Statistical issues in the assessment of health outcomes in children: a methodological review

Gillian A Lancaster, PhD, CStat, Senior Lecturer in Medical Statistics and Director of Postgraduate Statistics Centre, Lancaster University, UK

Address for correspondence: Dr G.A. Lancaster, Postgraduate Statistics Centre, Dept of Mathematics and Statistics, Lancaster University, Fylde College, Lancaster, LA1 4YF. Email: g.lancaster@lancaster.ac.uk, telephone: 01524 593943, fax: 01524 592681.

Summary

The lack of outcome measures that are validated for use on children limits the effectiveness and generalisability of paediatric health care interventions. Statistical epidemiology is a broad concept encompassing a wide range of useful techniques for use in child health outcome assessment and development. However the range of techniques available is often confusing and prohibits their adoption. In this paper an overview of methodology is provided within the paediatric context. It is demonstrated that in many cases assessment can be performed relatively straightforwardly using standard statistical techniques, although sometimes more sophisticated techniques are required. Examples of both physiological and questionnaire based outcomes are given. The usefulness of these techniques is highlighted for achieving specific objectives and ultimately for achieving methodological rigour in clinical outcome studies performed in the paediatric population.

Summary: 130 words

Keywords: child health, child mental health, health outcomes, quality of life, statistical methodology, study design

Running title: Statistical issues in the assessment of health outcomes in children

1. Introduction

The European Commission's proposal for a Regulation on Medicinal Products for Paediatric use (Commission of the European Communities, 2004) stated that 'The paediatric population is a vulnerable group with developmental, physiological and psychological differences from adults, which makes age and development related research of medicines particularly important… more than 50% of the medicines used to treat the children of Europe have not been tested and are not authorised for use in children: the health and therefore the quality of life of the children of Europe may suffer...' In the EU, the paediatric population (0-16 years) represents about 75 million people, that is 20% of the total population.

Child health outcome assessment can be defined as the procedures used to describe and quantify the effectiveness of all paediatric health care interventions including medicines. In this respect we want to be able to distinguish between positive and negative effects of treatment and quantify the magnitude of these effects. Ethical issues of research involving children (Helseth and Slettebo, 2004) emphasise the importance of assessment and quantification of health outcomes in the paediatric population. However, the lack of outcome measures that are validated for use on children limits the generalisability of treatment effectiveness results (Patrick and Chiang, 2000). In this paper we describe the main issues to consider in health outcome assessment during the process of development, illustrate various measurement methods using examples taken from different paediatric contexts, and highlight the appropriate statistical tools to use in each case. The emphasis is on providing an overview of the process to give researchers a context within which to work and with useful references for further study of specific techniques.

Paediatric assessments are derived from several sources. They may be based on child-reported outcomes, for example, pain measurement in children as young as 3 years old using smiley

faces (Wong and Baker, 1995), or parent or caregiver-reported outcomes, for example, to assess burden of care (Glasscoe et al., 2006a) or perceived Health Related Quality of Life (HRQL) (Eiser and Morse, 2001a). The use of proxy respondents prompts much debate in the literature (Eiser and Morse, 2001b, Janse et al., 2004) but may help to limit missing data when evaluating treatment over time for those unable to participate. Clinician or assessor-reported outcomes include physiological or pathological measures, for example, Body Mass Index (BMI) (Cleary et al., 2004), white cell count (Farrell et al., 2002), perceived pain (Stewart et al., 2004) or survival (Wong et al., 2000). Data may be collected in a variety of ways, for example, from medical records, laboratory reports or by direct observation and measurement, or through interviews, self-administered questionnaires and daily diaries.

All forms of assessment such as diagnostic and laboratory testing and psychometric testing, require instruments that have been shown to be reliable and valid (Gnecco and Lachenbruch, 2002). In this paper we are using the term 'instrument' to refer to any measuring device whether a mechanical device or a questionnaire. Development of measurement instruments is discussed in Section 2, and issues concerning reliability and validity of measurement in Section 3. A useful instrument should also be able to demonstrate that it can detect changes in a health outcome over time within subjects, and also distinguish or discriminate between subjects on the scale of interest (Guyatt et al., 1987, Chwalow and Adesina, 2002), and these issues are discussed in Section 4. The establishment of reference values for healthy children in the population of interest is another important aspect of instrument development (Jones et al., 1993, Marquis et al., 2004) that is addressed in Section 5. Some further considerations are given in the discussion in Section 6.

2. Development of a new instrument

There are several stages to go through in developing a new instrument and in demonstrating its reliability and validity, and therefore several phased studies (Asmussen et al., 1999) are usually required. The main stages in the development and assessment of a health outcome measure are summarised in Table 1.

Physiological measurement devices

If a new physiological measuring device has been developed and put on the market, for example, for measuring a child's blood glucose level or temperature, then it would be of interest to compare the performance of this device to one in current use, to consider its potential for adoption in practice. The new procedure may be less costly or painful, or the standard device too invasive to use in young children. A method comparison study can be carried out to determine whether the new method agrees sufficiently well with the reference standard. Bland and Altman (1999) have published a well-known methodology for quantifying systematic error (bias) and random error (limits of agreement), assuming that the measurements are recorded on the same continuous scale. If repeated measurements or replicates are available then methods based on Analysis of Variance (ANOVA) (Bland and Altman, 1999), confirmatory factor analysis (Dunn, 1992) or multi-level modelling (Snijders and Bosker, 1999) can separate out the different sources of variability. When the two measurements have not been recorded on the same scale then regression modelling can be used to calibrate one set of values to the same scale as the other set, provided that an estimate of repeatability is available from the same or a comparable sample (Chinn, 1990, Dunn and Roberts, 1999). Cohen's kappa (Cohen, 1960) is a popular measure of agreement for binary and ordinal data, which is discussed further in section 3.

