0013-7227/03/$15.00/0
Printed in U.S.A.

The Journal of Clinical Endocrinology & Metabolism 88(3):1223–1227

doi: 10.1210/jc.2002-021244

Validation and Calibration of the Kabi Pharmacia International Growth Study Prediction Model for Children with Idiopathic Growth Hormone Deficiency

MARIA A. J. DE RIDDER, THEO STIJNEN, AND ANITA C. S. HOKKEN-KOELEGA

Dutch Growth Foundation (M.A.J.d.R., A.C.S.H.-K.); and Department of Epidemiology and Biostatistics (M.A.J.d.R., T.S.), and Department of Pediatrics, Division of Endocrinology, Sophia Children's Hospital (A.C.S.H.-K.), Erasmus MC-University Medical Center Rotterdam, 3001 KB Rotterdam, The Netherlands

In 1999 a model was published for prediction of growth in children with idiopathic GH deficiency (IGHD) during GH therapy, derived using data from the Kabi Pharmacia International Growth Study (KIGS) database (Pharmacia & Upjohn, Inc., International Growth Database). We validated and calibrated this KIGS model for growth in the first year of GH therapy using data from 136 Dutch children with IGHD. Observed *vs.* predicted outcomes were plotted, and the fitted regression line was significantly different from the line of identity ($P = 0.03$). It appeared that the predictions were too extreme: relatively low predictions were too low, relatively high predictions were too high. This is a well known phenomenon in the context of prediction models, called overoptimism. For valid application to other data the KIGS predictions should be calibrated. Calibrated predictions are obtained using $Y_{cal} = Y_{orig} + (2.153 - 0.192 \times Y_{orig})$, where Y_{cal} is the calibrated prediction, and Y_{orig} is the KIGS prediction. The calibrated prediction will be higher than the original KIGS prediction when the original prediction is less than 11.2 cm/yr and will be lower otherwise. The variability of the prediction errors of the calibrated predictions was positively related to the value of the prediction ($P < 0.001$), described by the equation $SD_{pred\ err} = -1.017 + 0.286 \times Y_{cal}$. Our calibrated model will give better predictions for children with IGHD fulfilling the same criteria. (*J Clin Endocrinol Metab* 88: 1223–1227, 2003)

BECAUSE OF THE variability of response to GH treatment during the last decade several prediction models have been developed (1–4). It is well known that prediction models often perform less well than expected when applied to new patients, either of the same population or other populations. According to recent statistical insights this is in many instances due to the problem of overoptimism, which results in predictions that are too extreme (5, 6). This overoptimism, also called overfitting, is more pronounced if the prediction model is developed by selecting the predictor variables from a group of possible candidate predictors, and this selection is determined by the data. If a representative dataset is used, and a proper modeling method is applied, there is practically no doubt that in the final model the selected variables will be related to the outcome in that as well as in other datasets fulfilling the same criteria. However, the selection of exactly these predictor variables and exactly these values of the estimated coefficients of the predictors will partly be determined by the accidental characteristics of the dataset. This will give a model that is too much data-driven and will differentiate the subjects in predicted outcome too extremely. Even apart from this effect of variable selection, overfitting is likely to occur in each prediction model with two or more prediction variables (7). For valid application of a prediction model to new patients the predictions should therefore be calibrated by shrinking them to less extreme values.

In this study we used data from Dutch children with idiopathic GH deficiency (IGHD) to validate the prediction model derived on the database of the Kabi Pharmacia International Growth Study (KIGS; Pharmacia & Upjohn, Inc., International Growth Database) (8) for growth velocity of children with IGHD during the first year of GH therapy (2). This model, further referred to as the KIGS model, was developed on a dataset containing data from 593 children and is widely used in clinical practice. A validation of the model, using data from subsequent children from the KIGS database (temporal validation) as well as data from other studies (external validation), was described previously (2). Our study validated the model on a large external dataset with special attention to the problem of overfitting.

Abbreviations: HV, Height velocity; IGHD, idiopathic GH deficiency; PI, prediction interval.

