

2. METHODOLOGY

2.1 Design of the WHO Multicentre Growth Reference Study

The MGRS (July 1997–December 2003) was a population-based study that took place in the cities of Davis, California, USA; Muscat, Oman; Oslo, Norway; and Pelotas, Brazil; and in selected affluent neighbourhoods of Accra, Ghana and South Delhi, India. The MGRS protocol and its implementation in the six sites are described in detail elsewhere (de Onis et al., 2004a). Briefly, the MGRS combined a longitudinal component from birth to 24 months with a cross-sectional component of children aged 18–71 months. In the longitudinal component, mothers and newborns were screened and enrolled at birth and visited at home a total of 21 times on weeks 1, 2, 4 and 6; monthly from 2–12 months; and bimonthly in the second year. In the cross-sectional component, children aged 18–71 months were measured once, except in the two sites (Brazil and USA) that used a mixed-longitudinal design in which some children were measured two or three times at three-month intervals. Both recumbent length and standing height were measured for all children aged 18–30 months. Data were collected on anthropometry, motor development, feeding practices, child morbidity, perinatal factors, and socioeconomic, demographic and environmental characteristics (de Onis et al., 2004b).

The study populations lived in socioeconomic conditions favourable to growth and where mobility was low, $\geq 20\%$ of mothers followed WHO feeding recommendations and breastfeeding support was available (de Onis et al., 2004b). Individual inclusion criteria were: no known health or environmental constraints to growth, mothers willing to follow MGRS feeding recommendations (i.e. exclusive or predominant breastfeeding for at least 4 months, introduction of complementary foods by the age of 6 months, and continued partial breastfeeding up to at least 12 months), no maternal smoking before and after delivery, single term birth, and absence of significant morbidity (de Onis et al., 2004b).

As part of the site-selection process in Ghana, India and Oman, surveys were conducted to identify socioeconomic characteristics that could be used to select groups whose growth was not environmentally constrained (Owusu et al., 2004; Bhandari et al., 2002; Mohamed et al., 2004). Local criteria for screening newborns, based on parental education and/or income levels, were developed from those surveys. Pre-existing survey data for this purpose were available from Brazil, Norway and the USA. Of the 13 741 mother-infant pairs screened for the longitudinal component, about 83% were ineligible (WHO Multicentre Growth Reference Study Group, 2006d). Families' low socioeconomic status was the most common reason for ineligibility in Brazil, Ghana, India and Oman, whereas parental refusal was the main reason for non-participation in Norway and USA (WHO Multicentre Growth Reference Study Group, 2006d). For the cross-sectional component, 69% of the 21 510 subjects screened were excluded for reasons similar to those observed in the longitudinal component.

Term low-birth-weight (<2500 g) infants (2.3%) were *not* excluded. Since it is likely that in well-off populations such infants represent small but normal children, their exclusion would have artificially distorted the standards' lower percentiles. Eligibility criteria for the cross-sectional component were the same as those for the longitudinal component with the exception of infant feeding practices. A minimum of three months of any breastfeeding was required for participants in the study's cross-sectional component.

2.2 Anthropometry methods

Data collection teams were trained at each site during the study's preparatory phase, at which time measurement techniques were standardized against one of two MGRS anthropometry experts. During the study, bimonthly standardization sessions were conducted at each site. Once a year the anthropometry expert visited each site to participate in these sessions (de Onis et al., 2004c). Results from the anthropometry standardization sessions have been reported elsewhere (WHO Multicentre Growth Reference Study Group, 2006e). For the longitudinal component of the study, screening teams measured newborns within 24 hours of delivery, and follow-up teams conducted home visits until 24 months of age. The follow-up teams were also responsible for taking measurements in the cross-sectional component involving children aged 18–71 months (de Onis et al., 2004b).

