waist circumference and total or LDL cholesterol. QoL-AGHDA is a score consisting of 25 problem items to be answered with yes or no. A score of 25 indicates problems on all items and 0 no problems on any item. Analysis of KIMS has shown that the maximum effects of GH treatment, particularly on lipids, are seen after 2 years of treatment (5, 6, unpublished data). Therefore, we assessed the change (‘delta’) on the following variables before and after 2 years of GH treatment in KIMS: QoL-AGHDA, waist circumference and total cholesterol.

We chose total cholesterol rather than LDL cholesterol for reasons of simplicity: previous data have shown that cholesterol changes are mainly driven by changes in LDL cholesterol, whereas HDL cholesterol remains unchanged after GH therapy. Total cholesterol is the easiest to determine.

We tested linear correlations with Pearson’s and Spearman’s correlation analyses. We checked for normality by visual analysis.

Quintiles and their cut-off points were determined based on the 2-year delta for the respective variables. By adding one score point for each quintile (1 to 5) for each of the three delta variables, an overall response score
of 3–15 points was achieved. A minimum score of three indicates the lowest level of response (quintile 1) for each delta variable and a maximum score 15 indicates the highest level of response (quintile 5) for each delta variable. We considered all three parameters equally important. Therefore, we gave equal weight to each parameter.

Based on these scores, a dichotomous response variable \( Y \) was constructed, where response \( (Y=1) \) was defined as having a score of 7 or more and non-response as having a score of 6 or less \( (Y=0) \). The reason for this cut off was that the lowest two quintiles for a particular delta variable were associated with a worsening, whereas the upper three quintiles were associated with an improvement of that outcome (see Table 1). In clinical practice, a deterioration of the three response variables is observed over time without treatment \( (2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12) \). Therefore, we already considered an unchanged condition represented with a score of 7 as a response. This classification does not imply that those patients with a score of <7 did not show any clinical response to GH treatment.

In the aim of modelling the 2-year response probabilities, two logistic regression models were assessed: Model 1 examined eight KIMS baseline variables as potential response predictors. These were QoL-AGHDA, waist circumference, total cholesterol, height, age, gender, IGF1 SDS and the number of additional pituitary deficiencies. We also incorporated quadratic terms for all numerical variables. Analyses were performed by backward variable elimination. (Proc logistic, SAS 9.2). To test for potential inclusion bias we compared clinical variables of patients included in model 1 and patients excluded from model 1.

Model 2 incorporated all significant baseline variables from model 1 and three 6-month numerical delta variables. Final models were reached by backward elimination procedures. (Proc logistic, SAS 9.2). In further steps we incorporated only those baseline variables that remained significant in model 1 and in a step-by-step approach we consecutively tested each single 6-month numerical delta in QoL-AGHDA, cholesterol and waist circumference. Finally, we incorporated only the significant baseline predictors and the delta variable that contributed most to predictive power and had the lowest missing values.

All variables, except the categorical 2-year response variable, gender and the number of additional pituitary deficiencies, were numerical. A quadratic term was also assessed for numerical variables that were found as statistically significant.

A variable was eliminated from a model if the Wald-based significance value was >5%. Significance level was set to 5%. CI were Wald-based.

Areas under the curve (AUC) were calculated with ROC methods (Proc logistic, SAS 9.2). Generally, AUC values between 0.70 and 0.90 are considered ‘moderate accuracy’ and values between 0.90 and 1 as ‘high accuracy’ in classifying a diagnostic test (17). Goodness of fit of the models was performed with the Hosmer-Lemeshow test (18). For outlier influence assessment we used residual analyses and leave-one-out methods (19, 20, 21).

In sensitivity analyses, we tested models 1 and 2 using \( \geq 5 \) and \( \geq 9 \) instead of \( \geq 7 \) as alternative cut-off values for the response score.