Patients and Methods

Patients

For validation of the model we used data from the database of the Dutch Growth Foundation, containing data for all Dutch children who have been or are being treated with GH. We applied the same inclusion and exclusion criteria as were used for the cohort from which the KIGS model for children with IGHD was derived, although other brands of biosynthetic GH than Genotropin (Pharmacia & Upjohn Inc., Stockholm, Sweden) were used also. Height measurements for calculation of first year height velocity (HV) were allowed to have been performed between 0.8–1.25 yr after the start of treatment, comparable with the validation described in the paper in which the KIGS model was presented (2).

Observed and predicted outcomes

HV was computed correcting for the time interval between the start of treatment and the actual date of the height measurement after 1 yr of treatment. For the predictors used in the model, the SD scores were computed with the same reference data as those used for the KIGS cohort (9–11). The KIGS prediction was computed using the regression equa-

tion based on data at the start of treatment: predicted HV (cm/yr) = 14.55 − 1.37 × maximum GH response (ln; micrograms per liter) − 0.32 × age (yr) + 0.32 × birth weight SD score + 1.62 × GH dose (ln; international units per kilogram per week) − 0.40 × (height SD score − midparental height SD score) + 0.29 × weight SD score.

Statistical analysis

For the validation, we followed the method described by Van Houwelingen (12). The observed HV was plotted *vs.* the predicted HV in a calibration plot. A regression analysis was performed with observed HV as the outcome and predicted HV as the determinant, and it was tested whether the regression line was significantly different from the line of identity (where observed HV is equal to predicted HV). The estimated coefficients from the regression analysis were then used to determine the calibration correction. Using the calibrated predictions, we tested the homogeneity of variance of the prediction errors using the Spearman rank correlation between the absolute values of the prediction errors and the values of the predictions (13). In case of a significant correlation, the relation was modeled (14). In searching for the regression model for absolute residuals depending on predicted values, we allowed for fractional polynomials (15).

Results

One hundred and thirty-six cases of the Dutch database fulfilled the inclusion and exclusion criteria. The characteristics of this Dutch cohort were well within the ranges of the KIGS cohort (Table 1). The prediction errors (observed HV minus predicted HV) had a mean of 0.25 cm/yr and an SD of 1.95 cm/yr. The observed *vs.* predicted values are plotted in Fig. 1 together with the line of identity (where observed HV is equal to predicted HV; *dotted line*) and the fitted regression line. The figure shows that relatively low predictions tended to be underestimated, and relatively high predictions tended to be overestimated. The deviation of the regression line from the line of identity was statistically significant ($P = 0.03$), and the estimated slope was 0.808. To get a fitted regression line with a slope of 1 (line of identity), which is preferred, points at the left, relatively low predictions, should be shifted a little to the right, and points at the right, relatively high predictions, should be shifted a little to the left. This calibration correction is described by the formula:

$$Y_{cal} = Y_{orig} + 2.153 - 0.192 \cdot Y_{orig}$$

where Y_{cal} is the calibrated prediction, and Y_{orig} is the original KIGS prediction.

The correction term, $(2.153 - 0.192 \times Y_{orig})$, has a positive value for predictions below 11.2 cm/yr (point of intersection of the regression line and the line of identity) and a negative value for predictions above 11.2 cm/yr.

Using the calibrated model the prediction errors had a mean of zero and an SD of 1.91 cm/yr. When prediction errors were plotted *vs.* calibrated predictions (Fig. 2), an increase in variance of the prediction errors was seen with increasing value of the prediction. The absolute values of the prediction errors turned out to be positively related to the value of the predictions ($P < 0.001$). The relation between the calibrated predictions and the variability of their prediction errors was described with the model:

$$SD_{pred\ error} = -1.017 + 0.286 \cdot Y_{cal}$$

The differences between the predictions of the original KIGS model and the calibrated model are illustrated by two examples (Table 2). Example 1 shows that for a child with a relatively low prediction (predicted HV using the KIGS model, 7.0 cm/yr), the calibrated prediction is higher (7.8 cm/yr), whereas the 95% prediction interval (PI) is smaller than the PI using the KIGS model. Example 2 demonstrates that a relatively high prediction decreases after

