**Methods for determining pubertal status in research studies: literature review and opinions of experts and adolescents**

I. V. Walker,a,b C. R. Smith,c J. H. Davies,d H. M. Inskipa,e and J. Bairda,e

a Medical Research Council (MRC) Lifecourse Epidemiology Unit, University of Southampton,

Southampton General Hospital, Southampton, UK

b Primary Care and Population Sciences, University of Southampton, Southampton General Hospital, Southampton, UK

c Department of General Paediatrics, Southampton Children's Hospital, Southampton, UK

d Department of Endocrinology, Southampton Children's Hospital, Southampton, UK

e National Institute for Health Research (NIHR) Southampton Biomedical Research Centre,

University of Southampton and University Hospital Southampton NHS Foundation Trust,

Southampton, UK

Address for correspondence:

Dr I. V. Walker, c/o Professor J. Baird, MRC Lifecourse Epidemiology Unit, University of Southampton, Southampton General Hospital, Southampton, SO16 6YD, UK

tel. +442380 777624

fax +442380 704021

email for correspondence: i.v.walker@soton.ac.uk

Short title: Pubertal status assessment

**Abstract**

In lifecourse studies that encompass the adolescent period, the assessment of pubertal status is important, but can be challenging. We aimed to identify current methods for pubertal assessment and assess their appropriateness for population-based research by combining a review of the literature with the views of experts in the field. We searched bibliographic databases, extracted data and assessed study quality to inform a workshop with 21 experts. Acceptability of different approaches was explored with a panel of ten adolescents. We screened 11,935 abstracts, assessed 157 articles and summarised results from 38 articles. Combining these with the opinions of experts, self-assessment was found to be a practical method for use in studies where agreement with the gold standard of clinical assessment by physical examination to within one Tanner stage was acceptable. Serial measures of height and foot size accurately indicated timing of the pubertal growth spurt and age at peak height velocity, and were seen as feasible within longitudinal studies. Hormonal and radiological methods did not offer a practical means of assessing pubertal status. Assessment of voice maturation was promising, but needed validation. Young people thought that self-assessment, foot size and voice assessments were acceptable, and preferred an assessor of the same sex for clinical assessment.This review thus informs researchers working in lifecourse and adolescent health, and identifies future directions in order to improve validity of the methods.

*228 words*

*Keywords:* puberty, adolescence, pubertal status, pubertal assessment, Tanner stages

**Introduction**

Much epidemiological research on the developmental origins of health and disease requires following study participants from early life to assess their growth and development. However, during adolescence, interpretation of a variety of measurements, such as those of body composition and mental health, requires knowledge of the participants’ stage of puberty at the time the assessments are made.

Accurate evaluation of puberty is key in the assessment of growth in young people. In a clinical setting, a physical examination of secondary sexual characteristics by a trained health professional (hereafter ‘clinical assessment’) is undertaken to document the stage of puberty,1, 2 as changes in the timing and tempo of puberty may indicate delayed or precocious puberty, or adverse effects of an underlying disease or its treatment. The different aspects of the clinical assessment identify the degree to which the body has been exposed to different hormones: testosterone through its effect on pubic and axillary hair, and penile size, oestrogen through breast development and menarche, and gonadotrophins through testicular and ovarian volume. Clinical assessment is widely recognised as the gold standard method for assessing pubertal development.3-5 Traditionally attributed to Tanner (1962), this method uses a series of reference photographs accompanied by brief descriptions, which depict five stages of development of pubic hair growth for both sexes, and breast and external genitalia development, for girls and boys respectively.6 This resource is sometimes known as the sexual maturation scale (SMS) or sexual maturity rating (SMR).

In a research setting it may be useful to understand the effect of exposures on the timing of puberty or the confounding effects of sex steroids on various outcomes. Study participants are often healthy, and the acceptability and practicality of a standard clinical pubertal assessment is less certain. The use of clinical assessments in cohorts and large-scale longitudinal studies can be challenging due to difficulties obtaining consent to a physical examination. Studies often rely on self-assessment methods or parental reporting, either as the main or a backup method when clinical assessment is declined. Self-assessment methods are also not without their challenges – for example, schools and parents may be concerned about the use of images depicting the development of secondary sexual characteristics.7

There are few reviews of the different approaches to pubertal assessment in a research setting.3, 4, 8 In order to identify current methods for pubertal assessment and assess their appropriateness for population-based research, we combined a review of literature with the views of experts in the field, as well as the views of young people on acceptability of various methods.

**Methods**

***Review of evidence***

We sought studies that described and compared methods of pubertal assessment, specifically searching for those that reported validation of the methods. In December 2015, assisted by an information specialist, we searched the following online databases of abstracts: Medline, PsycInfo, Scopus, Sociological Abstracts, CINAHL and ERIC. Search results were imported into EndNote X7, and duplicates were removed. In addition, we consulted via email, and in person, clinical and epidemiology experts in the field for recommendations of relevant articles, and hand-searched key journals to identify further publications. Titles and abstracts were screened to determine the relevance of studies. Where these contained insufficient information, full articles were accessed. While we did not impose limits on publication dates or geographical location of studies, we excluded those articles where the full text was not available in English, due to a lack of funding for translation.

Whilst we employed systematic review methodology in using a standardised form for data extraction and study quality assessment, we did not aim for an exhaustive review; methodology is often described briefly within research articles and search techniques do not necessarily identify all such articles, because of the indexing procedures. Consistent with the purpose of the review, which was to inform a workshop of experts, we gathered information on the use of the methods to the point of saturation, but did not attempt to include every study that has used the methods. Data were extracted from the included studies, using a standard proforma adapted for this review, by authors CS and IW, following which IW and HI double-checked half of the extracted records. The risk of bias of each included study was assessed using a quality assessment checklist adapted from guidance of best practice in systematic reviews.9 Results were collated in a table format and summarised in a narrative synthesis.

***Expert opinion***

The preliminary findings of the literature review were presented at a workshop at the MRC Lifecourse Epidemiology Unit, University of Southampton in September 2016, which was attended by 21 clinicians and epidemiologists from the Cohort and Longitudinal Studies Enhancement Resources (CLOSER) and Society for Social Medicine (SSM) networks. The workshop incorporated discussions of the approaches to pubertal assessment most suitable for research studies. The workshop was facilitated and chaired by two of the authors (JB and HI). Following a presentation of the review findings, presentations were also given by experts in specific approaches to pubertal assessment – clinical examination, growth assessment to infer pubertal stage, and hormonal approaches. Three break-out groups were then formed to consider two specific questions:

1. What cross-sectional approach is most suitable to determine current pubertal stage?
2. What pubertal features should we assess longitudinally (e.g. peak height velocity), and how should we measure them?

A snowballing technique was used to reach consensus. The discussions of the break-out groups were recorded on flip charts and fed back to all workshop participants. Further discussion took place within the group as a whole. JB took notes throughout the meeting and produced a report of the workshop, including a synopsis of each presentation and a summary of the discussion points that arose during the snowballing exercise. This draft report was circulated to all participants, and they were invited to comment. We also held additional discussions with experts in the field of adolescent medicine outside the workshop.

***The views of young people***

We consulted six boys aged 12 to 16 years and four girls aged 10 to 13 years, who were part of the University Hospital Southampton Children’s Panel, which is available for consultations relating to research and clinical care. Three female researchers spoke to the panellists in two sex-specific groups, gathering their views on the acceptability, preferences and barriers in relation to the following types of pubertal assessment: physical examination, self-assessment questionnaires, shoe and foot size assessments, and voice maturation assessment for boys. In addition, the boys were asked about the use of age at first nocturnal emission and the use of orchidometer to assess testicular size. The girls were asked about the use of age at menarche. We used examples of self-assessment questionnaires with line drawings representing Tanner stages, and let both the boys and the girls use Speechtest, an app that reports boys’ pubertal stage derived from hearing the participant counting backwards from 20 to 1.10

**Results**

We screened 11,935 abstracts and assessed 157 articles in detail. Of these, 38 were deemed relevant, with most articles reporting data comparing two or more methods of pubertal assessment. Not all studies were validation studies, since some did not include a comparison with the gold standard method of clinical assessment.

The majority of studies had methodological weaknesses. Most did not include justification of the sample size. Selection bias was an issue in a number of studies with highly selective samples consisting of, for example, children from particular socio-economic backgrounds or those attending specialist clinics, where these substantially differed from the intended target populations. Many articles included insufficient details of the methods used. Eleven studies were conducted in the United States, nine in the United Kingdom, four in Sweden, two in Australia and two in Turkey, and one each from Chile, China, the Czech Republic, Denmark, France, India, Iran, Japan, the Netherlands and South Africa. Review findings were grouped into five categories according to the pubertal assessment method.

***Self-assessment***

This was the largest group, with 17 relevant studies (table 1). Most compared self-assessment using either the SMS or the pubertal development scale (PDS) against clinical assessment, with the PDS being an interview-based measure, consisting of questions about body changes and growth. Some studies used the photo version of the SMS, while others11, 12 used a line drawing version,13 in order to increase acceptability of the images to young people.