Method comparison studies are closely related to diagnostic test studies in terms of study design but enable more detailed preliminary evaluation to determine whether the methods

agree with each other across the whole spectrum of the measurement scale, or whether they do not agree but relate to each other by a constant amount above or below the other (Craig et al., 2002). In a diagnostic test study the new method is evaluated against a given single threshold or cut-off, and the second possibility cannot then be examined. In diagnostic test studies we are interested in establishing the sensitivity and specificity of the test compared to a gold standard criterion (see also Table 1), classifying for example whether a child is diseased or not diseased, and this type of study design and analysis is well documented in the literature (see for example Pepe (2003)). Prospective recruitment of children from the target population as a consecutive or random sample ensures measurements are taken across the whole spectrum of the disease. An alternative case-control approach has been shown to inflate estimates of sensitivity and specificity and test performance (Lijmer et al., 1999), because it concentrates only on the extremes of the sample (known cases and controls) and may not cover the middle of the spectrum. Altman (1991) gives guidelines for sample size for method comparison studies, NQuery Advisor 5.0 (Elashoff, 2003) is able to calculate sample sizes for confidence interval estimation of the kappa statistic, and sample size considerations for diagnostic test studies is given in Freedman (1987) and by Zhou et al. (2002).

In general, the methodological quality of these types of studies has been found to be lacking, particularly in children (Craig et al., 2000, Farrell et al., 2002). Common misconceptions included the use of correlation rather than agreement to compare the two methods, very small sample sizes that were boosted by including repeated measurements on the same individuals in the sample as independent observations, ignoring within-person correlation, and poor or no description of the procedural methods used to control for bias. Criteria for assessing methodological quality in these types of studies are given in Craig et al. (2000) and a summary of the related types of biases that have been identified are listed in Table 2.

Questionnaires

When the instrument takes the form of a questionnaire, the first stage of development is quite often the creation of questions or items through the use of focus groups and qualitative interviews with children or parents and carers (a topic in itself) to identify for example, the impact of caring for children with cystic fibrosis (Glasscoe et al., 2006a, Glasscoe et al., 2006b). The instrument may also be shown to users and experts to establish face and content validity. For example, in the development of a culturally appropriate developmental assessment tool for use on children in Malawi (Gladstone et al., 2008), problematic items were re-adapted or re-translated after cursory review by eight local research workers, five Malawian paediatricians, six medical students and a language expert from the University of Malawi. In later stages of development the instrument is piloted on a small sample of children or carers with preliminary evaluation of item performance. Item performance evaluation identifies mean, minimum and maximum responses, percentage missing, floor and ceiling effects (where majority of responses fall in lowest/highest category) and item-total correlations (Streiner and Norman, 2003). Optimal item performance occurs with large response ranges, few missing data and low percentages of minimum and maximum values (Asmussen et al., 1999, Kleinman et al., 2006).

In the paediatric context there are issues to address regarding the changing developmental status of a child. In particular, development of a carefully designed age-related instrument, incorporating the correct level of language comprehension with age-appropriate instrument formatting and design will minimize the likelihood of 'response sets'. The term 'response set' means that a child will respond in a certain type of way regardless of what is being asked, for example, through repetitive responses, or to please the interviewer or appear competent (Matza et al., 2004). However, an approach that addresses developmental status by using age-specific instruments poses a problem when measuring change over time, particularly in

consistency and context of measurement (see 'Measuring Health Related Quality of Life' in Section 4). Problems of translation have also been highlighted (Wittes, 2002), with scores on a measure of verbal fluency being affected by the range of vocabulary available within the different languages of the countries taking part in the trial. Whilst this example was based on an adult sample it does have similar implications for instrument development in children.

3. Establishing the reliability and validity of measurement

Reliability is concerned with the consistency of measurement, whether measurements are made by the same person on different occasions or different people on the same occasion, and validity is concerned with the accuracy of measurement and whether an instrument is actually measuring what it is supposed to be measuring. If an instrument exists but has never been used in a certain paediatric population, or has been adapted from adult studies, then it is important to establish that it is reliable and valid for use in that population before applying it in an intervention study. For example, an instrument that is reliable and valid in a clinic setting may be neither reliable nor valid in the community or hospital. Validity and reliability are therefore not fixed, immutable properties of an instrument but rather an interaction between the instrument and group completing it, and these properties may vary from one situation to another. Some attempts have been made to assess an instrument from within an intervention study. However, this has the potential for inflating false positive error rates, as well as a lack of reproducibility and generalisability to a more heterogeneous patient group to which the intervention would be applied in practice (Gnecco and Lachenbruch, 2002). It also runs the risk of conducting much of the study before discovering that the instrument is not valid or reliable.

Reliability

A reliable instrument is one whose scores shows stability and consistency of measurement when used by the same assessor on two separate occasions (intra-rater/test-retest reliability) or

different assessors on the same occasion (inter-rater reliability). Here we are interested in the accuracy of measurement in relation to the likely range of values of the instrument, for example for detecting wheeze in infants (Powell et al., 2002). Reliability is most commonly expressed as a ratio of the variability between individuals to the total variability in the measurements, which is the variability between individuals plus measurement error. This is called an intra-class correlation coefficient (ICC) and can be calculated easily using ANOVA (Streiner and Norman, 2003) or multilevel modelling (Snijders and Bosker, 1999). It is not a fixed characteristic and can change with the prevalence of the condition being studied between populations of children (Dunn, 1992), see for example Feeny et al.'s (2004) teenage comparisons using the health utilities index mark 2 (HUI2) and HUI3 scores. There has been much debate as to the most appropriate choice of reliability co-efficient. In particular, Cohen's kappa (Cohen, 1960) is often used to measure agreement in the medical literature (Powell et al., 2002, Elphick et al., 2004, Stewart et al., 2004), however if a quadratic weighting scheme is used, it can be shown to be exactly identical to an ICC (Fleiss and Cohen, 1973). It too is affected by prevalence. A generalised kappa statistic has been proposed for assessments made by multiple (>2) raters (Fleiss et al., 2003). The constraint underlying kappa, is that the probability of positively rating inconclusive items (random ratings) is equivalent to that for rating conclusive items (systematic ratings). When this constraint is violated then latent class analysis (Guggenmoos-Holzmann and Vonk, 1998) provides a more general framework within which to work. In general ICCs are used for continuous data and kappas for binary or categorical data. The design and analysis of reliability studies is discussed in Dunn (1992, 2000), and the ICC, kappa statistic, Pearson's product-moment correlation coefficient and Bland and Altman's method have been compared by Streiner and Norman (2003).

