Enhancing a Somatic Maturity Prediction Model

SARAH A. MOORE[1,2], HEATHER A. MCKAY[1,2,3], HEATHER MACDONALD[1,2,4], LINDSAY NETTLEFOLD[2], ADAM D. G. BAXTER-JONES[5], NOËL CAMERON[6], and PENELOPE M. A. BRASHER[2,7,8]

[1]Department of Orthopaedics, Faculty of Medicine, University of British Columbia, Vancouver, British Columbia, CANADA; [2]Centre for Hip Health and Mobility, Vancouver Coastal Health Research Institute, Vancouver, British Columbia, CANADA; [3]Department of Family Medicine, Faculty of Medicine, University of British Columbia, Vancouver, British Columbia, CANADA; [4]Child and Family Research Institute, BC Women's and Children's Hospital, Vancouver, British Columbia, CANADA; [5]College of Kinesiology, University of Saskatchewan, Saskatoon, Saskatchewan, CANADA; [6]Centre for Global Health and Human Development, Department of Sport, Exercise and Health Sciences, Loughborough University, Leicestershire, UNITED KINGDOM; [7]Centre for Clinical Epidemiology and Evaluation, Vancouver Coastal Health Research Institute, Vancouver, British Columbia, CANADA; and [8]Department of Statistics, Faculty of Science, University of British Columbia, Vancouver, British Columbia, CANADA

ABSTRACT

MOORE, S. A., H. A. MCKAY, H. MACDONALD, L. NETTLEFOLD, A. D. G. BAXTER-JONES, N. CAMERON, and P. M. A. BRASHER. Enhancing a Somatic Maturity Prediction Model. *Med. Sci. Sports Exerc.*, Vol. 47, No. 8, pp. 1755–1764, 2015. **Purpose**: Assessing biological maturity in studies of children is challenging. Sex-specific regression equations developed using anthropometric measures are widely used to predict somatic maturity. However, prediction accuracy was not established in external samples. Thus, we aimed to evaluate the fit of these equations, assess for overfitting (adjusting as necessary), and calibrate using external samples. **Methods**: We evaluated potential overfitting using the original Pediatric Bone Mineral Accrual Study (PBMAS; 79 boys and 72 girls; 7.5–17.5 yr). We assessed change in R^2 and standard error of the estimate (SEE) with the addition of predictor variables. We determined the effect of within-subject correlation using cluster-robust variance and fivefold random splitting followed by forward-stepwise regression. We used dominant predictors from these splits to assess predictive abilities of various models. We calibrated using participants from the Healthy Bones Study III (HBS-III; 42 boys and 39 girls; 8.9–18.9 yr) and Harpenden Growth Study (HGS; 38 boys and 32 girls; 6.5–19.1 yr). **Results**: Change in R^2 and SEE was negligible when later predictors were added during step-by-step refitting of the original equations, suggesting overfitting. After redevelopment, new models included age × sitting height for boys (R^2, 0.91; SEE, 0.51) and age × height for girls (R^2, 0.90; SEE, 0.52). These models calibrated well in external samples; HBS boys: b_0, 0.04 (0.05); b_1, 0.98 (0.03); RMSE, 0.89; HBS girls: b_0, 0.35 (0.04); b_1, 1.01 (0.02); RMSE, 0.65; HGS boys: b_0, −0.20 (0.02); b_1, 1.02 (0.01); RMSE, 0.85; HGS girls: b_0, −0.02 (0.03); b_1, 0.97 (0.02); RMSE, 0.70; where b_0 equals calibration intercept (standard error (SE)) and b_1 equals calibration slope (SE), and RMSE equals root mean squared error (of prediction). We subsequently developed an age × height alternate for boys, allowing for predictions without sitting height. **Conclusion**: Our equations provided good fits in external samples and provide an alternative to commonly used models. Original prediction equations were simplified with no meaningful increase in estimation error. **Key Words:** PEAK HEIGHT VELOCITY, MATURATION, ADOLESCENTS, GROWTH MODELING, CALIBRATION

The well-known variation in the tempo and timing of biological maturity for boys and girls of the same chronological age (35,36) necessitates the use of an accurate measure of maturation in research involving children and adolescents. The status and timing of biological maturity are commonly assessed using methods such as skeletal age, pubertal (e.g., Tanner) staging, age at menarche, percentage of adult height, and age at peak height velocity (APHV) (7,9,20). Given the concerns related to the invasiveness of some methods and the logistical challenges of others (7,9,20), simple noninvasive methods to assess biological maturity have been suggested (26).

The prediction of APHV using sex-specific regression equations offers one such noninvasive method (26). Estimation of actual APHV requires serial longitudinal data spanning the period from late childhood through adolescence, which is often unavailable. Thus, based on the known differential timings of growth in height, sitting height and leg length, Mirwald et al. (26) developed equations to predict years from APHV (maturity offset (MO)) in boys and girls from simple one-time anthropometric measures. These prediction equations have been well utilized, having been cited more than 250 times since 2002 (Web of Science, search

Address for correspondence: Heather A McKay, Ph.D., 775-2635 Laurel St, Vancouver, British Columbia, CANADA, V5Z 1M9; E-mail: heather.mckay@familymed.ubc.ca.

Submitted for publication July 2014.
Accepted for publication November 2014.
Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's Web site (www.acsm-msse.org).

0195-9131/15/4708-1755/0
MEDICINE & SCIENCE IN SPORTS & EXERCISE®

DOI: 10.1249/MSS.0000000000000588

date: September 12, 2014). However, results of three longitudinal studies published in the last decade highlight potential limitations associated with these equations. Specifically, in a 7-yr longitudinal study of 13 regional and national level female gymnasts age 6.0–17.6 yr, a systematic bias was reported such that predicted APHV was underestimated in gymnasts with a later observed APHV and overestimated in those with an earlier observed APHV (21). More recently, Malina and Koziel (23 evaluated the prediction equations and their errors in 391 Polish boys (22) and girls (23) age 8.0–18.0 yr from the Wroclaw Growth Study. In the boys, the prediction equations performed more variably in early- and late-maturing groups; mean differences were negative in late-maturing boys (i.e., prediction was early) and positive in early-maturing boys (i.e., prediction was late) (22). In the girls, the authors found that predicted MO and APHV were dependent on age at the time of prediction and performance was better when predicted before expected APHV (23). These findings were consistent with the error discussed in the original article (26). In all three studies by Malina and colleagues, APHV was determined using the Preece–Baines (P-B) model 1 curve-fitting procedure (29). Bland–Altman plots were then used to assess agreement between the criterion measure and predicted value. There are potential limitations associated with these approaches. First, there was a large age range in the sample. With any prediction, we would expect prediction to perform best nearer to the observed measure (e.g., closest to the time of actual APHV); this was reflected in all studies. Second, although the P-B model 1 provides a number of biologically meaningful parameters, it has been shown to underestimate APHV compared with raw height velocities or nonparametric modeling techniques (16). Third, when evaluating a prediction equation, we are interested in calibration and the possibility of recalibration to improve predictions in new populations (8). When evaluating the performance of a model, calibration refers to whether the predicted agrees with the observed measure (15,17). Calibration of the prediction equations was not investigated in any of these previous studies.