Agreement in Tanner stage between self-assessment and clinical assessment ranged from 43% to 81% of the sample. There was generally at least 84% agreement to within one pubertal stage for the SMS,5 and 85-100% for the PDS.5, 14, 15 Children earlier on in their development were reported to overestimate their stage of development, and children in the later stages were often found to underestimate this,14-17 although the opposite pattern was also observed.18 Girls tended to be more accurate in self-reporting than boys.12, 13, 16 Certain aspects of development were self-reported with more accuracy than others; for example, self-assessment of pubic hair development tended to correlate well with clinical assessment for both sexes, whereas self-reported breast and external genitalia stages were more weakly correlated with clinical staging.12, 15, 19

Self-assessment using the SMS perfectly agreed with that based on the PDS in 56% of females and 39% of males,20 but agreement to within one Tanner stage was observed in 97% and 89% respectively. Agreement between the two scales was also higher for 12-13 year olds compared with 10-11 year olds or 14-15 year olds. However, when compared with clinical assessment, boys tended to rate themselves as more mature using the SMS drawings than using the PDS score, and girls tended to rate themselves as less mature. This may suggest that viewing the images of pubertal stage might encourage a socially desirable response among young people.20 Colour drawings used in a clinic setting yielded assessments by children that were close to those of raters. Assessments were less reliable, however, when children were overweight or obese,12, 21, 22 although not all studies observed this.5, 21

When children were asked to compare their development with their classmates in a ‘global question’, taking into account all aspects of puberty, there was high concordance with answers to the same question provided by clinical examiners, with 95% for boys and 93.5% for girls, although this was in a self-selected subsample of those who volunteered to undergo clinical examination.23 A continuous Tanner visual analogue scale, used in one study, was found to be somewhat less accurate than the SMS or PDS.5

Four studies explored the use of parental assessments of own child’s pubertal status. One of the studies used a single question as to whether a particular child had entered puberty; however, parents were asked the question one year after the children, which limits interpretation of these findings.24 In another study, mothers of girls rated their daughters’ SMS-based Tanner stage higher than clinicians, but correlation with clinical assessment was 0.85, compared with 0.82 for self-assessment, indicating a potential for the use of this method.14 In the same study, mothers were more likely to overestimate breast staging at the beginning of breast development, but this was not found in relation to pubic hair.14 In a large study, mothers and children tended to underestimate pubertal state for girls and to overestimate it for boys.18 In another study, comparison of girls’ and mothers’ assessment of breast bud development using black and white Tanner photos with a clinical assessment showed that self-assessments by the girls themselves had low concordance with breast bud assessment by a trained nutritionist.25 Maternal assessment of breast bud development among the leaner girls was more accurate than that by the girls themselves.

While the experts in our workshop saw clinical assessment as the gold standard, they recognised that acceptability among study participants was often low. They considered self-assessment to be the easiest approach within cohort studies and the only practical method, other than clinical assessment, in cross-sectional studies. They agreed, however, that it was a rather crude approach, given the likelihood that it would only accurately categorise pubertal status to within one Tanner stage.

***Growth***

We identified nine relevant studies of growth (table 2). These generally focused on two parameters: age at height growth take-off, indicating the start of pubertal growth spurt, and age at peak height velocity, indicating the intensity of the pubertal growth spurt. Serial measurement of height was the most frequently reported approach, which can also help identify the pre-pubertal growth spurt.26 Height velocity may correlate with secondary sexual characteristics, with one study demonstrating high correlation between height velocity and testicular volume.27

One approach to growth curve analysis is the SuperImposition by Translation And Rotation (SITAR) model.28 It takes into account characteristics that differ from one individual to the next – namely mean height, timing and rate of puberty. The SITAR method produces three measurements, representing differences in mean size and growth tempo, and a measure of growth velocity. The method also could be applied to growth in other parameters, such as foot length, and to the development of secondary sexual characteristics, such as breast, testicular and pubic hair development. A study of the Edinburgh Longitudinal Growth Study cohort that used the SITAR method showed high correlations in relation to pubertal timing between such markers as height, genital and pubic hair stages, and testicular volume , although correlations for the shared markers were significantly higher in the girls.29 This method recently has been applied to the height measurements from children in the Avon Longitudinal Study of Parents and Children (ALSPAC) to enrich the dataset with measures that include age at peak growth velocity,30 a marker of pubertal timing. A rapid increase in foot length may correspond to the onset of puberty and Tanner stage 2 on clinical assessment.31 Peak increase in shoe size has also been found to precede peak increase in sitting height,32 although the comparison was with data from an unrelated sample. Others compared mean age at increase in foot velocity with mean age at take-off in height, and found no significant difference between the two parameters.33

The experts attending the workshop agreed that assessment of growth in height or foot size was a promising method, albeit that more evidence was required in relation to foot size. Growth assessment was thought to be feasible within cohort studies, provided measurements could take place with sufficient frequency, but by definition not suitable for cross-sectional studies.

Radiological methods of assessing the degree of maturation of the cervical vertebra, olecranon and digits have been proposed as pubertal assessment approaches. The Greulich and Pyle Atlas34 and Tanner-Whitehouse III score35 can be used to assess skeletal age. A cervical-vertebral index indicates level of skeletal maturation,36 and radiological images of the olecranon correlate with those of the digits, yet there was no comparison with a reference standard in these studies.37, 38 The experts agreed that radiological approaches were feasible in large-scale population-based studies, though required participants’ attendance at a clinic, limiting their use.

***Hormonal assessment***

*Gonadotrophins and gonadal hormones*

A number of articles discussed the potential for measurement of gonadotrophins as a means of assessing pubertal stage. At the onset of puberty, there is an increase in the overnight pulsatile release of luteinising hormone (LH), suggesting that early morning urinary LH might offer a way of determining Tanner stage in girls.39, 40 Serum testosterone is aromatised to oestrogen in fat, and it is oestrogen that triggers growth hormone (GH) secretion in both sexes.41 Inhibin B is released from the ovary in pubertal girls and rises in early puberty, whereas inhibin A is slower to rise.42 For boys, testosterone rises throughout puberty, with the steepest rise seen between Tanner stages 3 and 4.

We identified five studies that incorporated both hormonal measurement and reference to pubertal staging (table 3). In a study examining the correlation of three-monthly urinary oestradiol, testosterone and LH with self-reported SMS-based Tanner staging, the levels of all three hormones were correlated with the staging at baseline and 12 months later.43 In another study, staging based on physical examination reflected testosterone and DHEA levels in both sexes, but was only modestly related to oestradiol in girls.44 In boys, 24-hour testosterone levels in later puberty correlate reasonably well with testicular volume. Serum concentrations of testosterone increased progressively throughout puberty, with a marked increase occurring between early and mid-puberty. The onset of puberty was marked by accentuation of the diurnal rhythm of testosterone release due to increased release of testosterone at night.45 However, assays of urinary sex hormones both in boys and girls can be difficult to interpret due to within-person variability, and longitudinal sampling would be required to determine the hormonal changes associated with progression through pubertal stages.43, 46 It has been suggested that both inhibin A and B may have a potential as markers of pubertal development in boys and girls.46, 47 However, progressive changes in their levels are not consistent enough to enable pubertal staging. On the whole, there are significant challenges to interpretation of such data due to the complexity of relationships between gonadotrophins, sex hormones and inhibins.46, 47

*Leptin*

Leptin interacts with the reproductive axis at multiple sites with stimulatory effects on the hypothalamus and pituitary, and inhibitory action on the gonads.48 Leptin may affect the regulation of gonadotrophin-releasing hormone (GnRH) and LH secretion during puberty.49 There were five studies identified in this group (table 4). One study reported an increase of 50% in serum leptin levels just before the onset of puberty, and a decrease to approximately baseline after the initiation of puberty.50 However, these findings were based on a small sample and should be treated with caution, given that another study showed that leptin levels varied widely amongst schoolchildren.51 Another study showed no correlation between serum leptin levels and age in boys, but a significant correlation in girls.52

Others have examined urinary leptin, demonstrating that monthly urinary leptin levels were higher in girls than boys over a period of six months.53 Leptin was higher in children advanced in puberty, compared with children remaining prepubertal, but the measure of pubertal status was insufficiently described. Urine collection was a less invasive method for measuring leptin, compared with blood. In another study, urinary leptin showed day-to-day variability and correlated with serum leptin. Urinary leptin was similar in both sexes: in boys it increased significantly from Tanner stage 1 to stage 2, peaked in stage 3 and then declined for stage 4 and 5, while in girls there was a linear relationship between leptin levels and pubertal development.54

The experts in the workshop regarded assay of sex hormones and leptin as an area of interest for assessment of pubertal stage, but the need for repeated measurements and the potential cost of assays were seen as barriers to their use in research.

***Voice maturation assessment***

The maturation of the human voice is characterised by changes in pitch, loudness and a variety of tone qualities as the larynx grows in both sexes.55 Voice breaking in boys usually occurs as a distinct event during late puberty, with a rapid drop in voice occurring during Tanner stages 3 and 4, usually around 12-15 years.56, 57 One study has shown that the timing of voice breaking could act as a non-invasive marker of pubertal maturation, with moderate correlation with other markers, such as genital development, and pubic and axillary hair growth.57

Cooksey classification of voice is based on a six-stage pattern of pubertal voice development derived from the singing range in boys, and this method has been validated previously.58 A clear correlation was found between the Tanner stages and Cooksey classification of boys’ voices studied at three-month intervals. In addition, change in fundamental voice frequency was correlated with testicular volume, but not with serum testosterone levels.55 The experts saw assessment of voice maturation as a promising method, and acknowledged the need for further research in this area.

***The views of young people***

The panel of young people expressed a preference for questionnaires, rather than clinical assessment. They were shown line drawings representing pubertal stages, and said that they preferred these to the photographs. They also preferred paper, rather than digital, versions of questionnaires, and thought that measurements of height and foot size were acceptable and likely to be so for their peers. All stated that the voice maturation assessment via a Speechtest app was acceptable, and boys were unlikely to exaggerate the depth of their voice, provided the assessment was done in private. The boys felt that they would find a question about their first nocturnal emissions embarrassing and that most would not be able to recall the timing of this correctly. They also stated that they would need to self-examine in private in order to estimate their testicular size if asked to use an orchidometer. The girls did not object to being asked about age at menarche if the person asking was a professional, but felt that their mothers might be more accurate in their reporting of this. Both boys and girls suggested that young people would be much more likely to consent to a clinical assessment if this were carried out by a professional of the same sex. They noted the importance of clear communication that the assessment was brief and did not require complete removal of their clothes.

***Integration of review findings with the views of experts and young people***

As a result of the literature review it was possible to draw a number of conclusions for each category of assessment. These are summarised in table 6, alongside conclusions drawn from the discussions with experts and adolescents.