In diagnostic and laboratory tests that measure physiological outcomes, when a gold standard exists the rationale for the new instrument may be that the existing test is expensive or time-consuming to administer. When a gold standard is not available then assessment of agreement between devices can still be made. For example, Elphick et al. (2004) demonstrated the unreliability of the stethoscope for assessing respiratory sounds in infants, which had important implications for its use as a diagnostic tool for lung disorders, but equally they were unable to assess the reliability of acoustic analysis as an alternative diagnostic procedure because of the lack of a reliable gold standard.

In questionnaire-type measures, the instrument needs to have good internal properties for it to be of use in practice. As part of the reliability study therefore, item performance should be re-evaluated in this larger sample. Internal reliability should then be established, particularly if multiple items are contained within several domains. Internal reliability (Asmussen et al., 1999, Powell et al., 2002, Kleinman et al., 2006) can be assessed using several methods, for example, Cronbach's alpha, a split-half coefficient or the Kuder-Richardson 20 method (Streiner and Norman, 2003), all of which can be computed in standard statistical software. A factor analysis will help to determine the internal structure of the instrument by establishing the number of domains that are being measured and identifying redundant items (Powell et al., 2002). Item response analysis may be usefully applied to categorical or ordinal data (Drachler et al., 2007), and further information about factor analysis of binary and ordinal data (and mixtures) can be found in Bartholomew et al. (2008). Estimation for the methods can be carried out using Mplus (Muthén and Muthén, 1998-2007) or Stata (via GLLAMM) (Skrondal and Rabe-Hesketh, 2004). However, these types of analyses are not always possible, as they require a large sample size (Costello and Osborne, 2005).

Validity

A valid instrument is one which measures what it purports to be measuring in a particular group of children, for example, pain intensity in young children using smiley faces (Wong and Baker, 1995) or HRQL using the PedsQL™ 4.0 (Pediatric Quality of Life Inventory™) (Varni et al., 2003). Construct validity, a term that subsumes the various components of validity, is the focus here and refers to the extent to which the instrument conforms to the predicted theoretical properties that would be expected in its field of application. Convergent validity, for example, is demonstrated if the instrument correlates well with other known constructs to which it should be related, such as the correlation of the PedsQL™ school functioning scale with achievement scores. Conversely, divergent (or discriminant) validity is demonstrated if the instrument does not correlate well with unrelated constructs that should show no association. Construct validity can also be tested by comparing the instrument in different settings, either on two extreme groups of children, for example, those with and without pain, which is called extreme groups construct validity, or on several known diagnostic groups with varying levels of pain, for example, mild pain (minor head injury), moderate pain, or severe pain (compound displaced forearm fracture) which is called known-groups construct validity (Stewart et al., 2004). In this example assuming continuous data, significant differences between ordered diagnostic groups were tested using Cuzick's test for trend (Cuzick, 1985). Criterion validity, another type of construct validity, is demonstrated by comparing the new instrument to an external criterion, ideally a gold standard, well-established outcome, to demonstrate that the new instrument is both sensitive and specific in its diagnosis (see also diagnostic test studies in Section 2). There are two types of criterion validity, namely concurrent and predictive, with each defined according to how the external criterion is measured. For concurrent validity both the instrument and external criterion would be applied blind to the results of the other and measurements taken concurrently or immediately sequentially. For predictive validity the same principles of blinding should be

applied but in this case the external criterion will only become known, and measured, sometime in the future, for example after treatment or a biopsy, thus determining the usefulness of the instrument for predicting a likely outcome. In our pain example the external criterion might be level of analgesia (Stewart et al., 2004) or depth of burn wound (Beyer, 1998) following clinical examination or surgery. Another type of validity, respondent validity, may also be carried out at the end of the study to provide an assessment of the impressions of the users of the instrument.

Validating psychological measurements

Measures for assessing child mental health generally come in the form of questionnaires, sometimes delivered by trained interviewers. The assessments may require repeat visits by the parent and child increasing the likelihood of missed visits due to work commitments or an upset child. The assessments may be designed for use on adults and use language that is too difficult for a child to comprehend, or there may be no gold standard criterion available for use on children. In addition interviews are often time consuming and a child may not be able to tolerate too many questions.

The use of multiple respondents to validate responses in this context has been advocated as a better predictor of disorder (Young et al., 1987), although the agreement between child and parent in structured interviews has been shown to vary depending on type of disorder, with more agreement for behavioural symptoms, and less for anxiety (Hodges, 1993). In the Mental Health Survey of Children and Adolescents in Great Britain (Meltzer et al., 2000) for example, a range of assessment methods were utilised. Face to face interviews helped diagnose depression, anxiety, hyperactivity and conduct disorders, whilst self-completion questionnaires solicited information on the use of cigarettes, alcohol and drugs. In addition, some questionnaires such as the Strengths and Difficulties Questionnaire (SDQ), were

completed by parents, teachers and children aged 11-15 years for cross-informant comparisons. A powerful technique for assessing convergent and discriminant validity together, particularly with multiple informants is the multitrait-multimethod matrix (Streiner and Norman, 2003). Here two or more different traits (eg. childhood anxiety and depression (Cole et al., 1997)) are each assessed by two or more measurement methods (eg. parent, teacher, peer) and a correlation matrix is derived using confirmatory factor analysis. High cross-method, within trait correlations are evidence of convergent validity and low correlations (within-method, cross-trait and cross method, cross-trait) are indicative of discriminant validity (Dunn, 2000).

Many psychometric measures are designed for adults and require answering complex questions that a child may not understand. The Fairy Tale Test (Coulacoglou, 2002) is a novel way of rating, for example, fear of aggression, anxiety, or self-esteem in a quantitative manner from story-telling that has been standardised for use with 7-12 year olds. The use of doll role-play (Emde et al., 2003) is a similar approach that has been advocated to better understand a child's beliefs, experiences and personality. This approach has been used for separating out hypothesised constructs such as avoidance or aggression from children with behavioural difficulties (Hill et al., 2007). Relating psychological well-being to physiological changes is another way of validating psychological measurements when this is possible. For example, elevated salivary cortisol has been positively correlated with externalising behaviour and negatively correlated with internalising behaviour in boys (Zaslow et al., 2006). It has also been associated with aggression and poor self-control, and with the amount of stimulation and attention given to children in child care (Zaslow et al., 2006).