**Results**

**Correlations**

All variables except baseline AGHDA score were close to normally distributed. The Pearson correlation coefficient between baseline cholesterol and baseline waist circumference was \(-0.04\); between baseline cholesterol and baseline QoL-AGHDA score \(0.009\); between baseline waist circumference and baseline QoL-AGHDA score \(0.076\). The correlation between 2-year delta cholesterol and 2-year delta waist circumference was \(0.11\); between

<table>
<thead>
<tr>
<th>Quintile</th>
<th>Waist circumference</th>
<th>QoL-AGHDA</th>
<th>Total cholesterol</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quintile 1</td>
<td>Delta W &gt;3 cm</td>
<td>Delta A ( \geq 2 ) items</td>
<td>Delta C &gt;0.4 mmol/l</td>
</tr>
<tr>
<td>Quintile 2</td>
<td>0.0 (&lt;\Delta ) Delta W ( \leq 3 ) cm</td>
<td>0 (\leq \Delta ) Delta A ( \leq 1 ) items</td>
<td>(-0.1) (&lt;\Delta ) Delta C ( \leq 0.4) mmol/l</td>
</tr>
<tr>
<td>Quintile 3</td>
<td>(-2.0) (\leq \Delta ) Delta W ( \leq 0.0) cm</td>
<td>(-3) (\leq \Delta ) Delta A ( \leq -1 ) items</td>
<td>(-0.5) (&lt;\Delta ) Delta C ( \leq -0.1) mmol/l</td>
</tr>
<tr>
<td>Quintile 4</td>
<td>(-5.8) (\leq \Delta ) Delta W ( &lt; -2.0) cm</td>
<td>(-7) (\leq \Delta ) Delta A ( &lt; -4 ) items</td>
<td>(-1.1) (&lt;\Delta ) Delta C ( \leq -0.5) mmol/l</td>
</tr>
<tr>
<td>Quintile 5</td>
<td>Delta W ( \leq -5.8) cm</td>
<td>Delta A ( \leq -8 ) items</td>
<td>Delta C ( &lt; -1.1) mmol/l</td>
</tr>
</tbody>
</table>

Delta W, 2-year change in waist circumference; Delta A, 2-year change in QoL-AGHDA score; Delta C, 2-year change in total cholesterol.
2-year delta cholesterol and 2-year delta QoL-AGHDA score 0.03; and between 2-year delta waist circumference and 2-year delta QoL-AGHDA score −0.01. Spearman correlation analysis yielded similar results. This confirmed our primary assumption that the three clinical response variables were not or only weakly inter-correlated.

Quintiles
Out of 16 266 patients in KIMS at data close, 3612 patients had data at baseline and 2 years on QoL-AGHDA, or waist circumference or total cholesterol, and fulfilled the other inclusion and exclusion criteria. Quintiles were based on all patients with full baseline and 2-year data on the respective variable (waist circumference \( n = 2110 \), QoL-AGHDA \( n = 1545 \) total cholesterol \( n = 1417 \)). Table 1 displays the quintile values.

Predictors of response
A total of 844 patients fulfilled all inclusion and exclusion criteria and had full data at baseline and 2 years on QoL-AGHDA, waist circumference and total cholesterol. Table 2 displays the baseline characteristics of these patients. Note, however, that due to a varying number of missing values on candidate predictor variables, the different analyses were based on a different number of observations.

Model 1 was based on 791 patients due to 51 missing observations on IGF1 SDS and 2 on height at baseline. Comparison of clinical characteristics between the 791 patients included in model 1 and those not included showed no significant differences in 2-year changes of cholesterol, AGHDA score, or waist circumference and baseline age, cholesterol, IGF1 SDS, or the number of additional deficiencies. However there were clinically marginal but statistically significant differences in AGHDA score (excluded patients mean +1.2 scores points; \( P < 0.0001 \)), height (−2.5 cm; \( P < 0.0001 \)) and percentage of males (−4.8%; \( P = 0.0175 \)).

In the backward elimination procedure, baseline variables height, age, sex and the number of additional pituitary deficits were non-significant and eliminated. Baseline variables QoL-AGHDA, waist circumference, total cholesterol and IGF1 SDS remained in the model as predictors. The odds ratio (95% CI) per item QoL-AGHDA was 1.070 (CI 1.037–1.105); per centimetre waist circumference, it was 1.026 (CI 1.010–1.042); per mmol/l total cholesterol, 1.941 (CI 1.595–2.366); and per unit for IGF1 SDS, 0.874 (CI 0.765–0.992). Table 3 summarizes the model-building process.