The MGRS data included weight and head circumference at all ages, recumbent length (longitudinal component), height (cross-sectional component), and arm circumference, triceps and subscapular skinfolds (all children aged ≥ 3 months). However, this report presents only the standards based on length or height and weight. Observers working in pairs collected anthropometric data. Each observer independently measured and recorded a complete set of measurements, after which the two compared their readings. If any pair of readings exceeded the maximum allowable difference for a given variable (e.g. weight, 100 g; length/height, 7 mm), both observers once again independently measured and recorded a second and, if necessary, a third set of readings for the variable(s) in question (de Onis et al., 2004c).

All study sites used identical measuring equipment. Instruments needed to be highly accurate and precise, yet sturdy and portable to enable them to be carried back and forth on home visits. Length was measured with the portable Harpenden Infantometer (range 30–110 cm, with digit counter readings precise to 1 mm). The Harpenden Portable Stadiometer (range 65–206 cm, digit counter reading) was used for measuring adult and child heights. Portable electronic scales with a taring capability, calibrated to 0.1 kg (i.e. UNICEF Electronic Scale 890 or UNISCALE), were used to measure weight. Length and height were recorded to the last completed unit rather than to the nearest unit. To correct for the systematic negative bias introduced by this practice, 0.05 cm (i.e. half of the smallest measurement unit) was added to each measurement before analysis. This correction did not apply to weight, which was rounded off to the nearest 100 g. Full details of the instruments used and how measurements were taken are provided elsewhere (de Onis et al., 2004c).

2.3 Sample description

The total sample size for the longitudinal and cross-sectional components from all six sites was 8440 children. A total of 1743 children were enrolled in the longitudinal sample, six of whom were excluded for morbidities affecting growth (4 cases of repeated episodes of diarrhoea, 1 case of repeated episodes of malaria, and 1 case of protein-energy malnutrition) leaving a sample of 1737 children (894 boys and 843 girls). Of these, the mothers of 882 children (428 boys and 454 girls) complied fully with the MGRS infant-feeding and no-smoking criteria and completed the follow-up period of 24 months (96% of compliant children completed the 24-month follow-up) (Table 1). The other 855 children contributed only birth measurements, as they either failed to comply with the study's infant-feeding and no-smoking criteria or dropped out before 24 months. The reason for using these measurements was to increase the sample size at birth to minimize the left-edge effect. The size at birth of these 855 children was similar to that of the compliant sample (Table 2). The total number of records for the longitudinal component was 19 900.

Table 1 Total sample and number of compliant children in the longitudinal component

Site	N	Compliant ^a		
		Boys	Girls	Total
Brazil	309	29	37	66
Ghana	328	103	124	227
India	301	84	89	173
Norway	300	75	73	148
Oman	291	73	76	149
USA	208	64	55	119
All	1737	428	454	882

^a Compliant with infant-feeding and no-smoking criteria and completed the 24-month follow-up.

Table 2 Comparison of mean size at birth for compliant newborns and those that contributed only birth measurements

Measurement	Compliant ^a N=882	Non-compliant N=855
Weight (g)	3325	3306
Length (cm)	49.6	49.5
Head circumference (cm)	34.1	34.2

^a Compliant with infant-feeding and no-smoking criteria and completed the 24-month follow-up.

The cross-sectional sample comprised 6697 children. Of these, 28 were excluded for medical conditions affecting growth (20 cases of protein-energy malnutrition, five cases of haemolytic anaemia G6PD deficiency, two cases of renal tubulo-interstitial disease, and one case of Crohn disease) leaving a final sample of 6669 children (3450 boys and 3219 girls) (Table 3). The total number of records in the cross-sectional component was 8306 as some children in Brazil and the USA were measured two or three times at three-month intervals (Table 4). A full description of the MGRS sample with regard to screening, recruitment, sample attrition and compliance, as well as the baseline characteristics of the study sample is provided elsewhere (WHO Multicentre Growth Reference Study Group, 2006d).