**Discussion**

This review examined the main pubertal assessment methods that are currently in use in research. Clinical assessment can be used across all Tanner stages, from the pre-pubertal stage to complete maturation, in contrast with some of the other methods, which may be more useful in relation to specific changes through puberty, such as the pubertal growth spurt, specific hormonal changes or voice maturation. Physical examination is non-invasive compared with blood sampling required for many hormonal methods, and less harmful than radiological methods. Clinical assessment can also be used in cross-sectional studies, in contrast to some other methods. However, our work suggests low acceptability of the clinical examination in a research setting. The young people suggested that good communication, and sex-matching of the assessor and participant, could help improve acceptability of the method. However, it is possible that even when an assessor of the same sex is not available, this may not necessarily lead to refusal.59

The other methods considered in this review may have their place in pubertal assessment under different circumstances. The self-assessment approach may be suitable in studies where accuracy to within one Tanner stage is acceptable, in large scale studies lacking resources required to facilitate clinical assessment or in settings where there may be strong objections to physical examination. Self-reporting may be affected by social desirability or accepted norms,11, 14 and younger children may consider themselves older (and therefore, more developed) than they are, whereas older children might view themselves as younger, hoping for more development.17 Higher validity against the reference test of clinical assessment was observed when adolescents were allowed to self-examine before rating their own development, compared with self-recall from memory.21 Assessments of growth could be used in longitudinal studies, provided appropriately frequent measurements were feasible. The SITAR method in particular has the advantage over conventional approaches, in that it takes account of individual characteristics, such as mean height, timing and rate of puberty. The SITAR method could be applied to growth in other parameters, such as foot length, and to the development of secondary sexual characteristics. Assessing foot size is also a promising method. The absence of data on validity, along with the acceptability issues, mean that radiological methods of pubertal assessment are unlikely to be appropriate for use in research studies. Measurements of gonadotrophins, gonadal hormones and leptin can provide a detailed picture of the biological changes taking place during puberty. High laboratory-associated costs, the need for repeat measurements and the invasiveness of blood tests are some of the current barriers to the wider use of the hormonal methods. Between- and within-person variability and the lack of evidence of validity against clinical assessment also limit the utility of the hormonal methods. Nonetheless, advances in this field might yield acceptable and reliable methods in the future. Assessment of voice maturation presents a convenient, albeit insufficiently validated, method.

Clinical assessment remains a subjective method, prone to measurement error and bias, particularly if not diligently conducted. The most frequently used black-and-white reference photographs are old, and depict Caucasian adolescents only. This has been highlighted previously,4 and using more up-to-date resources would strengthen the method. Inter-rater variability could be improved by thorough training and rigorous protocols within research studies. It would be valuable to investigate the effect on rates of consent of improved communication and sex-matching of the clinical assessor and participant. More work is needed on the use of parental assessments. Studies of growth employing the SITAR method for parameters other than height would be of great value, in particular in relation to foot size. In addition, it would be helpful to determine whether there is scope for measuring foot growth using repeated self-administered questionnaires incorporating shoe size. There is a need for further research to assess the validity of hormonal assays as pubertal assessment methods, especially with regards to frequency of measurements and the use of less invasive methods, such as hair or urine sampling. Voice maturation assessment is another area where further evidence on validity of the method is needed, and development and evaluation of apps such as Speechtest might be useful, not least as they seemed acceptable to young people.

Previous reviews of methods for pubertal assessment agreed that clinical assessment constituted the gold standard method. Coleman and Coleman concluded that self- and parental assessment methods had lower validity than clinical assessment.3 An extensive review by Dorn et al. covered a range of objective and subjective methods and discussed important issues associated with each, proposing an argument that any methods could be used, provided they are appropriate for addressing the research questions in a study.8 Dorn and Biro highlighted the difficulties obtaining consent for clinical assessment and the fact that self-assessment is prone to significant bias and insufficient agreement with clinical assessment, whereas hormonal methods do not lead to adequate Tanner staging.4

One important consideration is that in research, accuracy implies proximity of results to the true value. Stages of puberty are artificial constructs, created in order to help to interpret the influence of the timing and tempo of puberty on health in adolescence and beyond. They are the best proxy we have for the ‘true value’ in pubertal development. Staging is an integral part of clinical assessment, which thus can be seen as the most accurate method, in contrast to, for example, growth, hormonal or voice methods, which sometimes attempt to arrive at Tanner stages indirectly. Depending on the method, within- and between-person variability in the parameters can affect the accuracy of estimating corresponding pubertal development stages. Furthermore, Tanner stages are well known and widely used by clinicians and researchers. It should therefore be borne in mind that the use of some other methods, such as the PDS, global question or a visual analogue method in self-assessment might complicate interpretation of study results and their use in comparative analysis and meta-analysis with studies that use Tanner staging.

In this review, we employed methodology that is conventionally used in systematic reviews, including extensive searches of the published literature, strengthened by input from an information specialist, careful screening of potentially relevant abstracts and papers, and detailed data extraction and quality assessment of included studies. This work adds to the existing reviews, incorporating the more recent studies and capturing methodologies not previously included, yet avoiding some of the areas already extensively discussed. We complemented our review with the opinions of experts, with whom the review findings were discussed. We found that there was considerable consistency between the views of the experts and overall conclusions emerging from the literature. Another strength of this work was the incorporation of the views of young people. This allowed triangulation with the review findings, which strengthened our interpretation.

Our review does not cover all research on approaches to pubertal assessment, given that our intention was to focus on methods that might be used in population-based research studies. We searched to saturation to identify relevant methods, but did not attempt an exhaustive review of all studies relating to each of the assessment methods. We did not search the grey literature or contact experts in order to identify relevant unpublished studies, hence publication bias is likely. Given that abstract screening was conducted by a single reviewer, it is possible that not all relevant published studies were identified. There were few validation studies. Descriptions of the methods in some articles lacked detail, presenting challenges for study quality assessment.

**Conclusions**

This review, complemented by the views of experts and young people, highlighted strengths and limitations of self-assessment methods in research studies, compared with the gold standard of clinical assessment of pubertal development. It has confirmed the barriers to the use of hormonal and radiological methods in this setting, and identified the need for further research into the validity of the promising growth assessment methods, such as foot size measurement, as well as voice maturation assessment. Improved assessment methods would enhance studies examining growth and development through adolescence.

**Acknowledgements**

We are grateful to Elizabeth Payne (Information Specialist, UK) for her assistance with literature searches, and to Tina Horsfall and Julia Hammond (MRC Lifecourse Epidemiology Unit, University of Southampton, UK) for their input in organising and running the University Hospital Southampton Children’s Panel discussions. Special thanks go to the young people for expressing their views on the methods. We are also grateful to Professor David Dunger (University of Cambridge, UK), Professor Tim Cole and Professor Russell Viner (both of University College London Great Ormond Street Institute of Child Health, UK) for their advice on the topic and presentations at the expert workshop, as well as to the workshop participants for their contributions.

**Financial Support**

The review of evidence and the work with young people were supported by Cohort and Longitudinal Studies Enhancement Resources (CLOSER), UK (grant reference ES/K000357/1), and the workshop with experts was supported both by CLOSER and the Society for Social Medicine (SSM), UK. Hazel Inskip is supported by the UK Medical Research Council and the NIHR Southampton Biomedical Research Centre, and her work receives funding from the European Union’s Horizon 2020 research and innovation programme under grant agreements no. 733206 (LifeCycle).