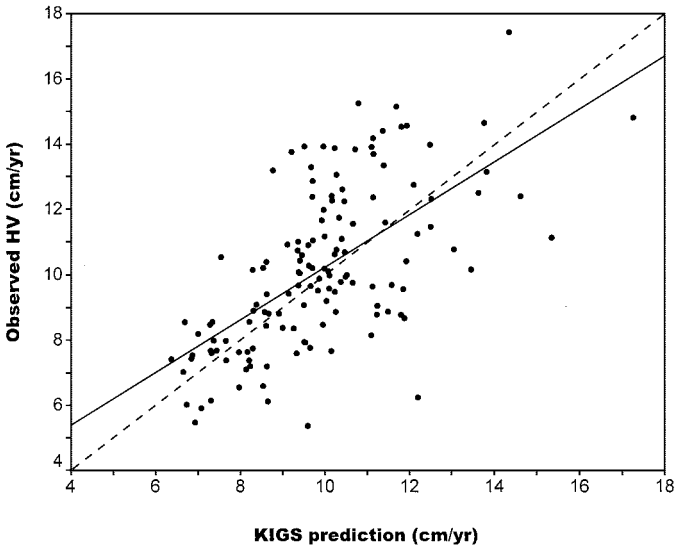

FIG. 1. First year HV (centimeters per year): observed *vs.* KIGS predicted values, together with the line of identity (*dotted*) and the fitted regression line (*solid*).

TABLE 1. Characteristics of children with IGHD from the KIGS cohort and the Dutch cohort

	KIGS cohort			Dutch cohort		
	Mean	Min	Max	Mean	Min	Max
Age (yr)	7.3	2.1	11.9	6.8	2.0	11.9
Height SD score	−2.6	−5.4	−0.4	−2.1	−5.3	−0.3
Weight SD score	−2.2	−3.8	−0.7	−1.9	−4.6	0.8
MPH SD score	−0.6	−3.4	2.3	0.4	−2.1	3.4
Height SD score − MPH SD score	−1.9	−6.2	1.7	−2.5	−6.6	0.0
BW SD score	−0.5	−2.0	3.6	−0.3	−1.9	3.1
Max GH-peak (μg/liter)	5.6	0.3	10.0	5.2	0.5	10.0
GH dose (IU/kg·wk)	0.6	0.2	1.3	0.6	0.3	1.3
HV in first year (cm/yr)	9.2	3.5	16.8	10.2	5.4	17.4

MPH, Midparental height; BW, birth weight.

calibration (KIGS prediction, 14.0 cm/yr; calibrated prediction, 13.5 cm/yr), whereas the 95% PI becomes much wider.

For the calculation of SD scores of the auxological characteristics used in the KIGS model, United Kingdom reference data (9, 10) were used. When for our validation in the Dutch cohort, national reference data were used [namely for height and weight at the start of treatment, the reference data from the 1997 study by Fredriks *et al.* (16) and for height of the parents the reference data from the 1965 study by Van Wieringen *et al*. (17)], the mean of the prediction errors was 0.23 cm/yr. The overfitting was again significant ($P < 0.001$), and the estimated slope in the calibration plot was 0.791. If the 1997 reference data were also used for height of the parents, the estimated slope was similar, but the mean prediction error increased to 0.60 cm/yr and was significantly different from zero ($P = 0.001$).

Discussion

In this paper for the first time a validation according to modern statistical insights into the behavior of prediction models was applied to a widely used prediction model for first year response to GH therapy. Special attention was given to the problem of overfitting. The analysis resulted in a calibrated model with a better fit when applied to new data.