When we considered calibration of the predictive equations, we acknowledged the possibility of overfitting. If overfitting is present, the model could perform poorly in external samples. Overfitting occurs when some of the covariance included in the model are based on spurious associations and/or coefficients are artificially large. This results from fitting the sample rather than the population and can lead to poor predictive performance (38). Development of the original prediction equations was based on 659 and 599 observations for boys and girls, respectively. The authors considered 15 potential predictors in their sex-specific models. This ratio of observations to predictors (40:1) suggests that overfitting should not be a problem (34). However, observations were not independent, as there were multiple observations per child, which reduces the effective sample size. Not acknowledging a within-child association may have resulted in standard errors (SE) and *P* values that were too small; this may have led to spurious variables being included in the prediction models. Overfitting can be minimized using four strategies: 1) prespecifying well-motivated predictors, 2) eliminating predictors without using the outcome, 3) cross-validating the target measure of the predictive errors using the outcome, and 4) shrinking the coefficient estimates using the outcome (38).

Given the widespread use of Mirwald et al. (26) prediction equations, recent queries regarding the fits of these equations (21–23), and the continued interest to develop non-invasive methods to assess maturity, we aimed to: 1) explore the possibility of over-fitting in the original development of the MO prediction equations; 2) modify the regression equations (as needed) employing cluster-robust variance techniques, 3) internally validate our equations using *k*-fold cross-validation; 4) create alternative equations that do not require sitting height; and 5) externally validate our equations using a cohort of Canadian children from the longitudinal Healthy Bones Study III (HBS-III) and a cohort of British children from the Harpenden Growth Study (HGS).

METHODS

Study participants. Our study includes participants from three well-known longitudinal growth studies: the Pediatric Bone Mineral Accrual Study (PBMAS), the HBS-III, and the HGS. We briefly describe each cohort in this subsection.

The original prediction equations were developed using data from the University of Saskatchewan's PBMAS (1991–1997), a mixed longitudinal study of 251 Canadian boys (n = 115) and girls (n = 136) recruited from two elementary schools in Saskatoon, Saskatchewan, Canada, between 1991 and 1993 (2,3,5). The study was designed to assess factors associated with bone acquisition in growing children. Participants were 8.0–15.0 yr of age at baseline; ages ranged between 8.0 and 21.0 yr across the initial 7-yr of the study. All participants were white (Caucasian) as determined by parents' place of birth (by questionnaire). Investigators created a subset of the PBMAS cohort for whom had sufficient measures to calculate APHV (n = 151 [60%]; 79 boys and 72 girls; see APHV protocol). All children were healthy with no conditions known to affect growth. Growth parameters were measured semiannually. Written informed consent was obtained from the parents of the participating children between 1991 and 1993. The University of Saskatchewan's Research Ethics Board approved all procedures. Professor Adam Baxter-Jones (coauthor), University of Saskatchewan, kindly provided all data from the original paper (26) for our analyses.

Our research group conducted the University of British Columbia's HBS-III (1999–2012), a mixed longitudinal study of 1071 Canadian boys (n = 515) and girls (n = 556) recruited from elementary schools in Vancouver and Richmond, British Columbia, Canada, between 1999 and 2009 (18,19,28). The HBS-III study was designed to assess factors associated with bone strength accrual in growing children. Participants

were 8.8–12.4 yr of age at baseline; ages ranged from 8.8 to 23.2 yr across the 14-yr study. The HBS-III sample was multiethnic; participants were of Asian (n = 533), white (n = 412), or "other" or "mixed" (n = 126) ancestry. We determined each participant's ethnicity based on their parents' or grandparents' place of birth as reported on a health history questionnaire at baseline. We classified participants as white if both parents or three of four grandparents were born in North America or Europe, and Asian if both parents or three of four grandparents were born in Hong Kong, China, Japan, Taiwan, Philippines, Korea, or India. Owing to the known ethnic-specific variation in the timing and tempo of growth (12), we excluded participants of Asian and other or mixed ethnicity from this analysis. From the white subsample, we included only those for whom we could accurately calculate APHV (n = 81 [20%]; 42 boys and 39 girls; see APHV protocol). All children were healthy and had no conditions known to affect growth. We performed anthropometry protocols in our laboratory annually. Written informed consent was provided by the parents or legal guardians, written assent from participants younger than 18.0 yr, and written informed consent from participants older than 18.0 yr. The University of British Columbia Clinical Research Ethics Board approved all procedures.

The Harpenden Growth Study (HGS; 1948–1971) was a mixed longitudinal study of 701 British boys (n = 419) and girls (n = 282) recruited from a national children's home in Harpenden, UK. The participants were 1.0–20.0 yr of age at baseline; with a range of 1.0–35.1 yr of age across the 24-yr study. The HGS was the first longitudinal study of human growth in Europe. The Harpenden team measured participants semiannually until age 12.0 yr, quarterly from the initiation of puberty, annually until age 20, and quinquennially thereafter. All measures were within ±3 wk of the target measurement date. For the present study, we included boys and girls who were first measured before 9.0 and 7.0 yr of age, respectively. Thus, our sample included 366 participants (238 boys and 128 girls) age 1.0–9.0 yr at baseline and whose ages ranged from 1.0 to 30.7 yr across the study's duration. From this subsample, we selected only those for whom we could accurately calculate APHV (n = 70 [19%], 38 boys and 32 girls; see APHV protocol). All children were white and were healthy, with no conditions known to affect growth. The study was carried out before the requirement of ethics committees and procedures or participants' consent. Professor Noël Cameron (coauthor), Loughborough University, kindly provided all HGS data for these analyses.

Anthropometry. The two Canadian Studies (PBMAS and HBS-III) used identical anthopometric protocols. The PBMAS and HBS-III studies assessed height (cm) during sitting and standing using standard stretch stature techniques (32) against a wall-mounted stadiometer and recorded the value to the nearest 0.1 cm. The PBMAS and HBS-III studies assessed weight (kg) to the nearest 0.1 kg on a calibrated electronic scale with the participants dressed in light clothing. All measures were taken in duplicate by trained research assistants at the respective laboratories. If measurements differed by more than the accepted measurement error (4 mm), a third measure was taken. Final height and weight were the mean of two (or three) measures. Leg length (cm) was determined by subtracting sitting height from standing height.