**Conflicts of Interest**

None

*(5078 words)*

Table 1: Description of studies using self-assessment methods

| First author, country, publication year | Study sample | Study focus and details | Validation or correlation, and related measures | Key findings and conclusions |
| --- | --- | --- | --- | --- |
| Rasmussen,18 Denmark, 2015 | 898 children (aged 5.6-14.9) and 1173 parents, based on data from an ongoing mother-child cohort | Validity of self-assessments and parental assessments of pubertal development stage using a questionnaire with drawings.  First examination: clinical and parental assessments; second examination: clinical and self-assessments. Breast palpation in addition to visual assessment for clinical breast staging. Parents not asked about sons’ genital stage. | Against blinded clinical assessment by one of six trained physicians  Percentage agreement, kappa and Kendall's correlation | Large study, white Danish mothers only, most from the highest two social classes (out of five). Majority in early stages of development.  Self-assessments and parental assessments inaccurate in a large proportion compared with clinical assessment. Children slightly more accurate than parents in identifying puberty onset. No significant effect of BMI. Half of the girls underestimated breast stage, and a quarter underestimated pubic hair. One third of the boys overestimated genital or pubic hair stage. Interrater variation for breast stage k = 0.78, but 100% agreement for onset of breast development, possibly because of palpation. Conclusion: Self-assessment and parental assessment may be unreliable, with girls tending to underestimate and boys overestimate, as previously shown. |
| Lum,24 UK, 2015 | Children aged 8-11 years in a multi-ethnic school in London, and their parents  246 parent-child pairs  Part of the Size and Lung function In Children (SLIC) study | Feasibility of assessing secondary sex characteristics as a proxy for pubertal status. Cross-sectional survey of children.  Exploration of ethnic differences in rates of pubertal attainment  Illustrated Tanner questionnaire, with a yes/no question on secondary sexual characteristics  Also answered by parents 12 months later | No clinical assessment  Child self-report compared with parental report  Percentage agreement and logistic regression | Comparing two different methods applied one year apart limits interpretation.  Agreement between children and parents of at least 68%. Overestimation by self-report compared to parents in 17% and underestimation in 15%  25% girls and 62% boys unsure of some aspects of their pubertal development.  Children of Black African origin more likely to have attained puberty at any of the ages than other ethnic groups. |
| Conclusion: Parental assessment may be more reliable than self-assessment |
| Pereira,25 Chile, 2014 | 481 girls, aged 6.8-9.2, and 341 mothers, within Growth and Obesity Chilean Cohort Study | Inspection and palpation of breasts by an assessor followed by visual self-assessment and parental assessment against standard black and white Tanner photos  BMI measured and z-score derived. Mothers’ education classified as less or more than eight years of formal education | Against blinded clinical assessment by a nutritionist trained in breast bud detection  Kappa, Spearman’s correlation | Low and middle class, three-way blinding. 43% girls overweight or obese. Poor concordance between self-assessment and clinical assessment (k=0.02), and good between mothers and clinical assessment (k=0.73). No significant differences by BMI in self-assessment, but mother of leaner girls one third less likely to overdiagnose breast bud development. No significant difference by mothers’ education.  Conclusion: Mothers, but not daughters, may be good assessors of breast bud development, irrespective of mothers’ education |
| Rabbani,21 Iran, 2013 | 190 boys aged 8-16 years from three contrasting regions | Comparison of Persian Tanner stages self-assessment questionnaire with physical examination | Against clinical assessment  Weighted kappa | Substantial agreement between self-assessment and clinical assessment. The boys were able to self-examine before completing questionnaires.  Complete agreement overall in 72% of cases, but varying depending on pubertal stage: 46% agreement in stage 3, 89% in stage 4 and 85% in stage 5 |
| Conclusion: Self-assessment can be used in Iranian populations for children in later stage of pubertal development. |
| Sun,22 China, 2012 | 9132 girls and 6924 boys, aged 7.9-18.9 years, from eight research sites, obese and normal BMI | Self-assessment using realistic colour drawings and descriptions proposed by Carel and Leger60, compared to physical assessment. Palpation used for breast and genital assessment.  Measured BMI classified as per Working Group on Obesity in China reference. | Against blinded clinical assessment by two trained examiners, same day as self-assessment  Kappa, Kendall’s rank correlation | Very large study, wide age range, Han ethnicity, urban and rural, representative of mainland China. Underweight children excluded from analysis.  Most children were able to self-assess accurately or close to raters. Obese children assessed less accurately than normal weight children. Weighted Kendall rank correlations between raters and children ranged from moderate to substantial (0.503-0.642), with the lowest for genital assessment in obese boys (k=0.352). Conclusion: Self-assessment using realistic colour images may yield high correlations with clinical examination, but are less reliable in obese children, especially boys. |
| Norris,11 South Africa, 2008 | 182 Black African participants (49% female) aged 10-18 years, urban setting | Comparison of self-assessment according to pubertal development scale (PDS) with clinical assessment using the sexual maturity scale (SMS).  The PDS is an interview-based continuous measure, involving a series of assessor-administered questions about growth, body hair and skin changes, with questions on facial hair and voice change for boys and breast development and menarche for girls.  Four response options: ‘not yet started’, ‘barely started’, ‘definitely started’ and ‘seems complete’. No pictures or diagrams. | Against clinical assessment  Kappa, Pearson’s and Kendall’s statistics | Self-assessment via a member of staff because of concerns over literacy and understanding. Females: 56% agreement between the PDS and the SMS, 28% underestimated and 16% overestimated  Males: 26% agreement between PDS and the SMS, 70% underestimated and 4% overestimatedConclusion: Self-assessment is less reliable in multi-ethnic community-based research. This differs from outcomes in studies of Caucasian populations. There is no shortcut to reliable pubertal assessment. |
| Bond,20 Australia, 2006 | 1486 girls and 1378 boys, aged 9-16 years, from primary and secondary schools, multi-centre | Comparison of two self-assessment measures: SMS and PDS  Proportion of missing data was considered to be an indicator of acceptability of the methods | No clinical assessment  SMS compared with PDS  Percent agreement, weighted kappa | Perfect agreement between the measures in 56% of girls and 39% of boys. 97% of girls and 89% of boys had concordance to within one category. Categorisation into early, middle and late puberty improved the agreement. Conclusion: Due to better response rate and acceptability in schools, the use of PDS is to be preferred for self-assessment. |
| Desmangles,17 USA, 2006 | 240 children aged 6-16, multicentre, urban setting | Self-assessed SMS compared with physician assessment  Pubic hair was used as surrogate marker of overall pubertal development in boys, i.e. not completely following the SMS | Against clinical assessment  Kendall correlation coefficient, kappa | 60% agreement between self-assessment and clinical rating for breast development. Of the 40% not in agreement, 48% overestimated and 52% underestimated.  Higher correlation at 77% for pubic hair development, with 61% of boys assessing pubic hair development correctly. Of the remainder, 73% overestimated and 27% underestimated. Higher degree of concordance amongst boys than other studies.Conclusion: Self-assessment did not appear reliable as a method of assessing pubertal development |
| Schmitz,5 USA, 2004 | Study 1: convenience sample of 178 children aged 8 to 18 years attending a diabetes and endocrine clinic  Study 2: 125 children aged 10 to 13 years recruited from an osteoporosis prevention study | Four measures: PDS, SMS, attenuated three-point Tanner and Tanner visual analogue scale. In the latter, to indicate perception of progress through puberty, participants drew a line on paper between the two extreme Tanner stages for each secondary sexual characteristic. The scale was scored as a continuous measure in 1/16-inch increments with a ruler.  Different versions of the self-assessment methods were used for the two study groups.  Study 1 children had a physician assessment; Study 2 children had a dual-energy X-ray absorptiometry (DXA) scan. 29 Study 1 participants repeated all four self-report measures three months later. | Against clinical assessment in study 1, against DXA bone mineral density in study 2  Pearson’s correlation coefficient, kappa, percentage agreement | Multiple comparisons between all measurements. Correlations calculated between pubertal measures and bone mineral density (all between 0.45 and 0.65). Complex design with every scale compared with every other one.  More than 85% agreement within one stage was obtained for all measures with the exception of breast assessment on the visual analogue scale.  None of the self-assessment measures were statistically different from one another in the proportion of agreement with physician rating. High correlations with physician rating (apart from the PDS), but low kappas (all below 0.5). Conclusion: The SMS self-assessment is possibly a more reliable marker of pubertal development than the PDS scale. |
| Bonat,12 USA, 2002 | 244 non-obese and obese healthy children, aged 6-12 years | Determine the reliability of Tanner stage self-assessments in both non-obese and obese children.  A standardised series of drawing with explanatory text for self-assessment. Breast and pubic hair for girls, pubic hair only for boys. Physical examination by a trained and experienced nurse or paediatrician. | Against clinical assessment  Kendall rank correlations, 1-sample sign test | Under-researched area, but limitations of selection bias and a young age range in the sample.  Interrater reliability 100% for breast and 98% for pubic hair due to substantial training. Obese girls were more likely to overestimate breast stage than non-obese. Non-obese girls were not any more likely to over- or underestimate breast stage, and both obese and non-obese girls were not any more likely to over- or underestimate pubic hair stage. Both obese and non-obese boys were more likely to overestimate pubic hair stage. Majority of overestimations were in Tanner stage 1 or 2. Conclusion: Self-assessment of breast Tanner stage in young non-obese girls, and of pubic hair in all girls, is a reasonable substitute for a physical examination. |
| Taylor,61 UK, 2001 | 62 boys and 41 girls attending a paediatric endocrinology clinic | New self-administered SMS questionnaire compared with clinician assessment  Sequentially recruited children | Against clinical assessment  Percentage agreement, weighted kappa | For pubic hair distribution agreement to within one Tanner stage in 88% of children. For female breast and male genitalia stage agreement to within one Tanner stage for 76% of children. Tendency to underestimate own level of development. Conclusion: Overall good agreement between SMS and clinical assessment for pubic hair, and moderate agreement for breast/genitalia stage. |
| Hergenroeder,15 USA, 1999 | 107 healthy girls aged 8-17 years | To evaluate interobserver reliability of physician assessments of pubertal maturation in girls in a multi-ethnic sample. Next, to evaluate the validity of self-assessment compared to physician assessments of pubertal maturation.  Assessment by one of two physicians in adolescent medicine, self-assessment using drawings and descriptions of five stages of development according to Tanner. | Against clinical assessment  Kappa | Small numbers of ethnic minority girls, which were excluded from some of the analyses.  Interobserver reliability: 84% for pubic hair and 76% for breast stage. Self-assessment: tendency to overestimate pubic hair development and no consistent trend for breast stage. Validity of self-assessment versus physical examination: kappa 0.34 for breast and 0.37 for pubic hair, i.e. marginal.  Conclusion: Self-assessment is not valid and interobserver reliability was low. |
| Berg-Kelly,23 Sweden, 1997 | 4516 young people, multi-centre | Cross-sectional study of three different school year groups (average age 13.5, 15.5 and 17.5 years).  A global question: ‘Considering your bodily development, how do you rate yourself compared to your classmates: very late, somewhat late, similar, somewhat early or very early?’  Global question was compared with the Q90 questionnaire – a set of questions about pubertal development, based on Tanner stages and other aspects of growth.  100 volunteers also had clinical assessment | Comparison of the categories of the global question (excluding ‘similar’) was made with self-assessment.  Validation against clinical assessment in a subgroup  Percentage agreement, sign test | Large study. Small subgroup validated, with selection bias. Little evidence of validation of the Q90 questionnaire.  Two-thirds rated themselves as average in development, similar across age ranges. Boys under-rated pubic hair compared with clinical assessment.  Agreement between global question and clinician rating was 94%. The question is simple to apply, has universal application and does not require clinical assessment.Conclusion: The study suggests that the global question has potential as a measure of pubertal development, although it does not provide an indication of time point within puberty. |
| Carskadon,16 USA, 1993 | 323 boys, 375 girls, their parents and teachers | A self-assessment scale, suitable for the classroom, adapted from the PDS, without use of pictures or the need for an interview  A parallel teacher and parent form of the scale was also developed | Comparison between assessments by children, parents and teachers  Against clinical assessment in a pilot study with 38 students  Spearman’s correlation coefficient, Cronbach’s alpha | Children who returned the questionnaire were more socially mature and more able academically. 43% boys and 46% of girls had missing data and were not included in the study.  Correlation with clinical assessment in the pilot r~=0.85. Significant correlations were found between parents and students for all the measures for 6th graders and 5th grade girls and several measures for 6th grade boys. Correlations between parents and children were better for girls. Teachers’ ratings had the lowest correlations with student and parent ratings. Conclusion: The scales worked better for older children, particularly for boys. May be useful where direct examinations, interviews or images not possible. |
| Petersen,7 USA, 1988 | Two successive birth cohort samples, Caucasian, middle and upper class, data on 252 children | To assess a PDS questionnaire with height measurements.  Twice annual assessments, over three years. Completed the same self-assessment according to the PDS. Interview-based assessment. | Not validated, but a subset had a comparison with growth measurements Correlations of peak height velocity with scores from the PDS  Alpha coefficients (assumed to be Cronbach’s alpha). | 335 boys and girls, data on 252 subjects. Correlations between age at peak height velocity and PDS scores at each age are all negative and hard to interpret. The longitudinal data indicated that assessment with the PDS reflected the sequence of pubertal Tanner staging.  10% of girls and 6% of boys decreased their level of development. The highest amount of regression was observed with skin changes and growth. Of note acne is a temporary phase and therefore regression may be expected.  Alpha coefficients ranged between 0.68 and 0.83 for the different ages of measurement, indicating good internal consistency.Conclusion: A self-report measure may be useful if the researcher is interested in an approximate assessment of pubertal status. |
| Brooks-Gunn,14 USA, 1987 | 151 girls aged 11-13 years recruited from a large cross-sectional study of female development | Self-completed PDS and SMS, and a maternal assessment using the SMS | Against clinical assessment. Correlations, alphacoefficient | The study considers the potential influence of social desirability and the existence of a set of norms.  The correlation with physician ratings was 0.82 for the SMS self-ratings, 0.85 for ratings with the mother, and 0.64 for self-reports on the PDS. Conclusion: The PDS was less valid than the SMS for self-assessment of pubertal development. Mothers could rate their daughters’ maturation well. The PDS may be useful in situations where the use of pictures is unacceptable. |
| Duke,62 USA, 1980 | 43 girls aged 9 to 17, 23 boys aged 11 to 18 | To demonstrate that adolescents can accurately assess their own developmental stage according to Tanners standard photographs.  Comparison of SMS by self-assessment with that by physician examination. | Against clinical assessment  Kappa | Agreement with the physician rating for breast development was 86% (kappa 0.81), female pubic hair 93% (kappa 0.91) and male genital stage 91% (kappa 0.88).Conclusion: Teenagers could accurately designate their own level of sexual maturation. |