The method of validation we have used is based on and illustrated by a calibration plot where observed values are plotted *vs.* predicted values, in our case observed HV in the first year of GH treatment *vs.* the prediction given by the KIGS model. Four examples of calibration plots are shown in Fig. 3. In all figures the line of identity, where the observed outcome is equal to the predicted outcome, is drawn. Perfect predictions will all lay on the line of identity (Fig. 3A). If the predictions are not perfect, but unbiased, the plot will show points scattered randomly around the line of identity (Fig. 3B). Any other pattern in the calibration plot indicates that the prediction model does not fit well to the validation data. As an example, Fig. 3C shows a scatter plot with points too much on the right side of the line of identity. Because the vertical position of the points is fixed, determined by the observed value, the horizontal position should be corrected by shifting all points a little to the left. This indicates that the original predictions are, on the average, too high. Another, frequently occurring pattern is that the scatter plot shows a relation between the observed and predicted outcomes with a lower slope than the line of identity (Fig. 3D). Now a correction should be made by shifting the points at the left side a little to the right, and the points at the right side a little to the left. This indicates underestimation of the lowest predictions and overestimation of the highest predictions. The model classifies too extremely into good and bad responses, and therefore this phenomenon is called overfitting or overoptimism. Overfitting is common when the selection of predictors and the estimation of their coefficients are both guided by the same dataset (18, 19). The occurrence of overfitting is not the result of differences between the modeling group and the validation group. Suppose we start with one dataset and split this randomly into a modeling group and a validation group, thus without systematic differences between the two groups. If we develop a prediction model on the modeling group, applying the model to the validation group will very likely show overfitting too.

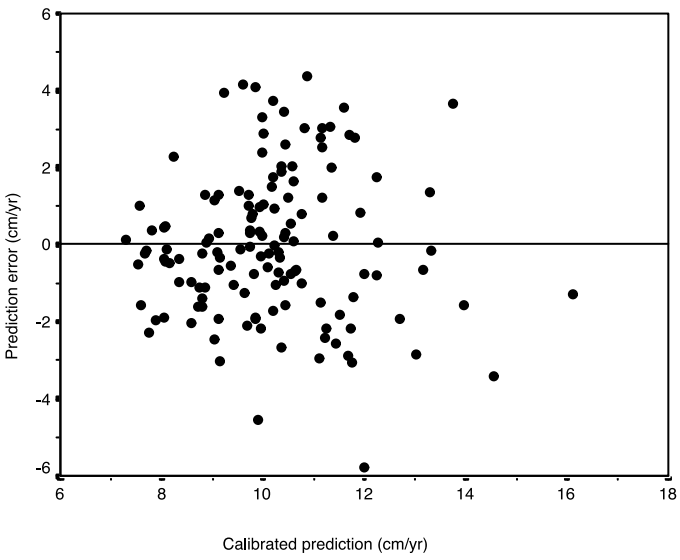

FIG. 2. Prediction errors *vs.* predicted value of the calibrated model.

In our validation of the KIGS model the calibration plot showed a pattern like that in Fig. 3D, although less extreme. The slope of the regression line was 0.808, and the line was significantly different from the line of identity. The overfitting could be corrected using the formula: $Y_{cal} = Y_{orig} + (2.153 - 0.192 \times Y_{orig})$. This leads to a correction of the predictions from the KIGS model ranging from +0.92 cm/yr for a prediction of 6.4 cm/yr (the lowest KIGS prediction in our validation group) to −1.2 cm/yr when 17.3 cm/yr was predicted (highest predicted value). Corrections for predictions of in-between values will be smaller, and for a predicted HV of 11.2 cm/yr the correction term is zero.

The validation performed on the KIGS model as described in the report by Ranke *et al.* (2), consisting of testing whether there was a significant difference between observed and predicted values, is only a check for a systematic prediction error (overall too low or too high predictions) and will not disclose overfitting.

In the Dutch cohort the SD of the prediction errors of the calibrated model was higher than the published SD of 1.46 cm/yr of the prediction errors found in the KIGS cohort (2). This was to be expected because a prediction model will be less precise for new data than for the data on which it was derived. Another important finding was the dependency of

TABLE 2. Examples of predictions and 95% PIs using the KIGS model and using the calibrated model

	Original KIGS model			Calibrated model		
	Prediction (cm/yr)	95% PI	Width of PI	Prediction (cm/yr)	95% PI	Width of PI
Example 1	7.0	4.1–9.9	5.7	7.8	5.4–10.2	4.8
Example 2	14.0	11.1–16.9	5.7	13.5	7.9–19.0	11.1