The HGS team measured height (cm) and sitting height (cm) using the same stretch stature techniques as described previously for the Canadian studies. Leg length was also determined in the same manner as the Canadian studies. Finally, weight was measured with participants in the nude on a portable beam balance to the nearest 0.1 kg. One technician (RH Whitehouse) took all measurements for the duration of the study.

Observed age at peak height velocity and maturity offset. For the PBMAS cohort, APHV was provided by Professor Baxter-Jones and was determined by fitting an interpolating cubic spline to empirical data using Prism (version 5.0, GraphPad, San Diego, CA). Age at peak height velocity was calculated from quotients of annual velocities and age differences (2). We used the same data set as in the original publication (26). In the HBS-III cohort, we chose to calculate APHV using the same statistical method as PBMAS (interpolating cubic splines). Briefly, we calculated APHV for those participants with sufficient height measurements during the identified pubertal period (five measures for boys between 11.5 and 16.5 yr and four measures for girls between 11.0 and 13.0 yr). We calculated running annual velocities and fit an interpolating cubic spline on a regular grid to identify maximum height velocity (APHV) for each participant. We performed this analysis in Stata (version 10.1, StataCorp, College Station, TX). We visually inspected all plots and selected those that had clear peaks during the pubertal spurt as well as pre- and post-APHV data. We provide a complete description of the HBS-III APHV protocol in Supplemental Digital Content 1 (See document, Supplemental Digital Content 1, Age at Peak Height Velocity Protocol for the Healthy Bones Study III, http://links.lww.com/MSS/A476).

Similar to the HBS-III protocol, we calculated APHV for the HGS participants with sufficient height measures during the pubertal period (15 measures for boys between 10.5 and 16.5 yr and for girls between 8.5 and 14.5 yr). We calculated running annual velocities and fit an interpolating cubic spline on a regular grid to identify maximum height velocity (APHV). We visually inspected all plots and selected those that had the most complete data during the pubertal growth spurt and where pre- and post-APHV could be clearly identified.

Finally, for all samples, we used APHV to calculate a biological maturity offset (MO, yr) by subtracting APHV from chronological age at the time of measurement. Thus, we generated a continuous measure of biological age (e.g., a MO = −1 yr is equivalent to 1 yr before APHV). Although APHV is our benchmark, we describe each measure as years

from APHV; where the difference is defined as MO (26). As MO allows for comparisons of somatic maturity between sexes, we considered the prediction of MO an advantage compared to that of APHV. Age at peak height velocity can easily be calculated after predicting MO.

Statistical analyses. We performed all statistical analyses in Stata (version 10.1, StataCorp, College Station, TX). We inspected data for potential errors and missing data; we cleaned and prepared all data for analyses. We removed any duplicate data. In the event of within-subject negative changes in growth parameters (<1% of observations), we checked the data and corrected with linear interpolation if a change in height was negative or assigned the previous year's value if we determined that adult height had been attained.

To determine whether overfitting occurred during development of original sex-specific prediction models, we first refit and evaluated the original equations. The original equations were developed using multivariable regression with hierarchical entry. Fifteen independent predictors (age, height, sitting height, leg length, weight, age × height, age × sitting height, age × leg length, age × weight, leg length × sitting height, weight by height × 100, body mass index (BMI; weight divided by height squared), sitting height by height × 100, leg length by height × 100, and leg length by sitting height × 100) were identified. As authors did not report the order in which predictors were entered, we reviewed the four models presented by Mirwald et al. (26) to surmise the most likely hierarchical order. To assess the potential for overfitting, we calculated change in R^2 and SEE with the addition of each predictor that Mirwald et al. (26) included in the original model.

We confirmed that overfitting was present after our step-by-step refitting of the original models. Thus, we proceeded to redevelop and create more parsimonious models. First, we reduced the number of potential predictors by examining correlations among predictors (15); when correlations were 0.99 or greater, we excluded the variable from the model that demonstrated the greatest measurement variability. Second, we addressed within-subject correlations by using a cluster-robust variance estimator in the regression model (30,39); this estimator is available in Stata via the vce (cluster id) option. We used a forward stepwise procedure to select other variables to include in the regression model (analysis 1) from the variables that remained. We recognize the limitations associated with automatic variable selection procedures so we performed an analysis akin to *k*-fold cross-validation (15). We randomly split the data set into seven subsets for boys and six subsets for girls, ensuring all subsets included 70 or more observations. We included only one observation for each individual. In each subset, we used forward stepwise regression to select variables. We repeated the random split five times (5 × 7- and 5 × 6-fold random splitting in boys and girls, respectively). We tabulated the number of times each predictor was selected for inclusion in the model (analysis 2). We used the variables that entered in analysis 1 and others consistently identified as important in the 35 samples for boys and 30 samples for girls to develop final sex-specific prediction equations using a forward step-by-step entry approach. Finally, with the interest of creating alternate equations not requiring sitting height, we fit a separate model (age × height) for boys.

We validated the new prediction models using the HBS-III and HGS data. We produced calibration plots (observed vs predicted values), calculated the coefficients of the calibration curve, and produced a descriptive and graphical summary of the differences between the observed and predicted values.

RESULTS

We provide a summary of the number of participants, measures, and a comparison between APHV (calculated by interpolating cubic spline as previously described) for boys and girls in the three studies (Table 1). In the boys, APHV did not differ between the HBS-III and PBMAS cohorts. However, APHV was significantly later in the HGS boys compared with the PBMAS boys (mean difference was 0.64 ± 0.05 yr [95% CI, 0.55–0.74]). In the girls, APHV was not significantly different between the three cohorts.

To assess the potential for overfitting of the original model, we provide the incremental change in R^2 and SEE, with the addition of each predictor variable to the sex-specific regression equations (Table 2). The original equations (26) had four predictors for both boys and girls. In the boys, the predictors were leg length × sitting height, age × leg length, age × sitting height, and leg length by height × 100. In the girls, the predictors were leg length × sitting height, age × sitting height, leg length by height ×100, and age × weight. In the boys, after the first three predictors were entered, the change in R^2 and the SEE was negligible (<1%). Similarly, in the girls, the change in R^2 and the SEE was negligible (<1%) after the first two predictors were entered. These small changes suggest that overfitting is present in the original equations and redevelopment of the regression equations may be useful.

In redeveloping the equations, we first assessed the original 15 predictor variables (26) and excluded variables where statistically and biologically appropriate. We identified three sets of variables with very high correlations (r = 0.99):

TABLE 1. Sample size, number of observations, test occasions, and mean APHV determined by interpolating cubic spline (in years) in the three studies: PBMAS, HBS-III, and the HGS.