4. Measuring change over time and discriminating between subjects

Since the goal of a treatment or intervention is to effect change in health status, then a useful instrument should be able to demonstrate change both within and between children. Yet this is one of the least studied areas in instrument development (Chwalow and Adesina, 2002). Evaluation of responsiveness over time within an individual might concern the measurement of pain or neurological functioning, for example, whereas measures that discriminate between children might focus on a child's height or intelligence (Guyatt et al., 1987). Instruments of measurement come in many different forms as do their scoring systems. Instruments that measure physiological measurements such as blood pressure or temperature, use scales (eg. mmHg or Celsius) that have well-known interpretative clinical meaning. For example, fever is often defined as a temperature greater than 38^{o}C (Dodd et al., 2006). With psychological instruments it is more difficult to determine what constitutes a meaningful change on the rating scale, and definitions of success can therefore be quite arbitrary, for example, a two-point change from baseline (Marquis et al., 2004). It is only with experience that meaningful magnitudes of change or cut-offs can be established for these types of instruments.

Magnitude of change and measurement error

Responsiveness to change may be viewed as a longitudinal type of construct validity, involving assessment of change within individuals and interpretation of meaning (Patrick and Chiang, 2000). Three measures are commonly used to estimate the magnitude of change from baseline to end point or pre-test to post-test (eg. Kleinman et al., 2006); Cohen's effect size (Cohen, 1988), Guyatt's responsiveness statistic (Guyatt et al., 1987) and the standard error of measurement (Wyrwich et al., 1999) (or variations on this approximately equivalent to the minimum important difference). However, measures based on change scores are only appropriate when the between-subject variance exceeds the within subject variance which is equivalent to an ICC≥0.5 (Streiner and Norman, 2003). Moreover, to reduce problems with regression to the mean caused by overly high (or low) pre-test scores, then Analysis of

Covariance (ANCOVA) provides a more robust analysis for adjustment of pre-test measurements (Vickers and Altman, 2001).

In clinical trials the primary interest is in what constitutes a clinically important difference, since changes may occur simply because of receiving attention, the Hawthorne effect, increasing knowledge, and measurement error. The Hawthorne effect is a well-known phenomena whereby participants in a trial have a better end result simply because of the effect of knowing that they are being studied, and dates back to studies done in the 1920s at the Western Electric Company's Hawthorne plant near Chicago. Therefore changes that exceed these types of variability are most important and as the variability increases larger treatment effects are needed to discriminate between treatment groups and demonstrate efficacy (Guyatt et al., 1987). Measurement error is a problem particularly apparent in dietary outcome assessment, for example, to measure energy intake or nutritional status after administering protein energy supplements to children with cystic fibrosis (Poustie et al., 2006). Multiple-day food records or 24-hour dietary recalls are commonly used as reference instruments to calibrate food frequency questionnaires (FFQs) and to adjust findings for measurement error (Kwiterovich et al., 1997). Biomarkers for energy (doubly labelled water) and protein (urinary nitrogen), for example, may also be used to calibrate measures, but these are limited, costly and may cause inconvenience. Correct adjustment requires that the errors in the adopted reference instrument be independent of those in the FFQ and of true intake (Kipnis et al., 2003). A novel approach used household itemised till receipts to calibrate dietary intake (Ransley et al., 2001), which proved an effective substitute for biomarkers (Greenwood et al., 2006). When assessing dietary interventions in children pilot studies are recommended to ensure the acceptability of the intervention (Lancaster et al., 2004). They are also helpful in determining the choice of appropriate instrument for assessing change, particularly when children are neurologically impaired (Bassi et al., 2004).

Multiple measurements over time

When multiple measurements are taken over time data analysis methods should address the longitudinal nature of the data and it is most efficient to use information at all time points to maximise the potential of the data without loss of information. However in clinical studies, sometimes to avoid the complexity of these types of analyses or because of missing data, the analysis may be restricted to two time points (baseline and the follow up time of primary interest), or a summary measure approach adopted to reduce the multiple measurements to one per child (Matthews et al., 1990). This may discard important information about trends within children or between groups of children. Questionnaire instruments often comprise of items that are rated on a Likert scale with 0 indicating no problem and 4 serious problems, for example. The analysis of ordinal data generated from questionnaire instruments rated on a Likert scale should be analysed using methods of ordinal longitudinal data analysis (Vermunt and Hagenaars, 2004). To overcome this level of complexity, ordinal data may be transformed from a Likert scale onto a common continuous scale ranging from 0-100 as in the PedsQL^{TM} 4.0 (Varni et al., 2003) and thus enable analysis by statistical methods for continuous data. Ordinal data may sometimes be dichotomized into a binary variable with 0 indicating no problem and 1 any type of problem, but this may result in the loss of a rich source of data about the spectrum of severity of the problem and statistical power. Choosing the right analysis strategy is therefore challenging.

Growth curve modelling (Singer and Willett, 2003) is a type of longitudinal data analysis, that allows for non-linear change, which is likely to be pertinent for children. Longitudinal data can be viewed as a hierarchical two-level structure with the measurements made over time at level 1 and children at level 2. Then adopting a multi-level mixed modelling approach, the level 1 model captures within-person change and the level 2 model between-person change. In this respect growth curve models are a special case of general mixed models, in which a

subject-specific trajectory is defined by allowing both a random intercept and random slope with time (or age) as the predictor variable. For example, DeLucia and Pitts (2006) studied the impact of growth over time in emotional autonomy from mothers in the development of adolescents with spina bifida. The models can incorporate higher order growth parameters for non-linear trajectories, and are estimated directly, using routines available in standard statistical software such as SAS or STATA.