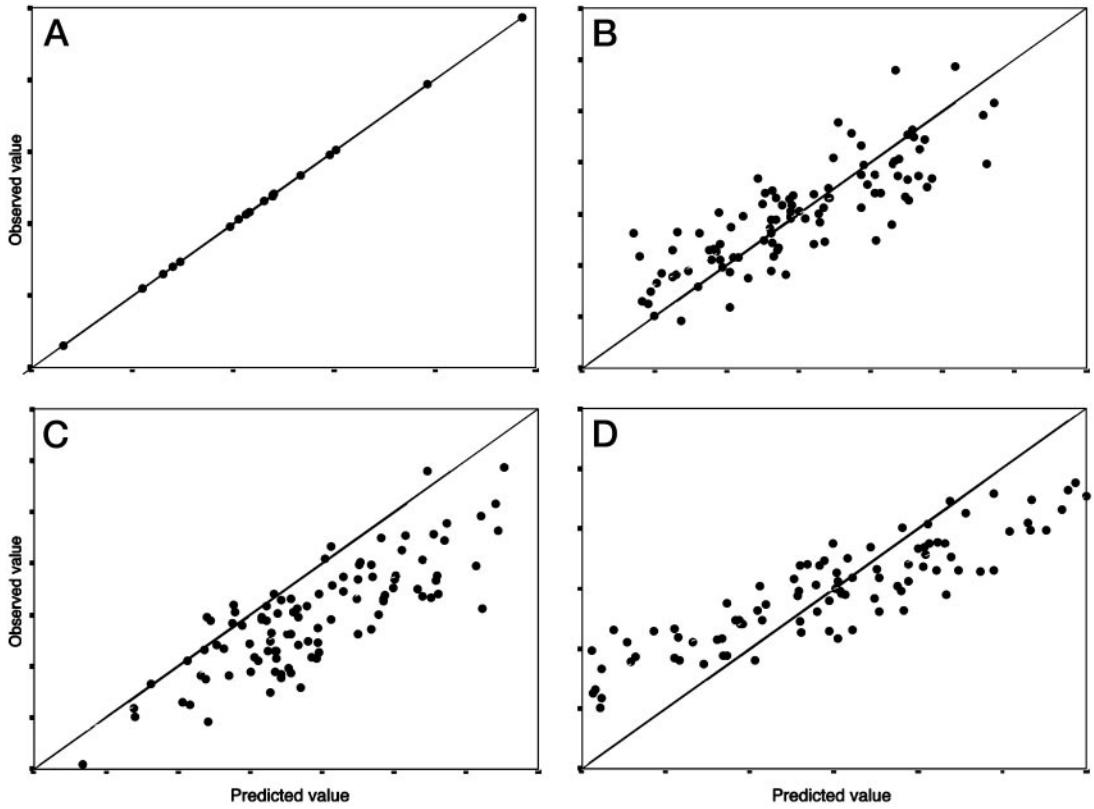

FIG. 3. Examples of calibration plots of prediction models for which the predictions are perfect (A), unbiased (B), too high (C), and overfitted (D).

variance of the prediction error on the magnitude of the prediction. We found that the prediction error SD was ranging from 1.07 cm/yr for a prediction of 7.3 cm/yr (our lowest calibrated prediction) to 3.58 cm/yr when the predicted outcome was 16.1 cm/yr (highest calibrated prediction). This means that the accuracy of a high prediction of HV is less than the accuracy of a low prediction, as reflected in the 95% PI accompanying the predicted response.

When developing the KIGS model, the analysts also checked for a relation between the variance of the prediction error and the prediction, by plotting Studentized residuals *vs.* predicted HV (2). In contrast to our findings, they did not find a relation. We cannot explain the discrepancy between their and our findings.