	Boys				
Study	Sample Size (*n*)	Observations (*n*)	Visits (*n*, Range)	APHV (Mean (SD))	APHV (Range)
PBMAS	79	659	5–13	13.4 (0.7)	11.1–15.6
HBS-III	42	427	7–15	13.5 (1.1)	10.9–15.9
HGS	38	745	14–24	14.0 (1.0)	11.3–16.2
	Girls				
	Sample Size (*n*)	Observations (*n*)	Tests (*n*, Range)	APHV (mean (SD))	APHV (Range)
PBMAS	72	592	6–12	11.9 (0.7)	10.3–13.6
HBS-III	39	335	4–15	11.6 (0.7)	10.5–13.4
HGS	32	676	15–26	12.1 (1.0)	9.8–14.2

SD, standard deviation.

TABLE 2. Changes in R^2 and the SEE with step-by-step refitting the original Mirwald et al. (26) equations.

Predictor 1	Predictor 2	Predictor 3	Predictor 4	R^2	Change in R^2	SEE	Change in SEE
Boys							
leg × sit				0.802	0.802	0.748	0.748
leg × sit	age × leg			0.872	0.069	0.605	0.143
leg × sit	age × leg	age × sit		0.912	0.041	0.499	0.106
leg × sit	age × leg	age × sit	Leg/ht × 100	0.915	0.003	0.490	0.009
Girls							
leg × sit				0.774	0.774	0.790	0.790
leg × sit	age × sit			0.898	0.124	0.532	0.258
leg × sit	age × sit	leg/ ht × 100		0.908	0.010	0.505	0.027
leg × sit	age × sit	leg/ ht × 100	Age × wt	0.910	0.002	0.500	0.005

Leg, leg length; sit, sitting height; ht, height; wt, weight.

age × sitting height, age × height, and age × leg length interaction variables, leg length by height ×100, leg length by sitting height ×100, and sitting height by height ×100 ratio variables and height, sitting height and leg length. We removed one variable from each set: age × leg length, leg length by height ×100, and leg length, given that leg length had the greatest measurement error (~6 mm). In our equation redevelopment, we included four candidate predictor variables (age, height, sitting height, and weight), four interactions (age × height, age × sitting height, age × weight, and leg length × sitting height), and four ratios (BMI, weight by height ×100, sitting height by height ×100, leg length by sitting height × 100) for a total of 12 potential predictors.

We accounted for within-subject correlation when redeveloping the model as previously described. We provide results from the forward stepwise variable selection procedure, which incorporated a cluster-robust variance, in Supplemental Digital Content 2 [see table, Supplemental Digital Content 2, Forward Stepwise Regression from the Original Pediatric Bone Mineral Accrual Study (PBMAS) Boys and Girls, http://links.lww.com/MSS/A477]. In Table 3, we provide the frequency of variable selection from the 5 × 7- and 5 × 6-fold random splitting exercise in boys and girls, respectively. In the boys, the most frequent predictors in both procedures were age × sitting height, and sitting height. In girls, the most frequent predictors were age × height, and leg length.

After reviewing the selected variables identified using these two approaches, we fit new sex-specific regression models to the PBMAS data. In the boys, the simplified MO regression equation is:

$$\begin{aligned}\text{Maturity offset} &= -8.128741 \\ &\quad + (0.0070346 \times (\text{age} \times \text{sitting height})); \\ \text{where } R^2 &= 0.906 \;\&\; \text{SEE} = 0.514\end{aligned}$$

In the girls, the simplified MO regression equation is:

$$\begin{aligned}\text{Maturity offset} &= -7.709133 \\ &\quad + (0.0042232 \times (\text{age} \times \text{height})); \\ \text{where } R^2 &= 0.898 \;\&\; \text{SEE} = 0.528\end{aligned}$$

All height measurements are in centimeter. The models here consider robust standard errors of coefficients as we used the cluster-robust variance estimator while developing these new equations. Using a step-by-step approach, we added a second predictor to each model, which yielded only minor changes (<1%) in the R^2 and SEE; in the boys, $R^2 = 0.910$; SEE = 0.504; and in the girls, $R^2 = 0.906$; SEE = 0.508.

We calibrated the new prediction equations using two external validation samples: the HBS-III and HGS cohorts. We plotted calibration curves (observed vs predicted) and calculated the coefficients for the calibration plots. This provided us a summary of the observed and predicted values. A calibration curve with an intercept of zero and a slope of one overlay the $y = x$ line; that is, the closer b_0 to 0 and b_1 to 1, the better the performance of the prediction on average.

We also provide a visual summary of the difference in the calibration curves between the original equations and the revised equations in boys and girls (Fig. 1). We quantified the difference between the calibration curves and present these results in Table 4. With fewer variables in the model, the redeveloped equations performed as well as the original equations in all cases except the HGS boys. The mean prediction error (in years) for the HBS boys was −0.05; HBS

TABLE 3. Results of the 5 × 7- and 5 × 6-fold random-splitting analysis with the original PBMAS boys and girls, respectively.

Potential Predictor Variables	Frequency of Variable Entering the Model (#/35)	Frequency of Variable Entering the Model (%)
Boys		
age × sitting height	35	100%
sitting height	16	46%
age	9	26%
height	8	23%
leg length × sitting height	5	14%
age × height	1	3%
leg length	1	3%
Potential Predictor Variables	**Frequency of Variable Entering the Model (#/30)**	**Frequency of Variable Entering the Model (%)**
Girls		
age × height	30	100
leg length	15	50
weight	8	27
BMI	7	23
height	4	13
leg length × sitting height	3	10
age × weight	3	10
age	2	7
weight / height ×100	1	3

This created 35 subsets in boys and 30 subsets in girls (with one observation per child; where $n \geq 70$). A forward stepwise regression procedure was used with $P \leq 0.10$ for entry and $P \geq 0.11$ for removal.