Table 3 Total sample of children in the cross-sectional component

Site	Boys	Girls	Total
Brazil	237	243	480
Ghana	684	719	1403
India	840	647	1487
Norway	725	660	1385
Oman	714	724	1438
USA	250	226	476
All	3450	3219	6669

Table 4 Total sample of children in the cross-sectional component by number of visits and total number of records

Site	Brazil	Ghana	India	Norway	Oman	USA	All
One visit	34	1403	1487	1385	1438	55	5802
Two visits	36	0	0	0	0	61	97
Three visits	410	0	0	0	0	360	770
No. of children	480	1403	1487	1385	1438	476	6669
No. of records	1336	1403	1487	1385	1438	1257	8306

2.4 Data cleaning procedures and exclusions

Data cleaning

The MGRS data management protocol (Onyango et al., 2004) was designed to create and manage a large databank of information collected from multiple sites over a period of several years. Data collection and processing instruments were prepared centrally and used in a standardized fashion across sites. The data management system contained internal validation features for timely detection of data errors and its standard operating procedures stipulated a method of master file updating and correction that maintained a clear trail for data-auditing purposes. Each site was responsible for collecting, entering, verifying and validating data, and for creating site-level master files. Data from

the sites were sent to WHO/HQ every month for master file consolidation and more extensive quality control checking. All errors identified were communicated to the site for correction at source.

After data collection was completed at a given site, a period of about 6 months was dedicated to in-depth data quality checking and master file cleaning. Detailed validation reports, descriptive statistics and plots were produced from the site's master files. For the longitudinal component, each anthropometric measurement was plotted for every child from birth to the end of his/her participation. These plots were examined individually for any questionable patterns. Query lists from these analyses were sent to the site for investigation and correction, or confirmation, as required. As with the data collection process, the site data manager prepared correction batches to update the master files. The updated master files were then sent to WHO/HQ and this iterative quality assurance process continued until all identifiable problems had been detected and corrected. The rigorous implementation of what was a highly demanding protocol yielded very high-quality data.

Data exclusions

To avoid the influence of unhealthy weights for length/height, observations falling above +3 SD and below -3 SD of the sample median were excluded prior to constructing the standards. For the cross-sectional sample, the +2 SD cut-off (i.e. 97.7 percentile) was applied instead of +3 SD as the sample was exceedingly skewed to the right, indicating the need to identify and exclude high weights for height. This cut-off was considered to be conservative given that various definitions of overweight all apply lower cut-offs than the one used (Daniels et al., 2005; Koplan et al., 2005).

To derive the above-mentioned cut-offs based on the sex-specific weight-for-length/height indicator, the weight median and coefficient of variation curves were modelled continuously across length/height using an approach that accounted for the sample's asymmetry as described below. The data were split into two sets: one set with all points above the median and another with all points below the median. For each of the two sets, mirror values were generated to create symmetrically distributed values around the median for the upper and lower sets. The generation of mirror data was necessary to simulate a symmetric distribution based on the distinct variabilities of the upper and lower sets. For each of the mirror data sets, median and coefficient of variation curves were estimated continuously across the length/height range using the LMS method (Cole and Green, 1992) fixing L=1, i.e. fitting a normal distribution to the data for each specific length/height value, to derive the corresponding cut-offs. In total, only a small proportion of observations were excluded for unhealthy weight-for-length/height: 185 (1.4%) for boys and 155 (1.1%) for girls, most of which were in the upper end of the cross-sectional sample distribution (Table 5).

Table 5 Number of observations by sex and study component included and excluded on the basis of weight-for-length/height

Boys		LS	%	CS	%	Total	%
Included		9233	99.3	4135	97.2	13 368	98.6
Excluded	Lower	11	0.1	2	0.1	13	0.1
	Upper	56	0.6	116	2.7	172	1.3
Total		9300	100.0	4253	100.0	13 553	100.0
Girls		LS	%	CS	%	Total	%
Included		9740	99.6	3886	97.2	13 626	98.9
Excluded	Lower	7	0.1	3	0.1	10	0.1
	Upper	35	0.3	110	2.7	145	1.0
Total		9782	100.0	3999	100.0	13 781	100.0

LS, Longitudinal study; CS, Cross-sectional study.