Table 2: Description of studies using growth methods

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| First author, country, publication year | Study sample | Study focus and details | Validation or correlation, and related measures | Key findings and conclusions |
| Canavese,37 France, 2014 | 44 boys and 78 girls in an orthopaedic clinic | To compare the accuracy of the simplified olecranon and digital methods during the pubertal growth spurt and investigate this in relation to the experience of the assessor.  X-ray radiographs of the hand and elbow to determine skeletal age and correlate this with the stage of pubertal growth.  These were taken during the follow-up of children whose standing height increased by more than 4cm in the preceding six months. | No clinical assessment of pubertal status.  Correlation between the two methods.  Interclass correlation coefficient | More girls than boys; decision to take radiographs made by the surgeon; method based on subjective judgement of images. No comparison to a reference standard test in a study attempting to compare accuracy. Did not include pubertal development staging or sitting and standing heights. Strong correlation between nine observers doing two reviews.  Correlation between the two methods was strong with r=0.84 for boys and 0.83 for girls. The pubertal growth spurt occurred at 11-13 years of skeletal age in girls and 13-15 years in boys.Conclusion: The olecranon method offered detailed information during the pubertal growth spurt. The digital method was as accurate, but less detailed, hence more useful after the growth spurt, once the olecranon has ossified. |
| Cole,29 UK, 2014 | 103 boys and 74 girls from Edinburgh Longitudinal Growth Study | To estimate and compare pubertal growth timing and intensity across several markers, namely height, Tanner stage (pubic hair, and breast and genital stage for girls and boys respectively), testis volume and menarche. A ‘toolbox of statistical methods, including SITAR’ is described, which efficiently estimate peak velocity (PV) and age at peak velocity (APV) for each marker, in individual subjects and overall. | Correlation of the markers in terms of their timing (as indicated by APV) and intensity (as indicated by PV), with timing and intensity estimated for each marker and then compared across markers within individuals, resulting in a matrix of six correlations between four markers | APV for Tanner stages corresponded to the mean age at Tanner stage 3. On the basis of standard deviations of PV, it appeared that greater measurement error was likely for the Tanner stages, than for height or testis volume. For height and testis volume, but not for Tanner stages, early maturers tended to have a higher peak velocity. In relation to pubertal timing, the correlations between the markers of height, genital and pubic hair stages, and testicular volume were high in boys (mean of 0.76), and significantly higher in girls (mean 0.87). For intensity the correlations were much lower, 0.1-0.6 in boys and close to 0 in girls.  Conclusion: This logistic growth model can be used for serial changes in Tanner stage and testis volume through puberty, allowing timing and intensity to be estimated for height and secondary sexual characteristics, and then compared across markers within individuals. |
| Busscher,32 Netherlands, 2011 | Shoe shop data, 242 girls and 104 boys in early puberty | To describe the increase in shoe size during adolescence and determine whether the timing of the peak increase could be an early indicator of the peak growth velocity of sitting height.  Use of longitudinal data of shoe sizes from two shoe shops. Mean length of follow-up after 8 years of age – 5.4 years for girls and 6.2 for boys | No clinical assessment.  Increase in shoe size from this study plotted against sitting height growth velocity from an unrelated data source in the Netherlands | Longitudinal data, albeit from shoe shops. Not clear whether retrospective or prospective study. Lack of demographic information and adjustment for confounding. Shoe size may be brand-dependent. Fewer boys than girls, and boys could have a second foot growth spurt not captured in this study.  Mean peak increase in shoe size for girls was 10.4 years and for boys was 11.5 years. Average age for girls (n=138) to reach a plateau phase in shoe size was 12 years. Average for boys (n=36) was 13.7 years. The mean peak increase in shoe size generally occurred 1.3 years and 2.5 years before the mean peak growth of sitting height, in girls and boys respectively, when correlating the results with sitting height from a different data source.63Conclusion: Shoe size increase can be a useful indicator for the timing of the sitting height growth spurt, however, this is based on cross-study correlation. |
| Mitra,31 India, 2011 | 973 middle income healthy children, urban, age 8-16 years | Correlation of foot length and full sexual maturity rating (SMR): breast and pubic hair development staging for girls and genital and pubic hair development staging for boys. | Against clinical assessment  Kappa, Student’s t-test | Cross-sectional study with only one measurement per subject. Insufficient information on selection and measurement. Full Tanner staging carried out by two types of secondary sexual characteristics, unlike many other studies, which rely on just one. High observer agreement, although the size of the subset, in which this was studied, is not specified.  Foot length increased rapidly in stage 2 of SMR, coinciding with the onset of puberty. No statistically significant variations in foot size in subsequent stages.Conclusion: Foot length measurement may be an acceptable and easy method, which can be used as an indicator of pubertal development. However, this study is limited by its cross-sectional design. |
| Cole,25 UK, 2010 | 3245 boys aged 9-19 years from Christ’s Hospital School (1936–1964) and 105 girls with Turner syndrome aged 9-18 years from a randomised controlled trial of oxandrolone to increase final height | To fit the SITAR (SuperImposition by Translation And Rotation) model and show how effectively it summarises height increase around the time of puberty.  To show how the estimated subject-specific parameters can be related to earlier exposures and later outcomes.  A shape invariant model combining three parameters of growth: size, tempo and velocity. Boys’ height was measured twice a term between 9 and 19 years of age. | No clinical assessment or correlation to stage of puberty. | Selection, attrition and measurement bias. Substantial sample size for boys. External validity in girls is limited.  The SITAR model explained 99% of the variance in both datasets, matching the fit of individually-fitted Preece-Baines curves. SITAR summarises individual growth curves with a single summary curve and subject specific random effects.  The random effects reflect each subject’s size, growth, tempo and growth velocity.Conclusion: A potentially useful approach to studying height in puberty, using a non-invasive method, but requiring frequent growth measurements |
| Ford,33 USA, 2009 | 86 girls aged 6-7 years when recruited | To determine whether change in foot size may be used as a marker of onset of puberty.  Height, weight, foot size and maturation ratings obtained every six months. Breast and pubic hair staging by inspection and palpation (the latter presumably only for breast) by a ‘trained healthcare professional’. | Correlation of velocity of foot size increase with height velocity increase and the age of onset of secondary characteristics.  Mixed model approach | Lack of information on recruitment. Weight measured, but not used. Six-month interval too long for estimating the onset of puberty to within months.  Onset of secondary sexual characteristics was at a mean age of 8.79 years; foot velocity increase – 8.40 years; height velocity increase – 8.27 years.  Statistically significant difference between age of onset of secondary sexual characteristics and age of foot velocity increase, as well as age of height velocity increase. No significant difference in age of foot velocity increase and height velocity increase.  African-American girls reached maximum height and foot velocities, and started developing secondary sexual characteristics on average of 0.42 years before their white American peers.Conclusion: Change in foot size could be an early marker of onset of puberty, but there was no difference between the age of increased foot velocity and height velocity. Suggests little benefit of using foot size over height. |
| Bundak,27 Turkey, 2007 | 1112 healthy boys aged 8-18 years, urban, high socioeconomic class | To provide normative data for the onset and tempo of puberty in Turkish boys, evaluate height velocity at each testicular volume stage and analyse the growth parameters in puberty.  Biannual growth measurements (weight and height) and calculation of growth velocity and final height (FH) – velocity of less than 0.5 cm/year. Measurement of pubertal development (testis volume, pubic and axillary hair). | Correlated growth measurements with clinical assessment  Linear correlations | Longitudinal design, but individual boys were followed up for different time periods. Only 30 boys followed up till FH. Significant selection bias. Axillary hair assessment is referred to, which is not normally part of Tanner staging.  Good recognition of the role of inter-observer variability and the need for training and consistency. Various subsamples of boys and girls are referred to, although the sample seemingly consisted of boys only.  Height at onset of puberty was positively correlated with FH and with duration of puberty. Maximal height velocity was obtained after the testes attained a volume of 10ml. The testes reached this volume at a mean age of 12.9 years and a mean height of 155cm. Conclusion: Height at onset of puberty was the most important determinant of FH. Height and height velocity were correlated with each testicular volume stage in this study. This may allow measurement of peak height velocity and determination of pubertal stage through the correlated testis volume, without directly measuring testis volume. Several study limitations. |
| Ozer,38 Turkey, 2006 | 150 boys, 9 to 19 years old, selected at random from 300 male patients referred for orthodontic treatment | To determine correlation of the cervical vertebrae maturation (CVM) index with the modified medial phalange index (MP3).  Use of lateral cephalometric and hand phalange X-ray radiographs | No pubertal assessment.  Kappa for correlation of CVM with MP3 | Selection bias due to recruitment from an orthodontic clinic. Small numbers in some of the maturation stages. Simple percentage agreement is given for both inter-rater reliability and correlation of the indices. Analysis of the study findings is limited.  CVM and MP3 maturation stages had 89.3% simple percentage agreement. Conclusion: The modified MP3 index correlates to the CVM index and can be a useful method to assess skeletal maturation. Several study limitations. |
| Karlberg,26 Sweden, 2003 | 81 boys and 64 girls aged 3-18 years, with annual examinations, from an international longitudinal growth study of healthy children | To introduce a pre-pubertal standard for the assessment of pre-pubertal height for children with late onset of puberty, to apply a final height prediction method, and to devise a method for assessing total pubertal height gain.  By plotting height values for each child in a chart containing pre-pubertal reference values one can identify the onset of the pubertal growth spurt, and thereafter apply a final height prediction method. | No pubertal assessment | The study uses data from another study and concentrates on a small proportion of the results, which limits quality assessment and interpretation.  Infancy-childhood-puberty (ICP)-constructed reference values for pre-pubertal growth agreed with those derived directly from the growth record data used in this study. Pre-pubertal growth and pubertal onset can be assessed in a more appropriate way using the ICP model than ordinary cross-sectional standards. Conclusion: A method is proposed by which clinicians can ascertain whether a child with late onset puberty is entering very early puberty, as identified by the prepubertal growth spurt. In addition proposed are a model of prediction of final height and a method of assessing total pubertal height gain. |