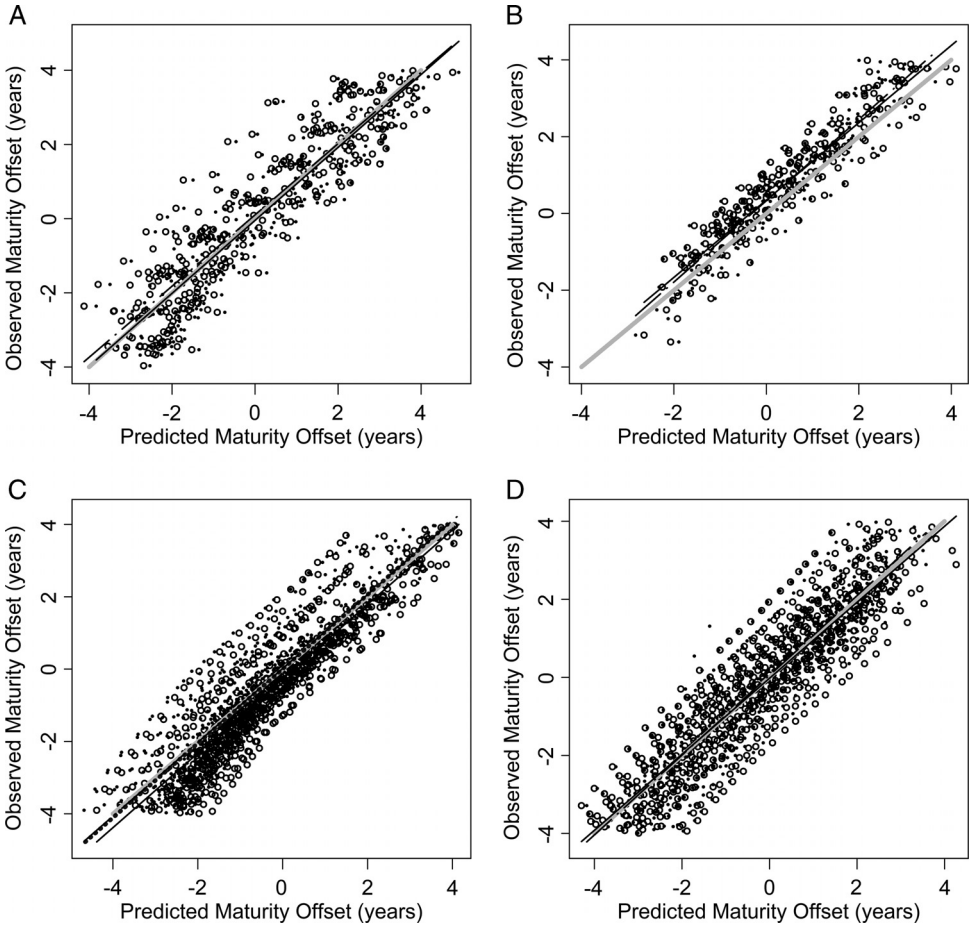

FIGURE 1—Calibration curves (observed vs predicted) in HBS-III boys (A), HBS-III girls (B), HGS boys (C), and HGS girls (D); where *thick light gray line* is $y = x$, *black solid line* is calibration line for the redeveloped equations; *black open dots* are predicted maturity offset (MO) by the redeveloped equations, *black dashed line* is calibration line for the original Mirwald et al. (26) equations, *black dots* are predicted MO by the original Mirwald et al. (26) equations, and *dark gray thick dotted line* (C only) is the recalibrated equations to account for significant difference in APHV. If $b_0 = 0$ and $b_1 = 1$, the calibration line would overlay the $y = x$ line; that is, the closer b_0 to 0 and b_1 to 1, the better the performance of the prediction on average.

girls, 0.35; HGS boys, −0.28; and HGS girls, −0.02. We attribute the poorer calibration in the HGS boys to their later APHV. Thus, we recalibrated the equation in this group by subtracting the calibration intercept (b_0), which yielded:

$$\text{Maturity offset} = (-8.128741 - 0.2683693) + (0.0070346 \times \text{age} \times \text{sitting height})$$

We describe the prediction error by MO category in Table 5 for both the redeveloped and original equations. These results indicate that prediction error is smaller for predictions made closer to APHV using both the redeveloped and original equations. The redeveloped and original equations performed similarly by MO category. With the redeveloped equations, mean differences between observed and predicted

TABLE 4. Summary results of the calibration curves and descriptive summaries of the prediction residuals (including 25th, 50th, and 75th percentiles) for the external validation samples: HBS-III and HGS.

Prediction Error	*n*	b_0 (SE)	b_1 (SE)	Min	p25	p50	p75	Max	Mean	RMSE
HBS-III boys										
Redeveloped	320	−0.05 (0.05)	0.98 (0.02)	−2.18	−0.75	−0.12	0.60	2.64	−0.05	0.8966
Original	320	0.10 (0.05)	0.96 (0.02)	−2.01	−0.58	0.02	0.74	2.81	0.10	0.9073
HBS-III girls										
Redeveloped	233	0.35 (0.03)	1.01 (0.03)	−1.28	0.01	0.41	0.77	1.80	0.35	0.6524
Original	233	0.33 (0.04)	1.06 (0.03)	−1.46	−0.01	0.40	0.75	1.98	0.40	0.6585
HGS boys										
Recalibrated*	745	−0.01 (0.01)	1.02 (0.01)	−2.00	−0.47	−0.15	0.29	2.47	0.01	0.802
Redeveloped	745	−0.20 (0.02)	1.02 (0.01)	−2.27	−0.74	−0.41	0.02	2.21	−0.27	0.8489
Original	745	0.02 (0.03)	1.02 (0.02)	−1.80	−0.46	−0.14	0.36	2.15	0.01	0.7789
HGS girls										
Redeveloped	676	−0.02 (0.03)	0.97 (0.02)	−2.05	−0.51	−0.04	0.55	2.28	−0.02	0.7006
Original	676	0.05 (0.03)	1.02 (0.02)	−2.00	−0.51	−0.02	0.58	2.68	0.05	0.8124

b_0 is the calibration curve intercept (standard error, SE), b_1 is the calibration curve slope (SE), and RMSE is the root mean squared error of the prediction.
*Recalibrated given the later APHV in the HGS boys compared with the PBMAS boys.

TABLE 5. Summary statistics (mean (SD)) for observed MO, predicted MOs, and prediction errors for the redeveloped equations and the original equations (26) (observed minus predicted) by observed MO category in HBS-III (top) and HGS (bottom) boys (left) and girls (right).