For calculation of the SD scores for the KIGS cohort, United Kingdom reference data (9, 10) were used. Dutch children and adults are known to be among the tallest on earth (20). Moreover, the secular trend in The Netherlands is relatively large (10, 16). If Dutch reference data were used, for parents dated back one generation, the predictions were not much different from the original predictions, but if we calculated the SD scores ignoring the secular trend, the predictions were significantly too low. In both situations the overfitting remained. Thus, to apply the KIGS model correctly, one should use the United Kingdom reference data.

The KIGS model is a well defined and easily applicable model developed on a population of sufficient size. The percentage of explained variability ($r^2 = 0.61$) might improve if more flexibility of the relations was allowed or other patient characteristics could be used, but the aim might have been to restrict to simple modeling with widely available characteristics. However, the problem of overfitting, which is very likely to be present if model selection and estimation are made using the same dataset, was not taken into account.

As a consequence of our calibration and of modeling the dependency of the prediction error SD, we propose that a KIGS prediction for the first year growth response in children with IGHD should be modified using the formula:

$$Y_{cal} = Y_{orig} + 2.153 - 0.192 \cdot Y_{orig}$$

where Y_{cal} is the calibrated prediction, and Y_{orig} is the original KIGS prediction.

The 95% PI is given by:

$$Y_{low} = Y_{cal} - 1.96 \cdot (-1.017 + 0.286 \cdot Y_{cal})$$

$$Y_{high} = Y_{cal} + 1.96 \cdot (-1.017 + 0.286 \cdot Y_{cal})$$

where again Y_{cal} is the calibrated prediction, Y_{low} is the lower limit of the PI, and Y_{high} is the upper limit of the PI. Figure 4 presents the calibrated predictions *vs.* the original KIGS predictions, with 95% PI. It shows that, for instance, an original prediction of 14 cm/yr should be modified to 13.5 cm/yr, with 95% PI of 8–19 cm/yr.

Our validation and calibration of the KIGS model are quite different from developing a new model. In our calibrated model we maintained the relative contributions of the predictor variables as determined for the original model, but the outcomes were adjusted using a correction term that depends on the value of the original prediction. Clearly these

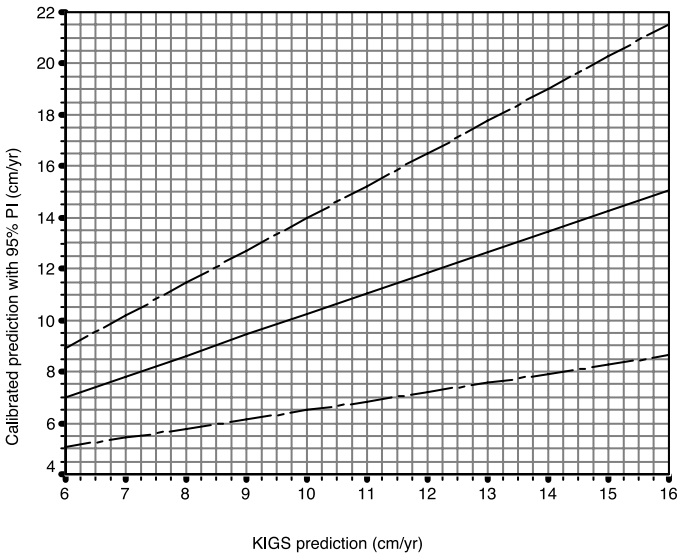

FIG. 4. Relation between the calibrated prediction and the value of the KIGS prediction, with 95% prediction interval.

calibrated predictions fitted better to the data of the Dutch cohort. Whether both the data from the KIGS cohort, used for modeling, as well as the data from the Dutch cohort, used for calibration, are representative of other cohorts of children with IGHD fulfilling the same in- and exclusion criteria is not yet known. We, however, postulate that our modification of the KIGS prediction rule will give better predictions for these children as well.

When new prediction models are developed in the future, they should be evaluated as to whether a calibration is required. The best way to examine this is by validation of the model using external data. This will improve the accuracy and benefit of prediction models for clinical practice.

Acknowledgments

Received August 6, 2002. Accepted December 2, 2002.

Address all correspondence and requests for reprints to: M. de Ridder, M.Sc., Dutch Growth Foundation, P.O. Box 23068, 3001 KB Rotterdam, The Netherlands. E-mail: deridder@epib.fgg.eur.nl.