In addition, a few influential observations for indicators other than weight-for-height were excluded when constructing the individual standards: for weight-for-age boys, 4 (0.03%) and girls, 1 (0.01%) observations and, for length/height-for-age boys, 3 (0.02%) and girls, 2 (0.01%) observations. These observations were set to missing in the final data set and therefore did not contribute to the construction of the weight-for-length/height and body mass index-for-age standards. The final number of observations used in the construction of the WHO child growth standards is shown in Table 6.

Table 6 Number of observations used in the construction of the WHO child growth standards by sex and anthropometric indicator

Indicator	Girls	Boys	Total
Weight-for-length/height	13 623	13 362	26 985
Weight-for-age	14 056	13 797	27 853
Length/height-for-age	13 783	13 551	27 334
BMI-for-age	13 623	13 362	26 985

2.5 Statistical methods for constructing the growth curves

The construction of the growth curves followed a careful, methodical process. This involved:

- detailed examination of existing methods, including types of distributions and smoothing techniques, in order to identify the best possible approach;
- selection of a software package flexible enough to allow comparative testing of alternative methods and the actual generation of the curves;
- systematic application of the selected approach to the data to generate the models that best fit the data.

A group of statisticians and growth experts met at WHO/HQ to review possible choices of methods and to define a strategy and criteria for selecting the most appropriate model for the MGRS data (Borghetti et al., 2006). As many as 30 construction methods for attained growth curves were examined. The group recommended that methods based on selected distributions be compared and combined with two smoothing techniques for fitting parameter curves to further test and provide the best possible approach to constructing the WHO child growth standards.

Choice of distribution. Five distributions were identified for detailed testing: Box-Cox power exponential (Rigby and Stasinopoulos, 2004a), Box-Cox t (Rigby and Stasinopoulos, 2004b), Box-Cox normal (Cole and Green, 1992), Johnson's SU (Johnson, 1949), and modulus-exponential-normal (Royston and Wright, 1998). The first four distributions were fitted using GAMLSS (Generalized Additive Models for Location, Scale and Shape) software (Stasinopoulos et al., 2004) and the last using the "xriml" module in STATA software (Wright and Royston, 1996). The comparison was done by age group, without considering the smoothing component. The Box-Cox-power-exponential (BCPE) distribution with four parameters — μ (for the median), σ (coefficient of variation), ν (Box-Cox transformation power) and τ (parameter related to kurtosis) — was selected for constructing the curves. The BCPE is a flexible distribution that offers the possibility to adjust for kurtosis, thus providing the framework necessary to test if fitting the distribution's fourth moment improves the estimation of extreme percentiles. It simplifies to the normal distribution when $\nu=1$ and $\tau=2$, and when $\nu \neq 1$ and $\tau=2$, the distribution is the same as the Box-Cox normal (LMS method's distribution). The BCPE is defined by a power transformation (or Box-Cox transformation) Y^V having a shifted and scaled (truncated) power exponential (or Box-Tiao) distribution with parameter τ (Rigby and Stasinopoulos, 2004a).

Apart from other theoretical advantages, the BCPE presents as good or better goodness-of-fit than the modulus-exponential-normal or the SU distribution.