HBS-III Boys					HBS-III Girls				
MO	*n*	Predicted MO Mean (SD)	Observed-Predicted MO Redeveloped Mean (SD)	Observed-Predicted MO Original Mean (SD)	MO	*n*	Predicted MO Mean (SD)	Observed-Predicted MO Redeveloped Mean (SD)	Observed-Predicted MO Original Mean (SD)
−4	6	−3.69 (0.15)	−1.48 (0.42)	−1.22 (0.47)	−4	—	—	—	—
−3	38	−3.04 (0.35)	−0.84 (0.53)	−0.64 (0.54)	−3	4	−2.94 (0.38)	−0.74 (0.41)	−0.64 (0.58)
−2	42	−2.03 (0.34)	−0.24 (0.73)	−0.05 (0.77)	−2	11	−1.93 (0.26)	−0.36 (0.36)	−0.34 (0.42)
−1	46	−2.06 (0.34)	−0.01 (0.74)	0.20 (0.76)	−1	35	−0.97 (0.29)	0.22 (0.43)	0.18 (0.47)
0	49	−0.02 (0.37)	0.22 (0.82)	0.36 (0.85)	0	57	−0.01 (0.31)	0.36 (0.48)	0.29 (0.49)
+1	42	1.05 (0.34)	0.02 (0.93)	0.10 (0.97)	+1	50	0.98 (0.29)	0.44 (0.54)	0.39 (0.46)
+2	42	2.05 (0.32)	0.09 (0.92)	0.19 (0.94)	+2	41	1.93 (0.32)	0.44 (0.54)	0.45 (0.53)
+3	39	3.05 (0.31)	0.33 (0.91)	0.46 (0.93)	+3	24	3.07 (0.29)	0.57 (0.55)	0.73 (0.47)
+4	16	3.72 (0.14)	0.40 (0.83)	0.48 (0.83)	+4	11	3.80 (0.13)	0.71 (0.63)	0.97 (0.49)
HGS Boys					**HGS Girls**				
MO	*n*	Predicted MO Mean (SD)	Observed-Predicted MO Redeveloped Mean (SD)	Observed-Predicted MO Original Mean (SD)	MO	*n*	Predicted MO Mean (SD)	Observed-Predicted MO Redeveloped Mean (SD)	Observed-Predicted MO Original Mean (SD)
−4	36	−3.73 (0.13)	−1.16 (0.60)	−0.87 (0.59)	−4	28	−3.73 (0.14)	−0.71 (0.72)	−0.73 (0.71)
−3	95	−2.98 (0.28)	−0.86 (0.65)	−0.56 (0.63)	−3	72	−3.00 (0.28)	−0.45 (0.76)	−0.46 (0.73)
−2	126	−1.98 (0.29)	−0.53 (0.59)	−0.21 (0.60)	−2	89	−1.99 (0.28)	−0.28 (0.75)	−0.27 (0.70)
−1	142	−0.97 (0.29)	−0.23 (0.63)	0.10 (0.62)	−1	103	−0.98 (0.28)	−0.05 (0.79)	−0.02 (0.74)
0	140	0.01 (0.28)	−0.07 (0.70)	0.20 (0.70)	0	113	0.00 (0.28)	−0.01 (0.78)	0.03 (0.71)
+1	95	0.96 (0.27)	−0.01 (0.79)	0.19 (0.79)	+1	119	0.98 (0.28)	0.09 (0.81)	0.22 (0.78)
+2	56	1.92 (0.27)	0.16 (0.90)	0.37 (0.85)	+2	94	1.93 (0.29)	0.22 (0.76)	0.33 (0.69)
+3	41	2.95 (0.28)	0.37 (0.94)	0.63 (0.86)	+3	43	2.93 (0.26)	0.61 (0.80)	0.77 (0.71)
+4	14	3.76 (0.14)	0.27 (0.70)	0.60 (0.62)	+4	15	3.70 (0.15)	0.90 (0.73)	1.13 (0.65)

were less than 0.5 yr between −2 and +4 MO in the HBS-III boys, −2 and +2 in the HBS-III girls, −1 and +4 MO in the HGS boys; and −3 and +2 MO in the HGS girls. With the original equations, mean differences between observed and predicted were less than 0.5 yr. There were the same as the redeveloped equation in the HBS-III boys, girls, and the HGS girls but slightly less in the HGS boys, were mean differences in the observed and the predicted, were less than 0.5 yr between −2 and +2 MO. That is, there was a larger window of prediction (with <0.5-yr error) using the redeveloped equation compared to the original. The ideal timing to predict is when a participant is closest to his or her expected APHV or a MO of zero.

Our sample was not large enough to rigorously assess variation in prediction error due to early- and late-maturing children. However, we identified a number of children in our sample who might be considered early or late maturers, and we used these data to illustrate how the equation performs in those cases. For the HBS-III boys whose APHV was between 13.5 and 14.5 yr, the mean (SD) error (MO; in years) was 0.57 (0.50), where *n* (observations) = 132. In the boys with APHV less than 12.5 yr (i.e., early maturers), the mean (SD) error was 0.87 (0.48), where $n = 39$; the mean (SD) prediction error using the original equation in early-maturing boys was 1.14 (0.86). In the boys with APHV greater than 15.5 yr (i.e., late maturers), the mean (SD) error was 1.35 (0.47), where $n = 14$. The mean (SD) prediction error was similar 1.36 (0.56) in late-maturity boys when using the original equation. In the HBS-III girls for whom APHV were between 11.5 and 12.5 yr, the mean (SD) error was 0.12 (0.42), where $n = 66$. In the girls with APHV less than 11.0 yr (i.e., early maturers), the mean (SD) error was 0.94 (0.31), where $n = 49$. Finally, in the girls with APHV greater than 13.0 yr (i.e., late maturers), the mean (SD) error was −0.76 (0.24), where $n = 14$. The mean (SD) prediction error was similar in early- and late-maturity girls when using the original equation: 0.95 (0.43) and −0.81 (0.42), respectively.

Finally, we acknowledge that sitting height may not be assessed in all growth studies. Therefore, we created an alternative model for boys that included an age–height interaction term for use when sitting height has not been documented. The alternate age × height MO regression equation for boys is:

$$\text{Maturity offset} = -7.999994 + (0.0036124 \times (\text{age} \times \text{height}));$$
$$R^2 = 0.896;\ \text{SEE} = 0.542$$

All height measurements are in centimeter. We provide a full description of the calibration for the alternative equation for boys in Supplemental Digital Content 3 (see document, Supplemental Digital Content 3, Alternate Maturity Offset Prediction Equation for Boys when there is No Documented Sitting Height and External Calibration Results, http://links.lww.com/MSS/A478).

DISCUSSION

Managing the effect of maturity and the substantial range in the tempo and timing of growth in studies of children continues to be a challenge. Valid and reliable maturity prediction equations using easily accessible noninvasive somatic growth measures may be one positive and effective solution. Our study identified and addressed issues of within-subject correlation and overfitting of the original Mirwald et al. (26) prediction equations. With these issues accounted for, the redeveloped sex-specific regression equations more appropriately predict MO and APHV in growing children. Our

findings suggest that the Mirwald et al. (26) equations can be simplified without a meaningful increase in estimation error and, importantly, without sacrificing prediction accuracy. We have also created equations that do not require sitting height; enhancing their usability across disciplines and by researchers, physicians, and other health professionals. Finally, although we saw similar predictions and prediction errors in our two external samples, the redeveloped equations should theoretically produce better fits across a range of external samples.