The area under the ROC curve (AUC) for this model was 0.736 (Fig. 1A). For prediction of 2-year response, the following equation for the predictive response probability was established:

\[
P = 1 - \frac{1}{1 + e^{bX}}
\]

where \( bX \) was estimated to \(-5.2271 + 0.0681A + 0.0257W + 0.6640C - 0.1377I \) (final model 1) where \( A \) stands for total AGHDA score, \( W \) for waist circumference in centimetre, \( C \) for total cholesterol in mmol/l and \( I \) for IGF1 SDS in SDS units, all at baseline.) If a sum score of \( \geq 5 \) and \( \geq 9 \) instead of \( \geq 7 \) was used as a response cut-off, the AUC for model 1 were 0.712 and 0.705 instead of 0.736 respectively. With the cut-off of \( \geq 5 \) and \( \geq 9 \) instead of \( \geq 7 \), only two and three instead of four variables remained significant predictors respectively.

For model 2, we performed backward elimination on all significant baseline variables and 6-month delta waist circumference only (454 non-missing observation), which ended up with AGHDA score, cholesterol and waist circumference at baseline and 6-month delta waist circumference as significant predictors. The AUC for this model was 0.815 (Fig. 1B). The Pearson’s correlation coefficients between baseline values and delta changes for total cholesterol, QoL-AGHDA score and waist circumference were −0.43, −0.41 and −0.10 respectively. Thus, waist circumference showed the weakest correlation between baseline and delta change. This supports the use of waist circumference as a 6-month variable.

For prediction probabilities of 2-year response, the following equation was formulated based on baseline and 6 months delta information:

\[
P = 1 - \frac{1}{1 + e^{bX}}
\]

This supports the use of waist circumference as a 6-month variable.

Table 2 Baseline characteristics of all patients with complete baseline and 2-year delta variables.

<table>
<thead>
<tr>
<th></th>
<th>Men</th>
<th>Women</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>462</td>
<td>382</td>
</tr>
<tr>
<td>Age (years)</td>
<td>50.6±13.0</td>
<td>46.9±13.2</td>
</tr>
<tr>
<td>Waist circumference (cm)</td>
<td>100.7±12.4</td>
<td>93.5±14.8</td>
</tr>
<tr>
<td>Height (cm)</td>
<td>177.4±8.2</td>
<td>164.2±7.4</td>
</tr>
<tr>
<td>QoL-AGHDA score items (median (range))</td>
<td>9 (0–25)</td>
<td>11 (0–25)</td>
</tr>
<tr>
<td>Total cholesterol (mmol/l)</td>
<td>5.70±1.20</td>
<td>5.82±1.24</td>
</tr>
<tr>
<td>IGF1 SDS</td>
<td>−1.55±1.56</td>
<td>−1.97±1.77</td>
</tr>
</tbody>
</table>

Data are shown as means ± s.d. unless otherwise indicated.
where $bX$ was estimated to $-6.1032 + 0.0706A + 0.0285W - 0.8391C - 0.1800D$ (where $A$ stands for total AGHDA score, $W$ for waist circumference in centimetre, $C$ for total cholesterol in mmol/l, all at baseline, and $D$ for 6-month change in waist circumference.)

Performing the corresponding exercise for 6-month delta AGHDA score (295 non-missing observations) resulted in an equation with waist circumference and cholesterol at baseline and 6-month delta AGHDA score as significant predictors (parameter estimates not shown) with AUC of 0.784. Correspondingly for 6-month delta cholesterol (243 non-missing observations) the model resulted in an equation with AGHDA score and cholesterol at baseline and 6-month delta cholesterol (parameter estimates not shown) with an AUC of 0.778. In none of these models was baseline IGF1 SDS of importance. Therefore, as final model 2 we selected the model with baseline QoL AGHDA score, baseline waist circumference, baseline total cholesterol and delta waist circumference, as it had a high AUC and the highest number of non-missing observations.

If a sum score of $\geq 5$ and $\geq 9$ instead of $\geq 7$ was used as a response cut-off, the AUC for model 2 were 0.726 and 0.732 respectively. With the cut-off of $\geq 5$ and $\geq 9$ instead of instead of $\geq 7$, only one and three instead of four variables remained significant predictors respectively.