Choice of smoothing technique. The expert group recommended two smoothing techniques for comparison: cubic splines and fractional polynomials (Borghi et al., 2006). Using the GAMLSS software, the two techniques were compared for smoothing length/height-for-age, weight-for-age and weight-for-length/height curves. For the fractional polynomials, a function in GAMLSS was used that estimates the best set of powers among $\{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$ within the choices of polynomials with the same number of terms. The best fractional polynomial for 1, 2 or 3 terms was fitted for each parameter curve. A number of combinations were tried among the different parameter curves, considering the Akaike Information Criterion (Akaike, 1974), *AIC*, defined as:

$$AIC = -2L + 2p,$$

where L is the maximized likelihood and p is the number of parameters (or the total number of degrees of freedom). According to this criterion, the best model is the one with the smallest *AIC* value.

The cubic spline smoothing technique offered more flexibility than fractional polynomials in all cases. For the length/height-for-age and weight-for-age standards, a power transformation applied to age prior to fitting was necessary to enhance the goodness of fit by the cubic spline technique.

Choice of method for constructing the curves. In summary, the BCPE method, with curve smoothing by cubic splines, was selected as the approach for constructing the growth curves. This method is included in a broader methodology, the GAMLSS (Rigby and Stasinopoulos, 2005), which offers a general framework that includes a wide range of known methods for constructing growth curves. The GAMLSS allows for modeling the mean (or median) of the growth variable under consideration as well as other parameters of its distribution that determine scale and shape. Various kinds of distributions can be assumed for each growth variable of interest, from normal to skewed and/or kurtotic distributions. Several smoothing terms can be used in generating the curves, including cubic splines, *lowess* (locally weighted least squares regression), polynomials, power polynomials and fractional polynomials. The simplified notation to describe a particular model within the class of the BCPE method is:

$$BCPE(x=x, df(\mu)=n_1, df(\sigma)=n_2, df(v)=n_3, df(\tau)=n_4),$$

where $df(\cdot)$ are the degrees of freedom for the cubic splines smoothing the respective parameter curve and x is age (or transformed age) or length/height. Note that when $df(\cdot)=1$, the smoothing function reduces to a constant and when $df(\cdot)=2$, it reduces to a linear function. The GAMLSS software was used to construct the WHO child growth standards. The main selected diagnostic tests and tools are available in this software. To complement and test the software, Dr Huiqi Pan and Professor Tim Cole provided the software LMS Pro, which offers the fitting of growth curves using the LMS method in a user-friendly and interactive way, including some of the available diagnostics for choosing the best set of degrees of freedom for the cubic splines and goodness-of-fit statistics. Wright and Royston's package "xriml", developed in the STATA environment, was used to test the fitting of fractional polynomials (Wright and Royston, 1996).

Diagnostic tests and tools for selecting the best model. The process for selecting the best model to construct each growth standard involved choosing, first, the best model *within* a class of models and, second, the best model *across* different classes of models. The set of diagnostic tests and tools was selected based on recommendations from the statistical expert group (Borghi et al., 2006), with additional contributions by Rigby and Stasinopoulos (2004a) and Pan and Cole (2004).

In most cases, before fitting the cubic splines, an age transformation was needed to stretch the age scale for values close to zero. Despite their complexity in terms of shape, even the flexible cubic splines fail to adequately fit early infancy growth with reasonable degrees of freedom. When the degrees of freedom are increased excessively, the function can fit well in infancy but it under-smoothes at older ages. The solution is to expand the age scale when growth velocity is high and to compress it when it is low (Cole et al., 1998). A power transformation applied to age, i.e. $f(\lambda)=age^\lambda$, was a good solution for the considered cases. Therefore, prior to determining the best degrees of freedom for the parameter curves, a search was conducted for the best λ for the age power transformation. For this, an arbitrary starting model was used to search for the best age-transformation power (λ) based only on the global deviance values over a preset grid of λ values, since the degrees of freedom remained unchanged. The grid of λ values ranged from 0.05 to 1 in 0.05 intervals, with the exception of the BMI-for-age standards for children younger than 24 months, for which the value 0.01 also was considered. No length/height transformation was necessary for weight-for-length/height.