As mentioned above, missing data are an additional burden when multiple measurements are taken. It is important therefore, to review the reasons for missingness (Fairclough, 2004). This will help determine whether certain domains are more difficult to measure than others, providing an indication that different methods of assessment may be required. Strategies for handling missing data are also important and necessary for minimising selection bias (Coste et al., 1995). For example, if it can be assumed that the data are missing completely at random (MCAR), that is, that there is no difference between children with observed scores and those with missing scores such that the missing assessments are unrelated to the outcome, then the missing data are ignorable. However, it may be more likely that the data are missing not at random (MNAR) and were caused by dropout due to severity of disease (ie. dropout is related to the unseen responses after dropout), then the missing data are non-ignorable, and more sophisticated methods of analysis, such as the use of pattern mixture models (Parsons et al., 2006), are required to obtain unbiased estimates. The less restrictive missing at random (MAR) assumption assumes dropout is only related to responses made at any occasion prior to dropout (Schafer and Graham, 2002), and is the basis for many statistical methods for adjusting for missing data. Note that under this assumption using growth curve modelling within a multi-level modelling framework, children with missing responses can be included without further adjustment. If multiple imputation is used to impute missing values (Carpenter

et al., 2006), then the imputation model should account for correlation of the responses from the same subject.

Measuring Health Related Quality of Life

Measuring HRQL over time is more complex in paediatric studies than for adult studies, because children will vary in their stages of development. This has prompted the adoption of multiple age-specific forms for use on children. However, these may compromise the stability of the HRQL outcome when taking measurements over time and need to be carefully developed and tested, see for example, the PedsQL^{TM} 4.0 (Varni et al., 2007) and work on health utility measures (Juniper et al., 1997, Feeny et al., 2004). Most HRQL findings to date have focused on cross-sectional studies, but as this field continues to evolve and instruments are used more routinely in clinical trials, then more literature in this area should begin to emerge (Landgraf, 2005). Moreover, it has been argued that the experiences and health concepts that a child can comprehend will not only be related to their age but also to their social context involving family, peer-relationships and community factors (Pal, 1996, McNunn et al., 2001). This may include the impact of the child's treatment on the HRQL of the care-giver (Clarke and Eiser, 2004). This view is upheld by the World Health Organisation Quality of Life Assessment Group (1996) who also include the cultural perspective in their definition of HRQL. It is an important consideration since whereas adults have a choice as to whether they change their environment, children have less power to change a problematic situation. Another issue is the youngest age at which children can reliably report their HRQL, and when proxy respondents such as the parent, carer or doctor should be used (Janse et al., 2004). Eiser and Morse (2001b, 2001c) found greater heterogeneity in measures of social and emotional compared to physical functioning between parents and children and advocated the use of parallel forms for completion by both child and parents whenever possible until there is more conclusive evidence as to which informant is more reliable. It may be that the different perspectives of the parent/carer and the child

interact in some way such that cross informant discrepancies may have important HRQL significance.

The appropriate choice of outcome from questionnaire items and subscales is an added complexity. Poor item performance may reduce the instrument's effectiveness. Multiple testing of many items can result in inflated Type I errors, and combination of subscales can affect statistical efficiency in terms of relative effect size (Vickers, 2004). Wittes (2002) also identified problems of scaling in multidimensional instruments such that some parts of the instrument may contribute to the total score much more heavily than others creating imbalance and therefore a biased outcome. Sometimes for this reason global summary scores may be adopted. With child/carer-reported outcomes such as HRQL, an external criterion such as death or disease severity may be brought in to help interpret the magnitude of change, for example, survival with HRQL assessment (Patrick and Chiang, 2000). However, Patrick and Chiang question the extent to which statistical methods for combining outcomes to obtain a net measure of effectiveness are successful, particularly in validation and interpretability, and this needs further debate.

5. Reference values

The availability of reference values (or 'norms') for comparison with measured physiological values, for example, serum immunoglobulin concentrations in pre-school children (Altman, 1991), and established 'cut-offs' for identifying conditions such as fever in infants (Jones et al., 1993, Dodd et al., 2006) is an essential requirement in clinical studies. The need for comprehensive interpretation strategies when using questionnaire-based rating scales has already been highlighted (Marquis et al., 2004). An instrument result is therefore not clinically meaningful or useful if appropriate data for comparison are not available. In their review of child outcome measures used in child care quality research, Zaslow et al. (2006)

describe ten methodological concerns, highlighting that many instruments were not established measures, which makes it difficult to clearly relate the content and gauge the strength of the measures across studies. They also highlighted poor reporting of validity and reliability information either in general or for an adapted instrument or in the culture in which the instrument was applied.

To obtain reference data in the simplest case, the instrument is applied to a large sample of healthy children, and the mean and standard deviation are used to calculate a 95% reference interval, containing the middle 95% of the distribution of values found in healthy individuals. However, it is important to remember that a result outside the corresponding health-related reference interval does not necessarily imply that the child is diseased or at risk. This simple approach assumes that the data are normally distributed, or that a suitable transformation will create the desired effect. Since children vary in their stages of development, results are often dependent on age and gender, and so it may be necessary to have separate reference intervals for different age and gender groups. If data are not normally distributed then percentile values, usually the 2.5^{th} and 97.5^{th} centiles, can be calculated directly without any distributional assumptions (Altman, 1991). Examples of reference ranges in paediatric rheumatology are given in Nugent et al. (2001). Large numbers are required for these types of studies with the minimum number needed generally at least 500. Bland (2000) gives more information on sample size calculation. Alternative questionnaire-type approaches have used logistic regression to develop reference values for assessment tools. Gladstone et al. (2008) examined a range of developmental items (assessed as 'pass' or 'fail') using logistic regression and, for badly fitting models, triple fit spline regression (Greenland, 1995) to establish norms for the age at which 25%, 50%, 75% and 90% of children in Malawi passed that item. Figure 1 shows how useful plots of the model fits were for judging item performance. However, sometimes more sophisticated methods may be warranted. For

example, in revising the scoring system for the Griffiths assessment tool (Luiz et al., 2006), age-specific reference values (created in the original manual using simple linear regression) were more precisely constructed using the LMS method, described below.

For population-based assessment, typically in nutritional surveillance, the z-score is widely recognized as the best system for analysis and presentation of anthropometric data because of its advantages compared to other methods. In the WHO global database on child growth and malnutrition (de Onis and Blössner, 2003), for example, weight-for-height, height-for-age and weight-for-age are interpreted by using the z-score classification system. The anthropometric values are expressed as a number of standard deviations or z-scores below or above the reference mean (or median) value for the age and gender. The scale is linear and a fixed z-score interval implies a fixed height or weight difference for children of a given age and gender, making results comparable across groups. A major advantage is that a group of z-scores can be subjected to summary statistics such as the mean and standard deviation to compare and contrast children's growth status between groups.