**Assessment of the models’ predictive probabilities**

Using the equations provided for final models 1 and 2 respectively we could calculate predicted response probabilities ranging between 0 and 1.

We used the whole data set for model building and did an internal cross-validation using the leave-one-out approach. With this approach, we examined how well we could predict the observation that was left out for each fit, repeating this step until all observation were left out once. With this approach the percentage of outliers were 4.7 and 3.7% in model 1 and model 2 respectively. This was below the cut-off of 5% considered acceptable and showed a good validation.

By classifying patients in the datasets for model 1 and 2 respectively, we could define a patient as a predicted responder if the calculated predicted probability was equal to or greater than a chosen predicted probability. We compared predicted responder status with actual responder status after 2 years of treatment. Thereby we could calculate sensitivity, specificity, positive predictive value and negative predictive value respectively for each selected cut-off of the model predicted probabilities, which are shown in Fig. 2. Positive predictive value corresponds to the probability that the patient actually responded to the treatment (by having a score of at least 7 at the 2-year visit), given the patient was predicted as a responder. The negative predictive value corresponds to the probability that the patient actually did not respond (by having a score of 6 or less at the 2-year visit), given the patient was predicted as a non-responder. Sensitivity measures the probability that the patient was predicted as a responder, given the patient is a responder (by having a score of at least 7 at the 2-year visit). Specificity measures the probability that the patient was predicted as a non-responder given that the patient is a non-responder. For

**Table 3** Summary of backward elimination and final Model 1 based on 791 observations.

<table>
<thead>
<tr>
<th>Step</th>
<th>Effects removed</th>
<th>DF</th>
<th>Number in</th>
<th>Wald $\chi^2$</th>
<th>$P$ value</th>
<th>AUC</th>
<th>S.E.M.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Age $^2$</td>
<td>1</td>
<td>14</td>
<td>0.0173</td>
<td>0.8952</td>
<td>0.757</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Cholesterol $^2$ (mmol/l)</td>
<td>1</td>
<td>13</td>
<td>0.2128</td>
<td>0.6446</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Gender</td>
<td>1</td>
<td>12</td>
<td>0.3226</td>
<td>0.5700</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>No. of additional deficiencies $^2$</td>
<td>1</td>
<td>11</td>
<td>0.4735</td>
<td>0.4914</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>No. of additional deficiencies</td>
<td>1</td>
<td>10</td>
<td>0.0033</td>
<td>0.9545</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>IGF1 SDS $^2$</td>
<td>1</td>
<td>9</td>
<td>0.9992</td>
<td>0.3175</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Height $^2$</td>
<td>1</td>
<td>8</td>
<td>1.4095</td>
<td>0.2351</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Height (cm)</td>
<td>1</td>
<td>7</td>
<td>0.8138</td>
<td>0.3670</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Age (years)</td>
<td>1</td>
<td>6</td>
<td>0.9989</td>
<td>0.3176</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Waist $^2$</td>
<td>1</td>
<td>5</td>
<td>3.6020</td>
<td>0.0577</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>AGHDA score $^2$</td>
<td>1</td>
<td>4</td>
<td>3.7838</td>
<td>0.0518</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Interceptor</td>
<td>1</td>
<td>$-5.2051$</td>
<td>27.2477</td>
<td>$&lt;$0.0001</td>
<td>0.736</td>
<td></td>
</tr>
</tbody>
</table>

*Values of number in for final model are an estimate.
the interpretation of the curves in Fig. 2, one can note that the data set for model 1 had a total number of actual non-responders at the 2-year visit of 126 (16%) and model 2 had 63 (14%). From the figures we can read that sensitivity and specificity curves intersects at around 65% for model 1 and 72% for model 2 at a predicted response probability of around 0.85. Positive predictive value is relatively high for all predicted response probabilities, whereas the negative predictive value decreases rapidly with increasing cut-off of the predicted response probability. In both models, a predicted response probability of 0.9 was already associated with a specificity and positive predictive value of more than 90%. A predicted response probability of 0.7 was associated with a sensitivity of 90% but a negative predictive value of only 40%. A predictive response score of 0.5 was also associated with negative predictive values of around 90%.