References

1. **Ranke MB, Guilbaud O, Lindberg A, Cole T** 1993 Prediction of the growth response in children with various growth disorders treated with growth hormone: analyses of data from the Kabi Pharmacia International Growth Study. International Board of the Kabi Pharmacia International Growth Study. Acta Paediatr Suppl 82(Suppl 391):82–88
2. **Ranke MB, Lindberg A, Chatelain P, Wilton P, Cutfield W, Albertsson-Wikland K, Price DA** 1999 Derivation and validation of a mathematical model for predicting the response to exogenous recombinant human growth hormone (GH) in prepubertal children with idiopathic GH deficiency. KIGS International Board. Kabi Pharmacia International Growth Study. J Clin Endocrinol Metab 84:1174–1183
3. **Albertsson-Wikland K, Kriström B, Rosberg S, Svensson B, Nierop AFM** 2000 Validated multivariate models predicting the growth response to GH treatment in individual short children with a broad range in GH secretion capacities. Pediatr Res 48:475–484
4. **Schönau E, Westermann F, Rauch F, Stabrey A, Wassmer G, Keller E, Brämswig J, Blum WF** 2001 A new and accurate prediction model for growth response to growth hormone treatment in children with growth hormone deficiency. Eur J Endocrinol 144:13–20
5. **Van Houwelingen JC, Le Cessie S** 1990 Predictive value of statistical models. Stat Med 9:1303–1325
6. **Harrell FEJ, Lee KL, Mark DB** 1996 Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 15:361–387
7. **Harrell FEJ** 2001 Regression modeling strategies. New York: Springer-Verlag
8. **Wallström A** 1999 KIGS: structure and organisation. In: Ranke MB, Wilton P, eds. Growth hormone therapy in KIGS: 10 years' experience. Heidelberg: Barth; 1–9
9. **Tanner JM, Whitehouse RH, Takaishi M** 1966 Standards from birth to maturity for height, weight, height velocity, and weight velocity: British children, 1965. II. Arch Dis Child 41:613–635
10. **Freeman JV, Cole TJ, Chinn S, Jones PR, White EM, Preece MA** 1995 Cross sectional stature and weight reference curves for the UK, 1990. Arch Dis Child 73:17–24
11. **Niklasson A, Ericson A, Fryer JG, Karlberg J, Lawrence C, Karlberg P** 1991 An update of the Swedish reference standards for weight, length and head circumference at birth for given gestational age (1977–1981). Acta Paediatr Scand 80:756–762
12. **Van Houwelingen JC** 2000 Validation, calibration, revision and combination of prognostic survival models. Stat Med 19:3401–3415
13. **Kleinbaum DG, Kupper LL, Muller KE, Nizam A** 1998 Applied regression analysis and other multivariable methods. Pacific Grove: Duxbury Press
14. **Altman DG** 1993 Construction of age-related reference centiles using absolute residuals. Stat Med 12:917–924
15. **Royston P, Altman DG** 1994 Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. Appl Stat 43: 429–467
16. **Fredriks AM, Van Buuren S, Burgmeijer RJF, Meulmeester JF, Beuker RJ, Brugman E, Roede MJ, Verloove-Vanhorick SP, Wit JM** 2000 Continuing positive secular growth change in The Netherlands 1955–1997. Pediatr Res 47:316–323
17. **Van Wieringen JC, Wafelbakker F, Verbrugge HP, De Haas JH** 1971 Growth diagrams 1965 Netherlands. Leiden, Groningen: Nederlands Instituut voor Praeventieve Geneeskunde/Wolters-Noordhoff; 1–69
18. **Altman DG, Royston P** 2000 What do we mean by validating a prognostic model? Stat Med 19:453–473
19. **Chatfield C** 1995 Model uncertainty, data mining and statistical Inference. J R Stat Soc A 158:419–466
20. **Roede MJ, Van Wieringen JC** 1985 Growth diagrams 1980: Netherlands third nation-wide survey. Tijdschr Soc Gezondheidsz 63(Suppl):1–34