BMI is a common measure of weight adjusted for height that may be used to diagnose overweight and obesity (Duran-Tauleria et al., 1995), or to assess poor nutritional status (Cleary et al., 2004), and has been shown to have a U-shaped association with death (Wong et al., 2000). Although it has been criticised because it does not distinguish overweight due to excess fat mass from overweight due to excess lean muscle mass, it does correlate with more direct fat measures and is the most commonly used measure for use in screening large populations (Must and Anderson, 2006). BMI z-scores are measures of relative weight adjusted for a child's age and sex, and are useful for measuring change over time. Because data in children are usually skewed, the International Obesity Task Force (Cole et al., 2000) used the LMS (Lambda, Mu, Sigma) method developed by Cole and Green (1991) to create

BMI z-scores. This is actually a special case of a more generalised additive approach (Rigby and Stasinopoulos, 2005), that also incorporates fractional polynomials (Royston and Wright, 1998). Other anthropometric measures such as waist circumference, have been advocated in addition to BMI, particularly as the waist circumferences of British children increased more than their BMI from 1987 to 1997, suggesting that BMI alone may not provide the full picture in relation to changes in body composition and obesity-related health (Must and Anderson, 2006).

Collecting reference data from hard to access populations

In general population studies it is sometimes difficult to find the subjects of interest because of the sensitivity of the topic, for example, child sexual abuse, or because an accurate diagnosis requires a detailed interview, as when measuring depression or parental neglect. The choice of using interviewer-based techniques as opposed to postal self-completion questionnaires has resource implications as well as issues of data quality. Interviewer-based questionnaires are costly and time-consuming to conduct but may obtain more detailed information, whilst postal questionnaires are cheaper at the risk of some inaccuracy or misinterpretation. Two-phase study designs (Dunn et al., 1999) have been advocated as a way of reducing interviewer costs. In the first phase of the study an initial screen of a large sample of the population is made utilising one or more postal questionnaires. In the second phase, the results of the first phase are used to stratify the subjects, then subsamples are taken from each strata and used for more detailed assessments involving an interviewer. By incorporating probability sampling weights reference value estimates can be calculated for the general population sample using the smaller stratified sample from the second phase (Thompson et al., 2001). However, this may cause difficulties if the screening questionnaire and interview methods are thought to measure different constructs. For example, in the Wirral Women's Health Survey, parenting style was measured by a postal screening questionnaire, and then a smaller stratified sample had a detailed interview by a trained expert to measure parental

neglect (Hill et al., 2001). As a consequence it was of interest to see whether the two methods were measuring the same underlying construct of parental neglect (Lancaster et al., 2007).

6. Discussion

In this paper we have attempted to take the reader through the myriad of methods, procedures and study designs necessary for development and assessment of efficient child health outcome measures. We have highlighted through examples the usefulness of these procedures for achieving specific objectives, and ultimately for achieving methodological rigour in clinical outcome studies performed in the paediatric population. We have also highlighted specific child related problems requiring special consideration.

Child Mental Health

Child mental health can often be overlooked when assessing child health outcomes, with most attention given to quality of life and quality of care. Yet children with chronic disease and disability have increased susceptibility to psychiatric disorder and social adjustment problems (Cadman et al., 1987). Psychiatric disorders and abnormalities of emotions, behaviour or hyperactivity are present in approximately 10% of children and adolescents in the general population (Meltzer et al., 2000). We have already seen some of the difficulties of making psychological assessments in a general paediatric context, and in child mental health the problems are compounded, particularly since the majority of children in the community with disorders are not under the care of psychiatric services (Rutter et al., 1970) and therefore need to be identified from within the general population before they can be studied.

In their review of outcome measures for child and adolescent mental health services, Hunter et al. (1996) found that the majority of outcome measurement tools understandably focused on recognising and diagnosing a problem, or on aspects of symptom intensity, levels of functioning, or quality of parenting, and that there was a dearth of tools that could be used in

routine clinical practice to cover all the important areas necessary to meaningfully rate the success of an intervention. In this context, the case characteristics (diagnosis, severity of associated disability) and case complexity (associated parental, family, medical, educational and social factors which may have an important influence on provision of treatment) may all influence the effectiveness of an intervention and so need measuring and accounting for in the analysis.

Health-related Quality of Life

The Convention on the Rights of the Child emphasised the child's right to adequate circumstances for physical, mental, spiritual and social development (United Nations, 1989). HRQL is an important health outcome for assessing a child's well-being, particularly when they are suffering from illnesses that require taking medicines. It has been increasingly used in adult randomised controlled trials to assess the impact of new and expensive treatments (Spiegelhalter et al., 1992). However, to date little attention has been given to a child's HRQL outcome, with most studies focussing on treatment efficacy and safety (Clarke and Eiser, 2004, Matza et al., 2004). In a review of HRQL assessment in paediatric oncology, only 3% of paediatric cancer clinical trials reported a HRQL assessment (Bradlyn et al., 1995). Barriers to the inclusion of a HRQL outcome have included attitudinal bias against using questionnaires on self-reported health, confusion as to which measure to use in a particular situation, the absence of a gold standard, burden and cost of assessment (Deyo and Patrick, 1989), and the physicians lack of confidence in use of procedures for detection in the case of emotional disorders (Snaith, 2003).

HRQL tools are multidimensional instruments designed to integrate a broad range of outcomes, for example, physical functioning, psychological well-being and social functioning. A plethora of generic and disease-specific HRQL instruments exist for use in adults but relatively fewer have been adapted for children (Eiser and Morse, 2001a, Harding,

2001). In Eiser and Morse's (2001c) extensive review of HRQL measures used on children they found that in tools developed in adult populations and adapted for children, certain parts of several different tools might be selected, and new questions added to construct an adapted instrument. As we have mentioned before, this may alter the psychometric properties of the tool and even render it invalid or unreliable for use in the paediatric population (Clarke and Eiser, 2004).