**Outlier and performance analysis**

Assessment of goodness of fit of the models using Hosmer-Lemeshow tests showed relatively good agreement between observed and expected numbers (model 1: $\chi^2 (8) = 12.21, P = 0.14$; model 2: $\chi^2 (8) = 7.37, P = 0.50$). Observations with a delta deviance ($\Delta D$) value greater or equal to 3.84 (5% percentile for $\chi^2 (1)$) were defined as outliers. A percentage of 5% or less outliers is generally considered expected. Model 1 had 37 outliers (4.7% of observations) and model 2 had 17 outliers (3.7%). All outliers (in both models 1 and 2) were observed non-responders with high predicted probabilities of being a responder. In model 1, 18 of 37 outliers (49%) and 391 of 754 non-outliers (52%) were reported with adverse events with diagnosis date before or around the 2 year visit, respectively. For model 2, the corresponding numbers were 11/17 (65%) for outliers and 259/437 (60%) for non-outliers respectively. Therefore, the adverse event rates were similar among outliers and non-outliers. Also, the rates of serious adverse events were similar among outliers and non-outliers (data not shown). One outlier patient was prescribed a zero GH dose at 2 years and one patient had an off-treatment period during the first 2 years of treatment (0.1 years, or approximately 1 month).

**Discussion**

We propose a simple clinical response score based on three clinical parameters that are routinely assessed during follow-up of GHRT. We additionally established two simple prediction models that allow the prediction of clinical response after 2 years, defined by a sum score of delta change in total cholesterol, waist circumference and QoL-AGHDA. In the baseline model, four variables were significant predictors of clinical response. These included the three variables that also compose the response score, namely, cholesterol, QoL-AGHDA, waist circumference, and IGF1 SDS. Worse values on baseline variables (higher cholesterol, AGHDA scores or waist circumference and lower IGF1 SDS) were associated with higher response probabilities. The combined model (model 2) with baseline variables QoL-AGHDA, waist circumference and cholesterol and 6-month delta waist circumference predicted clinical response well.

In secondary analyses, we showed that using alternative cut-offs reduced the number of significant predictors and decreased the AUCs of the models. This shows that the chosen cut-off is not only clinically meaningful but also statistically the most valuable cut-off.

Both models showed a good goodness of fit with similar observed and expected values according to the Hosmer-Lemeshow test. The discriminatory power of model 1 was fair, with an AUC of >0.7, and the discriminatory power of model 2 was good, with an AUC of >0.9.
Both models were robust and based on relatively high numbers of non-missing values ($n = 791$ and $n = 454$ for models 1 and 2 respectively).

Effects of GHRT on cardiovascular risk markers, body composition and quality of life have been well documented in both clinical trials and observational studies (1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12). Additionally, in a previous analysis of the German KIMS data, we have shown that cardiovascular risk assessed by conventional risk scores such as the Framingham score are beneficially influenced by GHRT (10). However, to our knowledge, until now there have been no data combining different dimensions of GH effects such as body composition, lipid markers and quality of life into a unified marker of clinical response.

Our data confirm previous studies showing that those patients with worse baseline characteristics respond best to GHRT (2, 5, 8, 11). In a previous KIMS analysis by Feldt-Rasmussen et al., sex, age and baseline IGF1 were predictors of IGF1 response and clinical changes to GHRT (11). In another study, sex was also the best predictor of favourable change in body fat and IGF1 response (9). In these studies, the association of different clinical baseline variables with IGF1 response and changes in clinical outcome variables such as weight, BMI, cholesterol and quality of life or body fat were analyzed. Our study confirmed IGF1, but not age or sex, as baseline predictors of clinical response. Both the different statistical approaches and different outcome variables might account for this difference. In the study by Feldt-Rasmussen et al., the main outcome variable was change in IGF1 SDS, and clinical outcomes were only studied in univariate analyses.

We analyzed combined clinical response variables using multiple logistic regression methods. Moreover, we intentionally did not include changes in IGF1 into our response score, as our aim was to establish a pure clinical response model, based on easy-to-measure characteristics of the patient. Also, GH doses are generally titrated according to IGF1 response. This might have additionally blunted the sex differences, as women were likely treated with higher GH doses in our cohort.