The original equations were developed to predict how far a child is away from APHV by providing the user a prediction of MO and a way to align children on a maturational timeline. The authors thoughtfully considered predictors based on differential timing of pubertal growth and developed sex-specific regression equations and identified four predictor variables that comprised the models for boys and girls, respectively. In the boys, the predictors were leg length × sitting height, age × leg length, age × sitting height, and leg length by height ×100. In the girls, the predictors were leg length × sitting height, age × sitting height, leg length by height ×100, and age × weight. In these equations, $R^2 = 0.891$ and SEE = 0.592 for the boys and the $R^2 = 0.890$ and SEE = 0.569 for girls. However, these values may have been compromised for three reasons: 1) the number of predictors was too large relative to the number of independent observations; 2) the R^2 was calculated from data used to create the model; and 3) the SEE may underestimate the prediction error in both equations, as the analyses did not account for within-subject correlation.

Our results attained similar predictions and prediction errors using simpler models. To minimize prediction error, an accurate but more conservative estimate of target prediction error is preferable (38). R^2 is the proportion of variance explained by the regression model and increases with the addition of each covariate. Importantly, overfitting may lead to a model that does not perform well with external data. Given that the usefulness of the original predictions (26) has been questioned in external samples (21–23), we aimed to determine if this lack of fit could be a result of overfitting. Thus, we used several statistical strategies to ensure the best fit for our redeveloped models: 1) we assessed the original prespecified predictors, 2) we reduced the number of predictors using biological and statistical rationale, 3) we used two methods that considered the within-subject clustering and identified final predictors; 4) we used a step-by-step forward-stepwise regression procedure (using robust standard errors) to enter variables in the model and assess the change in R^2 and SEE, and 5) with the final redeveloped models, we calibrated and compared observed versus predicted values using two external samples that had accurate measures of APHV and MO.

Many studies are using maturity prediction equations to classify maturational status and align children by MO or APHV. For example, equations are most commonly used to assess maturity status of athletes, for talent identification, or to assess changes in function or performance after activity-related interventions (11,24,27,31,33,37). It has historically been a challenge to provide a model that equitably classifies youth for sports participation and competition (4,6). Assessments are continually being sought at all levels of sport competition to represent a young athlete's maturity status. Malina and colleagues (21–23) cautioned those who might use these assessments for sports groups when there is the propensity for young participants to be early or late maturers (e.g., gymnastics). Although redevelopment of the original Mirwald et al. (26) equations addressed some sources of "systematic" error noted by Malina and Koziel (22,23), we note that prediction error is still likely to be slightly higher in early- or late- maturing children. Not surprisingly, prediction error also increases the farther a child is away from expected APHV.

We identified a small subset of children in our sample considered to be early or late maturers. When we applied the redeveloped prediction equation, prediction error was 2–3 times greater in the boys and 6–7 times greater in the girls, for early and late maturers. The mean prediction error for all groups with the exception of late-maturing boys was less than ±1 yr; perhaps still within the range to appropriately classify children into maturity groups for sports participation purposes.

Further, the Mirwald et al. (26) equations were developed using data acquired from a sample of white children. However, they have been applied, perhaps inappropriately, to predict biological age in ethnically diverse cohorts (1,10,13, 14,25). In light of the documented differences in the timing and tempo of growth and maturation between ethnicities (12), there is a need to confirm the usefulness of our new equations in other ethnic samples and, if needed, develop ethnic-specific equations.

We also revisited the findings of Mirwald et al. (26) that in 95% of cases, MO could be predicted accurately within ±1 yr. However, this suggestion may have been optimistic for the reasons of within-subject correlation and overfitting. We assessed the predictive ability of the original equations and found that in 80%–85% of cases, MO was predicted accurately within ±1 yr in two external samples. We then assessed the redeveloped equations presented here and found that in 90% of the cases MO was predicted accurately within ±1 yr, despite large ranges in APHV and MO. These estimates are more realistic, as they are calculated with external samples. We attribute the good fits (despite being simpler models) to the interaction terms included in the models, which consider both the measured aspect of linear growth (i.e., sitting height in boys; height in girls), and the timing (i.e., the interaction with age of the event). Mirwald et al. (26) also carefully considered differential timing of the various aspects of linear growth; thus, we found that the most influential predictors were these interaction terms.

We conclude that there is a profound need for accurate noninvasive approaches to predict maturity in children and youth. We found evidence of overfitting in the Mirwald et al. (26) prediction equations and subsequently

redeveloped the equations. The redeveloped equations perform similarly or better than the original equations; we found that 90% of predictions were within ±1 yr in two external samples. Importantly, our predictions provide alternatives for investigations in which sitting height has not been documented. Given the rigor with which we developed our prediction models, we propose that they be considered by clinicians, sports governing bodies, and in pediatric research to assess maturity and to align children on biological age. The predictions will prove most useful in children from similar populations. We acknowledge that external validation of models using data acquired from early or late maturers such as athletes, clinical populations, or ethnic groups is warranted.

This study received funding support from a number of agencies. The HBS-III was supported by the Canadian Institute for Health Research (CIHR; MOP-84575), the Michael Smith Foundation for Health Research (MSFHR), the British Columbia Ministry of Health, and 2010 Legacies Now. PBMAS was supported by the Canadian National Health and Research Development Program (NHRDP; 6608-1261) and CIHR (MOP-57671, MOP-83391, and MOP-85582). Additional support was provided by the Saskatchewan Health Research Foundation.

The authors thank the PBMAS, HBS-III, and HGS research teams, participants, and their families. The authors likewise thank the late Professor James Tanner and Mr RH Whitehouse for the careful collection and preservation of the HGS data. The authors also sincerely thank the original authors of the Mirwald et al. (2002) publication: Drs Robert Mirwald, Donald Bailey, and Gaston Beunen.

The results of the present study do not constitute endorsement by the American College of Sports Medicine.