Health utility measures

Utility measures are an alternative way of 'summarising' an individual's well-being by allocating a single score to indicate a person's preference for a particular health state or outcome (EuroQol Group, 1990, Petrou, 2003), and are used widely in health economics. They have been used on asthmatic children as young as 8 years using a 'feeling thermometer' (Juniper et al., 1997) and 12 years in a study of teenage survivors of extremely low birth weight using a standard gamble lottery approach, where the children have to choose between an intermediate certain health state, and a lottery ranging somewhere between perfect health and a least preferred state, to determine the point at which they become indifferent to getting the lottery or the sure thing (Feeny et al., 2004). This approach has, however, received criticism because of its lack of validation and inadequate conceptual basis (Carr-Hill and Morris, 1991). However, it does provide a different perspective on measuring health that may be of potential benefit to children (Petrou, 2003).

In conclusion, health technologies that are used on children must demonstrate their effectiveness, be shown to be safe and to have limited adverse effects. They should also have no detrimental impact on the well-being of the child and their family. To determine the success of interventions, methods of outcome assessment must be accurate and reliable, be able to measure responsiveness to change over time within and between children, and have good reference data available.

Acknowledgements

Many thanks to the referees and associate editor for their very helpful and useful comments. Thank you also to Melissa Gladstone for the data and Ashley Jones for plotting the graphs for the figure.

References

Altman, D.G. (1991) *Practical statistics for medical research*. London: Chapman and Hall/CRC.

Asmussen, L., Olson, L.M., Grant, E.N., Fagan, J. and Weiss, K.B. (1999) Reliability and validity of the Children's Health Survey for asthma. *Pediatrics*, **104**(6), e71 (Available from http://pediatrics.aappublications.org/cgi/reprint/104/6/e71).

Bartholomew, D., Steele, F., Moustaki, I. and Galbraith, J. (2008) *Analysis of multivariate social science data, 2nd edition*. London: Chapman and Hall/CRC.

Bassi, Z., Watling, R., Dalzell, M., Lancaster, G. and Rosenbloom, L. (2004) Nutritional intake and growth in children with neurodisability: a 6 month cohort. *Archives of Disease in Childhood*, **89** (Supplement 1), A21-22.

Beyer, J. (1998) Key issues surrounding the assessment of pain in children. *Paediatric and Perinatal Drug Therapy*, **2,** 3-13.

Bland, M. (2000) *An introduction to Medical Statistics, 3rd edition*. Oxford: Oxford University Press.

Bland, J.M. and Altman, D.G. (1999) Measuring agreement in method comparison studies. *Statistical Methods in Medical Research*, **8,** 135-160.

Bradlyn, A.S., Harris, C.V. and Spieth, L.E. (1995) Quality of life assessment in pediatric oncolcogy: a retrospective review of phase III reports. *Social Science and Medicine*, **41**, 1463-1465.

Cadman, D., Boyle, M., Szatmari, P. and Offord, D.R. (1987) Chronic illness, disability and mental social well-being: findings of the Ontario Child Health Study. *Pediatrics*, **79**, 805-813.

Carpenter, J., Kenward, M. and Vansteelandt, S. (2006) A comparison of multiple imputation and inverse probability weighting for analyses with missing data. *Journal of the Royal Statistical Society, Series A*, **169**, 571-584.

Carr-Hill, R. and Morris, J. (1991) Current practice in obtaining the 'Q' in QALYs – a cautionary note. *British Medical Journal*, **303**, 699-701.

Chinn, S. (1990) The assessment of methods of measurement. *Statistics in Medicine*, 9, 351-362.

Chwalow, A.J. and Adesina, A.B. (2002) Conception, development and validation of instruments for quality of life assessment: an overview. In *Statistical methods for quality of life studies* (eds M. Mesbah, B.F. Cole and M.-L. Ting-Lee), pp63-70. Netherlands: Kluwer Academic Publishers.

Clarke, S.-A. and Eiser, C. (2004) The measurement of health-related quality of life (QOL) in paediatric clinical trials: a systematic review. *Health and quality of life outcomes*, **2**, 66.

Cleary, G., Lancaster, G.A., Annan, F., Sills, J.A. and Davidson, J.E. (2004) Nutritional status of children with juvenile idiopathic arthritis. *Rheumatology*, **43**, 1569-1573.

Cohen, J. (1960) A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20**, 37-46.

Cohen, J. (1988) *Statistical power analysis for the behavioural sciences, 2nd edition.* Hillsdale, NJ: Lawrence Erlbaum.

Cole, D.A., Truglio, R. and Peeke, L. (1997) Relation between symptoms of anxiety and depression in children: a multitrait-multimethod-multigroup assessment. *Journal of Consulting and Clinical Psychology*, **65**, 110-119.

Cole, T.J., Bellizzi, M.C., Flegal, K.M. and Dietz, W.H. (2000) Establishing a standard definition for child overweight and obesity worldwide: international survey. *British Medical Journal*, **320**, 1240-1243.

Cole, T.J. and Green, P.J. (1991) Smoothing reference centile curves: the LMS method and penalised likelihood. *Statistics in Medicine*, **11**, 1305-1319.

Commission of the European Communities (2004) Proposal for a Regulation of the European Parliament and of the Council on Medicinal Products for Paediatric use and amending regulation (EEC) No. 1768/92, Directive 2001/83/EC and regulation (EC) No. 726/2004. Brussels: Commission of the European Communities.

Coste, J., Fermanian, J. and Venot, A. (1995) Methodological and statistical problems in the construction of composite measurement scales: a survey of six medical and epidemiological journals. *Statistics in Medicine*, **14**, 331-345.

Costello, A.B. and Osborne, J.W. (2005) Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation*, **10** (7), 1-9. (Available from http://pareonline.net/pdf/v10n7.pdf)

Coulacoglou, C. (2002) Construct validation of the Fairy Tale Test - standardization data. *International Journal of Testing*, **2**, 217-241.

Craig, J.V., Lancaster, G.A., Williamson, P.R. and Smyth, R.L. (2000) Temperature measured at the axilla compared with the rectum in children and young people: systematic review. *British Medical Journal*, **320**, 1174-78.

Craig, J.V., Lancaster, G.A., Taylor, S., Williamson, P.R. and Smyth, R.L. (2002) Infrared ear thermometry compared with rectal thermometry in children: a systematic review. *Lancet*, **360**, 603-609.

Cuzick, J. (1985) A wilcoxon-type test for trend. *Statistics in Medicine*, **4**, 87-9.