(a) Selecting the best model within a class of models

Models were grouped in classes according to the parameters to be modelled. The alternative to modelling parameters was to fix them, e.g. $\nu=1$ or $\tau=2$. The criteria used to choose among models within the same class were the *AIC* and the generalized version of it with penalty equal to 3 (*GAIC(3)*) as defined in Rigby and Stasinopoulos (2004a):

$$GAIC(3) = -2L + 3p,$$

where L is the maximized likelihood and p is the number of parameters (or the total number of degrees of freedom). While the use of the *AIC* enhances the fitting of local trends, smoother curves are obtained when the model's choice is based on the *GAIC(3)* criterion. Consistency in the use of these two criteria was attempted across all indicators. For selecting the best combination of $df(\mu)$ and $df(\sigma)$, both criteria were used in parallel. In cases of disagreement, *AIC* was used to select $df(\mu)$ and *GAIC(3)* to select $df(\sigma)$, overall favouring the options which offered a good compromise between keeping estimates close to the empirical values and producing smooth curves. Only *GAIC(3)* values were examined to select $df(\nu)$ and, whenever needed, $df(\tau)$. In rare cases, other age-specific diagnostic tools were considered for selecting the model with an adequate number of degrees of freedom for the cubic splines fitting the parameter curves. Worm plots (van Buuren and Fredriks, 2001) and Q-test (Royston and Wright, 2000) were used conjointly for this purpose.

Group-specific Q-test statistics resulting in absolute values of z_1 , z_2 , z_3 or z_4 that were larger than 2 were interpreted to indicate a misfit of, respectively, mean, variance, skewness or kurtosis. The overall Q-test statistics combining all groups were based on a Chi-square distribution, which assumes that observations from different groups are independent. In this case, however, given the repeated measurements in the longitudinal study component, the resulting test's p-values could be distorted slightly. To minimize this potential problem, age groups were designed to avoid repeated measurements of the same child within the same age group. The age groups were formed in time intervals (days) to achieve an approximately even sample size distribution across the entire age range of interest, especially in the cross-sectional component, where sample sizes are smaller than in the longitudinal data.

For the longitudinal component, i.e. the first 24 months, time intervals were selected to preserve the longitudinal follow-up structure and avoid having multiple measurements of a given child within one age group. Note that for the longitudinal sample, age ranges were defined to correspond to specific visits, although visits did not always take place at the exact targeted age. For this reason, the constructed age group sample sizes were sometimes slightly different from the designed follow-up

visit sample sizes. Moreover, cross-sectional observations were added to the longitudinal sample between 18 and 24 months. In the cross-sectional data, it is possible that in a few cases more than one measurement from the same child occurs because of the multiple visits in Brazil and the USA, combined with the lower data density in this component. Similarly, it was impossible to break the sample into independent groups for the weight-for-length/height indicators. For this reason, the Q-test results required a conservative interpretation.

Overall, Q-test results were interpreted with caution and considered simultaneously with results of worm plots (van Buuren and Fredriks, 2001) which do not require any assumption and still offer very specific information about the goodness of fit for each group. The same age grouping was used as defined for the Q-test. Interpretation of results requires careful review of the shapes of the worms formed by a cubic polynomial (the red line in all worm plots) fitted to the points of the detrended Q-Q plots based on z-score values derived from the model being evaluated. A detrended Q-Q plot is presented for each age group. Confidence intervals (95%) are displayed for each of the worms (dotted curves in all worm plots). Table 7 summarizes the interpretation of various worm plot patterns. Flat worms indicate an adequate fit. The Q-test combined with the worm plot patterns provide a robust assessment of a model's goodness of fit, especially in terms of evaluating local fit.