*Table 3: Description of studies using gonadotrophin or gonadal hormone methods*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| First author, country, publication year | Study sample | Study focus and details | Validation or correlation, and related measures | Key findings and conclusions |
| Singh,43 Australia, 2015 | 104 healthy children (57 girls), age 10-12 years | To examine the feasibility of three-monthly urine collection in young adolescents, as well as the utility of liquid-chromatography-tandem mass spectrometry assays for urine and serum sex hormones  To examine the association between the changes in anthropometry and self-reported Tanner stage with changes in hormones in both urine and serum over a 12-month period | No clinical assessment  Linear regression for association of changes in anthropometry and self-rated Tanner stage, and changes in urine and serum hormones | Little information on recruitment and participants; pubic hair assessment not included in self-assessment.  The levels of all three hormones positively correlated with Tanner staging at baseline and 12-month follow-up. Change in height was associated with changes in serum testosterone, and serum and urine luteinising hormone (LH) in females and both serum and urine testosterone in males. Change in weight was associated with changes in urine oestradiol, serum testosterone and serum LH in females. Serum testosterone and LH was associated with self-rated Tanner stage in males. Hormone concentrations increased through each Tanner stage and each year of age. High compliance, low attrition rates. Conclusion: Morning urine sampling can be advantageous over single blood sampling due to pulsatile hormone release, in particular in early puberty. However, frequent longitudinal sampling may be necessary. Urinary hormone changes may not be progressive, indicating within-subject variability. Tanner stage and anthropometrical changes lag behind hormonal changes. Yet urine sampling over 2-3 years may be useful in explaining biological basis for patterns of pubertal changes. |
| Shirtcliff,44 USA, 2009 | 82 boys and 78 girls, aged 9-14 years | To evaluate agreement between physical examination and different methods of self-report, the associations between hormones, and the extent to which self-report methods parallel physical examination in relation to hormonal measures  PDS and a picture-based interview about puberty (PBIP) were used for self-report. Standard Tanner stages used for all methods. Breast palpation was used in addition to inspection in clinical examination. Orchidometer was used in boys along with usual Tanner parameters.  Salivary samples for testosterone and DHEA, plus estradiol in girls, were provided throughout a 12-hour day | Against a clinical assessment by a paediatric nurse practitioner  Percentage agreement, Pearson correlation, kappa | Diverse ethnic and socio-economic backgrounds, albeit insufficient sample size to explore differences. Estradiol cyclical before and after menarche, hence limited interpretation. 13% adolescents refused physical exam, but none refused self-assessment.  Modest concordance between PDS and physical exam, and the same between PBIP and physical examination for breast and genital stage, but better for pubic hair (k=0.43). Low to modest concordance between PDS and PBIP. Neither sex nor BMI influenced accuracy. Adolescents tended to report stages that were most typical for their age.  Physical examination reflected basal testosterone and DHEA well in both sexes, but only modestly in relation to estradiol in girls.  PDS genital stage was more closely associated with testosterone and DHEA than in the physical examination in boys, but pubic hair stage was not as highly related to DHEA as in the physical examination. In girls, PDS breast stage performed better than in the examination, and PDS pubic hair was not associated with basal hormones, which was unexpected. PBIP mapped onto basal hormones in parallel to physical exam, or occasionally slightly better. Nonetheless, even the best measure captured only half of the variability in basal hormones.  Conclusion: Different measures may best capture different aspects of pubertal development, and their use may depend on the research question. |
| Ankarberg-Lindgren,45 Sweden, 2004 | 55 boys, age 5.0-18.6 years, healthy (n=25) or investigated for short (n=26) or tall (n=4) stature | To establish levels for comparison for 24-h total and free serum testosterone before and throughout puberty, and to relate these values to testicular volume | All underwent orchidometry and clinical assessment  Curves of serum testosterone at each pubertal stage, Mann-Whitney test for comparison of different pubertal stages | More than half of the sample had growth abnormalities; lack of information on recruitment and baseline characteristics; inconsistent frequency of blood sampling and Tanner staging.  Based on the changes in testosterone and calculated free testosterone (calc-FT) in relation to testicular volume, the authors classified puberty into six stages. Serum testosterone increased progressively to a constant high value in late puberty, with a marked increase between early and mid-puberty. Serum calc-FT increased continuously throughout puberty, with a striking increase between early puberty and the penultimate stage of puberty. The onset of puberty was characterised by an accentuation of the diurnal rhythm due to an early nocturnal increase in testosterone as compared with the prepubertal diurnal rhythm. Late puberty is characterized by a diminished diurnal rhythm compared with early and mid-puberty.Conclusion: This is a small study consisting of detailed investigations of diurnal variation and levels of testosterone before and during puberty, which correlated these to testicular volume and pubic hair stage. |
| Chada,46 the Czech Republic, 2003 | 78 healthy boys throughout childhood and adolescence | To determine changes in serum inhibin B, follicle stimulating hormone (FSH), luteinizing hormone (LH) and testosterone concentrations during childhood and puberty in males and investigate the relationship between these with regards to age and stage of puberty in boys. | Hormone levels correlated with Tanner staging – genital only, no pubic hair assessment. Not known how staging was carried out.  Correlation coefficient | Insufficient detail for study quality assessment. Many values did not reach a statistically significant level for the difference of means of age categories.  Inhibin B increased with pubertal progression and reached a peak in stage G3. A rise in serum gonadotropins and testosterone was observed during pubertal development. The correlation of inhibin B to FSH, LH and testosterone changed during puberty in agreement with previous studies. In early puberty, serum inhibin B concentrations had a positive association with testosterone and LH, but little relationship to FSH. In contrast, around mid-puberty (stage G3), inhibin B lost its positive correlation with LH and testosterone, but developed a strong negative correlation with FSH, which persisted into adulthood. Conclusion: The study may help inform investigations of gonadal disorders in children, but there is less scope of using its findings for pubertal assessment in epidemiological studies, due to complicated methodology and individual variations in hormone levels. |
| Crofton,47 UK and Ireland, 2002 | 345 healthy girls 0-18 years for age-related reference data and 80 premenarchal girls with full pubertal staging: 51 of these investigated for familial short stature, 40 out of which in addition treated with growth hormone (GH) | To investigate how dimeric inhibins change from birth to late adolescence in girls, to derive reference data and to explore their relation with pubertal stage, FSH, oestradiol and each other | Against clinical assessment (seemingly breast staging only) in girls investigated for familial short stature  Kruskall-Wallis ANOVA and Mann-Whitney U-test for differences in hormones between pubertal stages | Lack of information on recruitment and variation in subjects’ baseline characteristics limit interpretation. Only breast staging is used in the analysis, but ‘full pubertal assessment’ is referred to in places. Most girls with pubertal staging were early on in pubertal development. Half the girls with pubertal staging were being treated with GH as part of their usual care, and all the girls taking GH in stage B2 were removed from the analysis due to higher levels of inhibin B and FSH in this group.  Both inhibins varied markedly with age, and inhibins A, B, FSH and oestradiol all showed a marked relationship with pubertal stage. Detailed results of the hormone concentration fluctuations by age and puberty stage also provided.Conclusions: Several conclusions are offered by the authors in relation to correlations and changes in the levels of the hormones studied. ‘To summarise, in girls in late puberty, there may be complex interactions among FSH, oestradiol and both inhibins that would require more detailed longitudinal and interventional studies to elucidate’. |