We used a model with a dichotomized response. We are aware that this approach may cause some loss of information. However, we think that this approach is most useful for the purpose of our study, as we aimed for a simple and clinically useful response model.

Our study extends previous work by proposing a unified clinical response score and by providing two simple prediction models for clinical response: a baseline model that can be helpful for the decision to initiate GH treatment (model 1) and a combined model that helps to
decide on continuation of GH treatment 6 months after GHRT start (model 2).

The results provided in this study might be useful for future clinical research and clinical practice in several aspects. The clinical response score was established on a large data set of 1417–2110 patients. This score can be a helpful tool in assessing clinical response in clinical studies on the effect of GHRT. It can also be used in more detailed future studies on response by subgroups such as sex, genetic background, baseline IGF1 levels, GH peak or others.

The prediction models established here allow predicting individual 2-year response both by baseline characteristics and 6-month changes using only a small number of variables. The combined prediction model with baseline and 6-months data showed a particularly good predictive ability and is easy to use in every-day clinical practice, as it only requires the routine measurements on waist circumference, cholesterol, the common and simple quality-of-life tool QoL-AGHDA and 6-months changes in waist circumference. It is of note that our prediction models should only be used after the other possible deficiencies are adequately replaced.

In clinical practice, for model 1, the four baseline prediction parameters waist circumference, QoL-AGHDA score, total cholesterol and IGF1 can be inputted in the regression equation provided. The regression equation yields a probability value between 0 and 1. Correspondingly, model 2 can be used after 6 months GH replacement. A probability value of 0.9 or higher indicates that the patient is highly likely to respond to treatment, whereas a probability of 0.5 or lower indicates that the patient is unlikely to respond to treatment.

Some limitations of our study need addressing. We do not know if our data are generalizable to patients with Cushing's disease, craniopharyngioma or previous acromegaly, as they were excluded from our study. Also, patients from the UK were excluded, as QoL-AGHDA data from UK might be biased since prescription depends on baseline QoL-AGHDA score. Therefore, we do not know if our data are generalizable to patients from the UK as well as from other countries that did not contribute to KIMS.

As in most observational studies, the number of missing values was high. We cannot exclude a potential exclusion bias. However, we expect this to be marginal, as only height, sex distribution and AGHDA score differed marginally. There were no differences in the other baseline values and in the 2-year changes of the response score variables.

Quality of life was measured with the QoL-AGHDA questionnaire. There are several other additional measurement tools of health-related quality of life. We do not know if our data can be generalized to cohorts with different measures of quality of life. In general, we do not know if our data can be generalized to populations treated with other GH formulations. However, we think that large differences are unlikely, as the effects of all GH formulations can be expected to be comparable. Also, possible changes in eating habits, physical activities and socio-economics status in the patients during the follow up period might affect accuracy of our prediction model.

Our data were not validated in an external cohort. However, using the leave-one-out method, we had less than 5% acceptable outliers and thus a good validation. In both models, the number of outliers was in the acceptable range below 5%. However, all outliers tended into one direction, being non-responders with high predicted response probabilities. The number of adverse events was similar among outliers and non-outliers. Therefore it is unlikely that this was influential. Additional factors such as non-adherence may also play a part.

In summary, we have proposed a clinical response score that allows individual clinical monitoring based on simple clinically meaningful variables. Moreover, we have established two simple clinical prediction models that allow predicting response in an individualized approach. This can be directly useful in helping to decide whether to initiate and whether to continue GH therapy.

Declaration of interest
H J Schneider received an independent research grant and statistical support for this study. During the work on this study, H J Schneider, M Buchfelder, H Wallaschofski and Peter H Kann were members of the German KIMS board, M Buchfelder, A Luguer and G Johannsson were members of the KIMS strategic advisory board, and A Mattsson is an employee of Pfizer, Inc., Sweden. A Luguer has received honoraria for presentations and/or participating in advisory boards from Pfizer, Ipsen, Novo Nordisk and Merck/Serono.

Funding
This study was supported by an independent research grant and by statistical support from Pfizer.

Acknowledgements
We thank all KIMS recruiting centres for contributing their data and all KIMS patients for allowing their data to be used. KIMS was funded by Pfizer, Inc.
References