Table 7 Interpretation of various patterns in the worm plot^a

Shape	Moment	If the worm	Then the
Intercept	Mean	passes above the origin,	fitted mean is too small.
		passes below the origin,	fitted mean is too large.
Slope	Variance	has a positive slope,	fitted variance is too small.
		has a negative slope,	fitted variance is too large.
Parabola	Skewness	has a U-shape,	fitted distribution is too skew to the left.
		has an inverted U-shape,	fitted distribution is too skew to the right.
S-curve	Kurtosis	has an S-shape on the left bent down,	tails of the fitted distribution are too light.
		has an S-shape on the left bent up,	tails of the fitted distribution are too heavy.

^a Reproduced from van Buuren and Fredriks (2001) with permission from © John Wiley & Sons Limited.

Pan and Cole (2004) proposed using a new tool to guide the choice of degrees of freedom for cubic splines fitting each of the parameter curves. They suggested plotting standardized Q-statistics against the number of age groups minus the corresponding degrees of freedom, for each of the L, M and S curves of the LMS method (Cole and Green, 1992). If fitting is adequate, the Q-statistics should be normally distributed with values within the range -2 to 2. This tool provides a global rather than a local test of significance and gives an accurate impression of the underlying goodness of fit because it does not depend on the precise choice of the number of groups. The proposed test is very useful for cross-sectional data where the choice of the number of groups can affect the Q-test results considerably. For example, points that are close in age but in opposite tails of the distribution generate opposing skewness when they fall into separate groups but cancel each other out when they are in the same group. This test was not implemented for the MGRS sample for two reasons. First, the largest number of observations was obtained in the study's longitudinal component, i.e. data were collected frequently at relatively well-defined ages from birth to 24 months. Second, splitting age intervals in a manner that failed to follow the study design, e.g. from birth to one month (which includes

measurements taken at birth, and at 7, 14 and 28 days) would group together four measurements per child, thereby reducing the reliability of the Q-test results.

(b) Selecting the best model across different classes of models

The search for the best model was done in an add-up stepwise form, starting from the simplest class of models comprising the age transformation, if any, and the fitting of the μ and σ curves, while keeping fixed $v=1$ and $\tau=2$ as described in section (a) above. The next step was to fit the v curve, fixing only $\tau=2$ and using the $df(\mu)$ and $df(\sigma)$ selected in the previous step. Once the best model within this class of models was selected, Q-test and worm plot results were evaluated to inform the decision on whether or not to select the more complex model. In a few cases when Q-test and worm plots were not sufficient to assess the improvement offered by the more complex model, comparison of observed and fitted percentiles was used to determine if differences were of clinical significance.

The fit of τ was considered only when Q-test or worm plots indicated misfit with respect to kurtosis. In this case, a third class of models was considered and comparison of observed against fitted percentiles was done to assess the improvement in the final curves. Among the rare cases where this occurred, fitting the fourth parameter always led to change that was negligible in practical terms. Therefore, all the models fitted had at most 3 non-fixed parameters (μ , σ and v).

With $df(v)$ thus selected (i.e. when v was not fixed to value 1), a new iteration was done to re-search for $df(\mu)$ and $df(\sigma)$. However, none of the additional iterations indicated any need to change either $df(\mu)$ or $df(\sigma)$. A further iteration was carried out to investigate if it was necessary to change the age-transformation power λ . This exercise did not lead to any changes in the selected models.

The methodology described above was used for all the indicators. Methodological aspects that are specific to the construction of each of the standards are described hereafter in relevant sections.

As part of the internal validation for each indicator, a detailed examination was made of the differences between empirical and the fitted centiles resulting from the selected model. Comparisons were also made between the observed and expected proportions of children with measurements below selected centiles across age (or length/height for weight-for-length/height) groups. For these two diagnostic tools, evidence of systematic patterns indicative of biases and the magnitude of deviations were examined.

Length/height-for-age, weight-for-age and BMI-for age curves were constructed using all available data (i.e. from birth to 71 months) but final age-based standards were truncated at 60 completed months to avoid the right-edge effect (Borghi et al., 2006). The weight-for-length standards go from 45 to 110 cm and weight-for-height from 65 to 120 cm.