Table 4: Description of studies using leptin methods

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| First author, country, publication year | Study sample | Study focus and details | Validation or correlation, and related measures | Key findings and conclusions |
| Maqsood,53 UK, 2007 | 20 children (13 boys) judged to be close to the initiation of puberty | Three consecutive first morning urine samples collected from each subject each month over 6 months.  The children were subsequently classified into those who remained physically prepubertal (n = 7) and those that had advanced in puberty (n = 13).  Leptin and gonadotrophins were measured by immunoradiometric and immunofluorometric assay respectively. | Arbitrary measure of puberty, looking at whether the child has remained physically pre-pubertal or advanced in puberty  Correlation coefficient | Early morning urine samples collected for leptin. Pubertal assessment oversimplified. Focused more on correlation between follicle stimulating hormone (FSH), luteinising hormone (LH) and leptin, than pubertal stage or development and leptin. Small numbers.  In children approaching and progressing into puberty, leptin was associated with changes in LH and FSH over the same timeframe.Conclusion: The data imply that leptin is an important facilitator in the early stages of pubertal development. |
| Wang,51 Japan, 2004 | 906 schoolchildren, aged 9 to 17 years, urban | Cross-sectional study, incorporating height and other measurements, as well as a questionnaire.  Serum leptin measured in fasting blood samples | Age of menarche in girls and maximum increment in height (MIA) used in boys.  Pubertal stage estimated on the basis of growth.  One-way ANOVA for differences in leptin at different pubertal stages | The method of pubertal assessment does not appear to have been validated and the only references are to papers from their team published in local reports  The paper establishes standard age variation curves of serum leptin levels in healthy adolescents by calculating 25th, 50th and 75th centiles for each age in both boys and girls. Leptin rises in boys before MIA then falls, but steadily rises in girls.Conclusion: A potentially useful method for evaluating serum leptin levels in adolescents, considering the impact of gender and growth, but the ranges of measurement at each age are wide, particularly for boys, and these are presented by chronological age, not pubertal stage. |
| Zaman,54 UK, 2003 | 188 healthy school children aged 5-19 years, recruited as part of a larger study. | Cross-sectional study of urine leptin, measured in the first morning void urine and expressed as nanograms excreted overnight, and serum concentrations of leptin, IGF-1, IGF-11, IGFBP-3 and IGFBP-1. | Against pubertal status by SMS  Non-linear regression for relationship between urinary leptin and pubertal staging | No information on recruitment into this or the larger study, from which the data were used; not stated if pubertal staging was by self-assessment or clinical assessment.  Numbers are higher than is often the case in studies on hormones, but small numbers in some pubertal stages.  Urinary leptin showed similar changes through puberty to those of serum leptin, with levels rising in females throughout puberty, whereas levels in males peaked in stages 2/3 and then decreased.Conclusion: Urinary leptin is a valid marker of serum leptin concentration. |
| Mantzoros,50 USA, 1997 | 8 boys aged 9.75-11.9 at entry, recruitment method unclear | A longitudinal assessment of leptin and testosterone levels measured every four months.  Correlation of leptin was with the onset of puberty as determined by rise in testosterone levels. | Against clinical assessment using SMS  Rise in testosterone was taken as the onset of puberty  Repeated measures ANOVA to compare means and trends across puberty | Small numbers. Examined longitudinally until Tanner stage 5. Frequent sampling. Testosterone and dehydroepiandrosterone levels also obtained.  Leptin levels rose about 100% before or at the time of the onset of puberty, and decreased to baseline after the initiation of puberty.Conclusion: This method would be difficult to use as a one-off measure because levels before and after puberty are similar. |
| Carlsson,52 Sweden, 1997 | 252 healthy children aged 1.7 to 18.6 years, source of participants unclear | Cross-sectional study, with a longitudinal element for 15 girls with known dates of menarche. Blood samples obtain between 10am and 2pm from which serum leptin concentrations were obtained.  Subgroup of 15 girls was followed up with two to seven repeated observations. | Against Tanner staging by SMS (assumed by clinical assessment) and testicular volume  Mann-Whitney U-test | Different findings to other studies. Many samples for boys were taken in pre-pubertal stages, with low numbers for the pubertal stage (124 vs 38).  30-fold variation in leptin concentration between participants.Conclusion: In girls leptin increased through puberty, whereas there was no change in leptin with pubertal development in boys. The paper hypothesises this may be due to amount of adipose tissue. |

Table 5: Description of studies using voice maturation methods

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| First author, country, publication year | Study sample | Study focus and details | Validation or correlation, and related measures | Key findings and conclusions |
| Ong,57 UK, 2012 | 2008 boys aged 14 years from the 1946 birth cohort | In 1961, the boys underwent a pubertal assessment at school, which included secondary sexual characteristics and subjective assessment of voice.  The child’s school doctor rated the development of genitals, pubic hair and voice on a *three-option rating scale, and appearance* of axillary hair on a dichotomous scale.  Assessed voice-breaking at 14 years and height difference in SD scores (HD-SDS) as markers of timing of pubertal maturation | Against clinical assessment with a simplified rating scale. Voice change timing used as a marker of pubertal development and cross-referenced to later outcomes in adulthood  Intercorrelations between voice breaking and other markers | Large sample of White European descent. Crude pubertal development rating scale. Subjective method of voice assessment.  Similar to females with earlier menarche, the trajectory to earlier sexual maturation in males was preceded by faster early postnatal growth and weight gain, and lead to higher adult BMI. Timing of pubertal maturation was seen to have potential relevance to adult disease risks in males.Conclusion: This study demonstrated how voice could be used as a marker of timing of pubertal maturity in males. |
| Harries,55 UK, 1997 | 26 school boys aged 13-14 years | To investigate the characteristics of the male speaking and signing voice in relation to other biological changes in puberty.  Measurement of standing height, weight, pubertal stage by Tanner, and testicular volume; salivary testosterone profiling; acoustic and musical recordings. Three-monthly follow-up over 12 months. | Correlation between the speaking and signing voice and testosterone levels, testicular volume and Tanner stages of puberty (assumed by clinical assessment).  Correlation coefficient | Potential for selection bias – no details of recruitment are given, and observer bias – not clear who carried out the measurements. Small numbers. Perhaps a wider range of ages and a longer time-period would have been beneficial, although the group appear to have collated information throughout puberty.  Changes in voice fundamental frequencies correlated with testis volume, but not with testosterone levels. There was a clear correlation between Tanner stages and a Cooksey musical classification during male puberty. Voice-breaking is a late event in male puberty.Conclusion: This study shows a good correlation between Tanner stages and Cooksey stages of pubertal voice development. It provides an alternative method to assessing puberty in cohort and longitudinal studies. |

Table 6: Conclusions from the literature review and opinions of experts and adolescents

|  |  |  |  |
| --- | --- | --- | --- |
| Category of assessment | Literature review | Opinions of experts | Opinions of adolescents |
| Self-assessment | Children earlier on in their development may overestimate their pubertal stage, and children in the later stages may underestimate their stage.  Girls tended to be more accurate than boys.  Some aspects, e.g. pubic hair development, reported more accurately than others.  Higher validity if adolescents self-examined before self-rating, compared to recall from memory.  Generally high agreement between SMS and PDS to within one Tanner stage.  Colour drawings led to closer correlation between self-assessments and clinical examination.  May be less reliable in overweight or obese children.  Possible over- and underestimation of development stages in parental assessment, compared to clinical assessment. | A rather crude approach, given that accuracy may be to within one Tanner stage.  May be the easiest method within cohort studies and the most practical method within cross-sectional studies. | Questionnaires preferred to clinical assessment.  Line drawings preferred to photographs.  Paper questionnaires preferred to online assessments. |
| Growth | Age at height growth take-off indicates pubertal growth spurt.  Age at peak height velocity indicates the intensity of the pubertal growth spurt.  The SITAR method produces measurements representing differences in mean size and growth tempo, and a measure of growth velocity.  SITAR can also be applied to foot length and secondary sexual characteristics.  Rapid increase in foot length may precede peak increase in sitting height or correspond to the onset of puberty.  Radiological methods of assessing olecranon and digital maturation may be used in the assessment of skeletal maturity, but pubertal staging was not done. | Assessments of growth in height or foot size could be promising, but more evidence is needed in relation to foot size.  Growth assessment with adequate measurement frequency may be suitable for cohort studies, but by definition not for cross-sectional studies.  Limited use of radiological methods, in part due to the need for clinic attendance. | Measurements of height and foot size would be acceptable. |
| Hormonal assessment:  *Gonadotrophins and gonadal hormones* | Urinary testosterone, oestradiol and luteinizing hormone positively correlated with SMS-based Tanner staging.  Staging by clinical examination reflected testosterone and DHEA in both sexes.  Increased release of testosterone at night corresponded to the onset of puberty in boys.  Progressive changes in inhibins A and B were associated with pubertal staging, but these were not consistent enough.  Overall, interpretation is challenging due to a complex relationship between gonadotrophins and gonadal hormones. | Interesting area, but application is limited due to the need for repeat measurements and cost of assays. | Not discussed |
| Hormonal assessment:  *Leptin* | Possible relationships with the onset of puberty.  Urinary leptin correlated with serum leptin.  Inconsistent results and intra- and inter-person variability. | As above, interesting area, but application is limited due to the need for repeat measurements and cost of assays. | Not discussed |
| Voice maturation assessment | Timing of voice breaking showed moderate correlation with genital development, and pubic and axillary hair growth.  Correlation between Cooksey classification of boys’ voices and Tanner stages. | A promising method, but more research is needed. | Voice maturation assessment would be acceptable. |

**References**

1. Marshall WA, Tanner JM. Variations in pattern of pubertal changes in girls. *Arch Dis Child*. 1969;44(235), 291-303.

2. Marshall WA, Tanner JM. Variations in the pattern of pubertal changes in boys. *Arch Dis Child*. 1970;45(239), 13-23.

3. Coleman L, Coleman J. The measurement of puberty: a review. *J Adolesc*. 2002;25(5), 535-550.

4. Dorn LD, Biro FM. Puberty and Its Measurement: A Decade in Review. *J Res Adolesc*. 2011;21(1), 180-195.

5. Schmitz KE, Hovell MF, Nichols JF, et al. A Validation Study of Early Adolescents' Pubertal Self-Assessments. *J Early Adolesc*. 2004;24(4), 357-384.

6. Tanner JM. *Growth at Adolescence*, 2nd edn, 1962. Blackwell: Oxford.

7. Petersen AC, Crockett L, Richards M, Boxer A. A self-report measure of pubertal status: Reliability, validity, and initial norms. *J Youth Adolesc*. 1988;17(2), 117-133.

8. Dorn LD, Dahl RE, Woodward HR, Biro F. Defining the Boundaries of Early Adolescence: A User's Guide to Assessing Pubertal Status and Pubertal Timing in Research With Adolescents. *Appl Dev Sci*. 2006;10(1), 30-56.

9. Centre for Reviews and Dissemination. Systematic reviews: CRD's guidance for undertaking reviews in health care. University of York; 2009 [cited 2019 20 February]; Available from: <https://www.york.ac.uk/media/crd/Systematic_Reviews.pdf>.

10. Howard DM. Speechtest. 2014 [cited 2018 24 July]; Available from: <https://itunes.apple.com/gb/app/speechtest/id904673964?mt=8>.

11. Norris SA, Richter LM. Are there short cuts to pubertal assessments? Self-reported and assessed group differences in pubertal development in African adolescents. *J Adolesc Health*. 2008;42(3), 259-265.

12. Bonat S, Pathomvanich A, Keil MF, Field AE, Yanovski JA. Self-assessment of pubertal stage in overweight children. *Pediatrics*. 2002;110(4), 743-747.

13. Morris NM, Udry JR. Validation of a self-administered instrument to assess stage of adolescent development. *J Youth Adolesc*. 1980;9(3), 271-280.