REFERENCES

1. Anderson KD, Baxter-Jones AD, Faulkner RA, Muhajarine N, Henry CJ, Chad KE. Assessment of total and central adiposity in Canadian aboriginal children and their Caucasian peers. *Int J Pediatr Obes*. 2010;5(4):342–50.
2. Bailey DA. The Saskatchewan Pediatric Bone Mineral Accrual Study: bone mineral acquisition during the growing years. *Int J Sports Med*. 1997;18(3 Suppl):S191–4.
3. Bailey DA, McKay HA, Mirwald RL, Crocker PR, Faulkner RA. A six-year longitudinal study of the relationship of physical activity to bone mineral accrual in growing children: the university of Saskatchewan bone mineral accrual study. *J Bone Miner Res*. 1999;14(10):1672–9.
4. Baxter-Jones AD. Growth and development of young athletes. Should competition levels be age related? *Sports Med*. 1995; 20(2):59–64.
5. Baxter-Jones AD, Mirwald RL, McKay HA, Bailey DA. A longitudinal analysis of sex differences in bone mineral accrual in healthy 8- to 19-year-old boys and girls. *Ann Hum Biol*. 2003; 30(2):160–75.
6. Baxter-Jones ADG, Eisenmann JC, Sherar LB. Controlling for maturation in pediatric exercise science. *Pediatr Exerc Sci*. 2005; 17(1):18–30.
7. Beunen GP, Rogol AD, Malina RM. Indicators of biological maturation and secular changes in biological maturation. *Food Nutr Bull*. 2006;27(Suppl Growth Standard 4):S244–56.
8. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*. 1986; 1(8476):307–10.
9. Cameron N, Bogin B. *Human Growth and Development*. 2nd ed. London, UK: Elsevier; 2012.
10. Carvalho HM, Coelho-e-Silva M, Valente-dos-Santos J, Goncalves RS, Philippaerts R, Malina R. Scaling lower-limb isokinetic strength for biological maturation and body size in adolescent basketball players. *Eur J Appl Physiol*. 2012;112(8):2881–9.
11. Erlandson MC, Kontulainen SA, Baxter-Jones AD. Precompetitive and recreational gymnasts have greater bone density, mass, and estimated strength at the distal radius in young childhood. *Osteoporos Int*. 2011;22(1):75–84.
12. Eveleth PB, Tanner JM. *Worldwide Variation in Human Growth*. 2nd ed. Cambridge, UK: Cambridge University Press; 1990.
13. Farr JN, Chen Z, Lisse JR, Lohman TG, Going SB. Relationship of total body fat mass to weight-bearing bone volumetric density, geometry, and strength in young girls. *Bone*. 2010;46(4): 977–84.
14. Fukunaga Y, Nishizono H. Longitudinal height velocity during adolescence and its relationship to sports injury. *Japanese Journal of Physical Fitness and Sports Medicine*. 2010;59(5): 521–8.
15. Harrell FE Jr., Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med*. 1996; 15(4):361–87.
16. Hauspie RC, Cameron N, Molinari L. Methods in human growth research. In: Molinari L, Gasser T, editors. *The Human Growth Curve*. Cambridge: Cambridge University Press; 2004.
17. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. *Ann Intern Med*. 1999;130(6): 515–24.
18. Macdonald HM, Kontulainen SA, Khan KM, McKay HA. Is a school-based physical activity intervention effective for increasing tibial bone strength in boys and girls? *J Bone Miner Res*. 2007; 22(3):434–46.
19. Mackelvie KJ, McKay HA, Khan KM, Crocker PR. A school-based exercise intervention augments bone mineral accrual in early pubertal girls. *J Pediatr*. 2001;139(4):501–8.
20. Malina RM, Bouchard C, Bar-Or O. *Growth, Maturation, and Physical Activity*. 2nd ed. Champaign (IL): Human Kinetics; 2004.
21. Malina RM, Claessens AL, Van Aken K, et al. Maturity offset in gymnasts: application of a prediction equation. *Med Sci Sports Exerc*. 2006;38(7):1342–7.
22. Malina RM, Koziel SM. Validation of maturity offset in a longitudinal sample of Polish boys. *J Sports Sci*. 2014;32(5):424–37.
23. Malina RM, Koziel SM. Validation of maturity offset in a longitudinal sample of Polish girls. *J Sports Sci*. 2014;32(14):1374–82.
24. Matthys SP, Fransen J, Vaeyens R, Lenoir M, Philippaerts R. Differences in biological maturation, anthropometry and physical performance between playing positions in youth team handball. *J Sports Sci*. 2013;31(12):1344–52.
25. Mendez-Villanueva A, Buchheit M, Kuitunen S, Douglas A, Peltola E, Bourdon P. Age-related differences in acceleration, maximum running speed, and repeated-sprint performance in young soccer players. *J Sports Sci*. 2011;29(5):477–84.
26. Mirwald RL, Baxter-Jones AD, Bailey DA, Beunen GP. An assessment of maturity from anthropometric measurements. *Med Sci Sports Exerc*. 2002;34(4):689–94.
27. Mitchell C, Cohen R, Dotan R, Gabriel D, Klentrou P, Falk B. Rate of muscle activation in power- and endurance-trained boys. *Int J Sports Physiol Perform*. 2011;6(1):94–105.
28. Nishiyama KK, Macdonald HM, Moore SA, Fung T, Boyd SK, McKay HA. Cortical porosity is higher in boys compared with girls at the distal radius and distal tibia during pubertal growth: an HR-pQCT study. *J Bone Miner Res*. 2012;27(2):273–82.
29. Preece MA, Baines MJ. A new family of mathematical models describing the human growth curve. *Ann Hum Biol*. 1978;5(1):1–24.
30. Rogers WH. Regression standard errors in clustered samples. *Stata Technical Bulletin*. 1993;13:19–23.

31. Rogowski I, Ducher G, Brosseau O, Hautier C. Asymmetry in volume between dominant and nondominant upper limbs in young tennis players. *Pediatr Exerc Sci*. 2008;20(3):263–72.
32. Ross WD, Marfell-Jones MJ. Kinanthropometry. In: Green H, editor. *Physiological Testing of the High Performance Athlete*. Champaign: Human Kinetics; 1991.
33. Sherar LB, Baxter-Jones AD, Faulkner RA, Russell KW. Do physical maturity and birth date predict talent in male youth ice hockey players? *J Sports Sci*. 2007;25(8):879–86.
34. Tabachnick B, Fidell L. *Using Multivariate Statistics*. Needham Heights (MA): Allyn & Bacon; 2001.
35. Tanner JM. *Growth at Adolescence; with a General Consideration of the Effects of Hereditary and Environmental Factors upon Growth and Maturation from Birth to Maturity*. 2nd ed. Springfield (IL): Oxford, Blackwell Scientific Publications; 1962. p. 325.
36. Tanner JM. *Foetus into Man*. 2nd ed. Herts (UK): Castlemead Publications; 1989.
37. Vandorpe B, Vandendriessche JB, Vaeyens R, et al. The value of a non-sport-specific motor test battery in predicting performance in young female gymnasts. *J Sports Sci*. 2012;30(5):497–505.
38. Vittinghoff E, Glidden DV, Shiboski SC, McCulloch CE. Predictor selection. In: *Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models*. New York: Springer; 2012. pp. 395–429.
39. Williams RL. A note on robust variance estimation for cluster-correlated data. *Biometrics*. 2000;56(2):645–6.