14. Brooks-Gunn J, Warren MP, Rosso J, Gargiulo J. Validity of self-report measures of girls' pubertal status. *Child Dev*. 1987;58(3), 829-841.

15. Hergenroeder AC, Hill RB, Wong WW, Sangi-Haghpeykar H, Taylor W. Validity of self-assessment of pubertal maturation in African American and European American adolescents. *J Adolesc Health*. 1999;24(3), 201-205.

16. Carskadon MA, Acebo C. A self-administered rating scale for pubertal development. *J Adolesc Health*. 1993;14(3), 190-195.

17. Desmangles J-C, Lappe JM, Lipaczewski G, Haynatzki G. Accuracy of pubertal Tanner staging self-reporting. *J Pediatr Endocrinol Metab*. 2006;19(3), 213-221.

18. Rasmussen AR, Wohlfahrt-Veje C, Tefre de Renzy-Martin K, et al. Validity of self-assessment of pubertal maturation. *Pediatrics*. 2015;135(1), 86-93.

19. Norris SA, Richter LM. Usefulness and Reliability of Tanner Pubertal Self-Rating to Urban Black Adolescents in South Africa. *J Res Adolesc*. 2005;15(4), 609-624.

20. Bond L, Clements J, Bertalli N, et al. A comparison of self-reported puberty using the Pubertal Development Scale and the Sexual Maturation Scale in a school-based epidemiologic survey. *J Adolesc*. 2006;29(5), 709-720.

21. Rabbani A, Noorian S, Fallah JS, et al. Reliability of pubertal self assessment method: an Iranian study. *Iran J Pediatr*. 2013;23(3), 327-332.

22. Sun Y, Tao FB, Su PY, China Puberty Research C. Self-assessment of pubertal Tanner stage by realistic colour images in representative Chinese obese and non-obese children and adolescents. *Acta Paediatr*. 2012;101(4), e163-166.

23. Berg-Kelly K, Erdes L. Self-assessment of sexual maturity by mid-adolescents based on a global question. *Acta Paediatr*. 1997;86(1), 10-17.

24. Lum S, Bountziouka V, Harding S, et al. Assessing pubertal status in multi-ethnic primary schoolchildren. *Acta Paediatr*. 2015;104(1), e45-48.

25. Pereira A, Garmendia ML, Gonzalez D, et al. Breast bud detection: a validation study in the Chilean growth obesity cohort study. *BMC Womens Health*. 2014;14, 96.

26. Karlberg J, Kwan C-W, Gelander L, Albertsson-Wikland K. Pubertal growth assessment. *Horm Res*. 2003;60(Suppl 1), 27-35.

27. Bundak R, Darendeliler F, Gunoz H, et al. Analysis of puberty and pubertal growth in healthy boys. *Eur J Pediatr*. 2007;166(6), 595-600.

28. Cole TJ, Donaldson MDC, Ben-Shlomo Y. SITAR--a useful instrument for growth curve analysis. *Int J Epidemiol*. 2010;39(6), 1558-1566.

29. Cole TJ, Pan H, Butler GE. A mixed effects model to estimate timing and intensity of pubertal growth from height and secondary sexual characteristics. *Ann Hum Biol*. 2014;41(1), 76-83.

30. Frysz M, Howe LD, Tobias JH, Paternoster L. Using SITAR (SuperImposition by Translation and Rotation) to estimate age at peak height velocity in Avon Longitudinal Study of Parents and Children. *Wellcome Open Res*. 2018;3, 90.

31. Mitra S, Samanta M, Sarkar M, Chatterjee S. Foot length as a marker of pubertal onset. *Indian Pediatr*. 2011;48(7), 549-551.

32. Busscher I, Kingma I, Wapstra FH, et al. The value of shoe size for prediction of the timing of the pubertal growth spurt. *Scoliosis*. 2011;6, 1-1.

33. Ford KR, Khoury JC, Biro FM. Early Markers of Pubertal Onset: Height and Foot Size. *J Adolesc Health*. 2009;44(5), 500-501.

34. Greulich W, Pyle SI. *Radiographic atlas of skeletal development of the hand and wrist.*, 2nd edn, 1959. Stanford University Press: Stanford, CA.

35. Tanner JM, Whitehouse RH, Cameron N. *Assessment of skeletal maturity and prediction of adult height (TW3 Method).* 3rd edn, 2001. W.B Saunders: London.

36. Cericato GO, Bittencourt MAV, Paranhos LR. Validity of the assessment method of skeletal maturation by cervical vertebrae: a systematic review and meta-analysis. *Dentomaxillofac Radiol*. 2015;44(4), 20140270.

37. Canavese F, Charles YP, Dimeglio A, et al. A comparison of the simplified olecranon and digital methods of assessment of skeletal maturity during the pubertal growth spurt. *Bone Joint J*. 2014;96-B(11), 1556-1560.

38. Ozer T, Kama JD, Ozer SY. A practical method for determining pubertal growth spurt. *Am J Orthod Dentofacial Orthop*. 2006;130(2), 131.e131-136.

39. Apter D, Bützow T, Laughlin GA, Yen SS. Accelerated 24-hour luteinizing hormone pulsatile activity in adolescent girls with ovarian hyperandrogenism: relevance to the developmental phase of polycystic ovarian syndrome. *J Clin Endocrinol Metab*. 1994;79(1), 119-125.

40. Grumbach MM, Roth JC, Kaplan SL, Kelch RP. Hypothalamic-pituitary regulation of puberty in man: Evidence and concepts derived from clinical research. In *Control of the onset of puberty*. 1974; pp. 115-166. John Wiley & Sons: New York.

41. Søeborg T, Frederiksen H, Mouritsen A, et al. Sex, age, pubertal development and use of oral contraceptives in relation to serum concentrations of DHEA, DHEAS, 17α-hydroxyprogesterone, Δ4-androstenedione, testosterone and their ratios in children, adolescents and young adults. *Clin Chim Acta*. 2014;437, 6-13.

42. Sehested A, Andersson AM, Müller J, Skakkebaek NE. Serum inhibin A and inhibin B in central precocious puberty before and during treatment with GnRH agonists. *Horm Res*. 2000;54(2), 84-91.

43. Singh GKS, Balzer BWR, Kelly PJ, et al. Urinary Sex Steroids and Anthropometric Markers of Puberty - A Novel Approach to Characterising Within-Person Changes of Puberty Hormones. *PLoS One*. 2015;10(11), e0143555.

44. Shirtcliff EA, Dahl RE, Pollak SD. Pubertal development: correspondence between hormonal and physical development. *Child Dev*. 2009;80(2), 327-337.

45. Ankarberg-Lindgren C, Norjavaara E. Changes of diurnal rhythm and levels of total and free testosterone secretion from pre to late puberty in boys: testis size of 3 ml is a transition stage to puberty. *Eur J Endocrinol*. 2004;151(6), 747-757.

46. Chada M, Prusa R, Bronsky J, et al. Inhibin B, follicle stimulating hormone, luteinizing hormone and testosterone during childhood and puberty in males: changes in serum concentrations in relation to age and stage of puberty. *Physiol Res*. 2003;52(1), 45-51.

47. Crofton PM, Evans AEM, Groome NP, et al. Dimeric inhibins in girls from birth to adulthood: relationship with age, pubertal stage, FSH and oestradiol. *Clin Endocrinol (Oxf)*. 2002;56(2), 223-230.

48. Rockett JC, Lynch CD, Buck GM. Biomarkers for assessing reproductive development and health: Part 1--Pubertal development. *Environ Health Perspect*. 2004;112(1), 105-112.

49. Brann DW, Wade MF, Dhandapani KM, Mahesh VB, Buchanan CD. Leptin and reproduction. *Steroids*. 2002;67(2), 95-104.

50. Mantzoros CS, Flier JS, Rogol AD. A longitudinal assessment of hormonal and physical alterations during normal puberty in boys. V. Rising leptin levels may signal the onset of puberty. *J Clin Endocrinol Metab*. 1997;82(4), 1066-1070.

51. Wang T, Morioka I, Gowa Y, et al. Serum leptin levels in healthy adolescents: Effects of gender and growth. *Environ Health Prev Med*. 2004;9(2), 41-46.

52. Carlsson B, Ankarberg C, Rosberg S, et al. Serum leptin concentrations in relation to pubertal development. *Arch Dis Child*. 1997;77(5), 396-400.

53. Maqsood AR, Trueman JA, Whatmore AJ, et al. The relationship between nocturnal urinary leptin and gonadotrophins as children progress towards puberty. *Horm Res*. 2007;68(5), 225-230.

54. Zaman N, Hall CM, Gill MS, et al. Leptin measurement in urine in children and its relationship to other growth peptides in serum and urine. *Clin Endocrinol (Oxf)*. 2003;58(1), 78-85.

55. Harries ML, Walker JM, Williams DM, Hawkins S, Hughes IA. Changes in the male voice at puberty. *Arch Dis Child*. 1997;77(5), 445-447.

56. Hodges-Simeon CR, Gurven M, Cardenas RA, Gaulin SJC. Voice change as a new measure of male pubertal timing: a study among Bolivian adolescents. *Ann Hum Biol*. 2013;40(3), 209-219.

57. Ong KK, Bann D, Wills AK, et al. Timing of voice breaking in males associated with growth and weight gain across the life course. *J Clin Endocrinol Metab*. 2012;97(8), 2844-2852.

58. Cooksey JM. The male adolescent changing voice: Some new perspectives. In *Research symposium on the male adolescent voice*. (ed. Runfola M), 1984; pp. 4-59. Buffalo: State University of New York Press.

59. Dorn LD. Measuring puberty. *J Adolesc Health*. 2006;39(5), 625-626.

60. Carel JC, Leger J. Clinical practice. Precocious puberty. *N Engl J Med*. 2008;358(22), 2366-2377.

61. Taylor SJ, Whincup PH, Hindmarsh PC, et al. Performance of a new pubertal self-assessment questionnaire: a preliminary study. *Paediatr Perinat Epidemiol*. 2001;15(1), 88-94.

62. Duke PM, Litt IF, Gross RT. Adolescents' self-assessment of sexual maturation. *Pediatrics*. 1980;66(6), 918-920.

63. Gerver WJ, de Bruin R. *Paediatric morphometrics: a reference manual*, 2001. UPM Maastricht.