DeLucia, C. and Pitts, S.C. (2006) Applications of individual growth curve modeling for pediatric psychology research. *Journal of Pediatric Psychology*, **31** (10), 1002-23.

Deyo, R.A and Patrick, D.L. (1989) Barriers to the use of health status measures in clinical investigation, patient care and policy research. *Medical Care*, **27** (3), S254.

Dodd, S.R., Lancaster, G.A., Craig, J.V., Smyth, R.L. and Williamson, P.R. (2006) In a systematic review, infrared ear thermometry for fever diagnosis in children finds poor sensitivity. *Journal of Clinical Epidemiology*, **59** (4), 354-357.

Drachler, M. de L., Marshall, T. and Carlos de Carvalho Leite, J. (2007) A continuous-scale measure of child development for population-based epidemiological surveys: a preliminary study using item response theory for the Denver. *Paediatric and Perinatal Epidemiology*, **21**, 138-153.

Dunn, G. (1992) Design and analysis of reliability studies. *Statistical Methods in Medical Research*, **1**, 123-157.

Dunn, G. (2000) *Statistics in Psychiatry*. London: Arnold.

Dunn, G., Pickles, A., Tansella, M. and Vazquez-Barquero, J. (1999) Two-phase epidemiological surveys in psychiatric research. *British Journal of Psychiatry*, **174**, 95–100.

Dunn, G. and Roberts, C. (1999) Modelling method comparison data. *Statistical Methods in Medical Research*, **8**, 161-179.

Duran-Tauleria, E., Rona, R.J. and Chinn, S. (1995) Factors associated with weight for height and skinfold thickness in British children. *Journal of Epidemiology and Community Health*, **49**, 466-473.

Eiser, C. and Morse, R. (2001a) A review of measures of quality of life for children with chronic illness. *Archives of Disease in Childhood*, **84**, 205-211.

Eiser, C. and Morse, R. (2001b) Can parents rate their child's health-related quality of life? Results of a systematic review. *Quality of Life Research*, **10**, 347-57.

Eiser, C. and Morse, R. (2001c) Quality of life measures in chronic diseases of childhood. *Health Technology Assessment*, **5**, 1-157.

Elashoff, J.D. (2003) *Nquery Advisor 5.0*. Ireland: Statistical Solutions Ltd. (Available from http://www.statsol.ie/)

Elphick, H.E., Lancaster, G.A., Solis, A., Majumdar, A., Gupta, R. and Smyth, R.L. (2004) Reliability and validity of acoustic analysis of respiratory sounds in infants. *Archives of Disease in Childhood*, **89**, 1059-1063.

Emde, R.N., Wolf, D.P. and Oppenheim, D. (eds) (2003) *Revealing the inner worlds of young children. The MacArthur Story Stem Battery and Parent-Child Narratives.* New York: Oxford University Press.

EuroQol Group (1990) EuroQol – a new facility for the measurement of health-related quality of life. *Health Policy*, **16**, 199-208.

Fairclough, D.L. (2004) Patient reported outcomes as endpoints in medical research. *Statistical Methods in Medical Research*, **13**, 115-138.

Farrell, M., Devine, K., Lancaster, G.A. and Judd, B. (2002) A method comparison study to assess the reliability of urine collection pads as a means of obtaining urine specimens from non-toilet trained children for microbiological examination. *Journal of Advanced Nursing*, **37** (4), 387-393.

Feeny, D., Furlong, W., Saigal, S. and Sun, J. (2004) Comparing directly measured standard gamble scores to HUI2 and HUI3 utility scores: group- and individual-level comparisons. *Social Science and Medicine*, **58**, 799-809.

Fleiss, J.L. and Cohen, J. (1973) The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, **33**, 613-19.

Fleiss, J.L., Levin, B. and Paik, M.C. (2003) *Statistical methods for rates and proportions 3rd edition***.** Hoboken, NJ: Wiley.

Freedman, L.S. (1987) Evaluating and comparing imaging techniques: a review and classification of study designs. *The British Journal of Radiology*, **60**, 1071-1081.

Gladstone, M., Lancaster, G.A., Jones, A.P., Maleta, K., Mtitimila, E.**,** Ashorn, P. and Smyth, R.L. (2008). Can Western developmental screening tools be modified for use in a rural Malawian setting? *Archives of Disease in Childhood*, **93**, 23-29.

Glasscoe, C., Quittner, A.L., Evans, J., Burrows, E.F., Cottrell, J.J., Heaf, L., Jones, S., Hope, H.F., Lancaster, G.A., Smith, J.A., Hill, J. and Southern, K.W. (2006a) Developing a tool to assess the impact on a family of caring for a child with cystic fibrosis. *Pediatric Pulmonology*, Supplement 29, 405-406.

Glasscoe, C., Smith, J.A., Hope, H.F., Jones, S., Cottrell, J.J., Burrows, E.F., Heaf, L., Evans, J., Lancaster, G.A., Quittner, A.L., Hill, J. and Southern, K.W. (2006b) The challenge of living with cystic fibrosis: a collective response from parents. *Pediatric Pulmonology*, Supplement 29, 406.

Gnecco, C. and Lachenbruch, P.A. (2002) Regulatory aspects of quality of life. In *Statistical methods for quality of life studies* (eds M. Mesbah, B.F. Cole and M.-L. Ting-Lee), pp9-19. Netherlands: Kluwer Academic Publishers.

Greenland, S. (1995) Dose-response and trend analysis in epidemiology: alternatives to categorical analysis. *Epidemiology*, **6** (4), 356-364.

Greenwood, D.C., Ransley, J.K., Gilthorpe, M.S. and Cade, J.E. (2006) Use of itemised till receipts to adjust for correlated dietary measurement error. *American Journal of Epidemiology*, **164**, 1012-1018.

Guggenmoos-Holzmann, I. and Vonk, R. (1998) Kappa-like indices of observer agreement viewed from a latent class perspective. *Statistics in Medicine*, **17**, 797-812.

Guyatt, G.H., Walter, S. and Norman, G. (1987) Measuring change over time, assessing the usefulness of evaluative instruments. *Journal of Chronic Disease*, **40**, 171-178.

Harding, L. (2001) Children's quality of life assessment, a review of generic and health-related quality of life measures completed by children and adolescents. *Clinical Psychology and Psychotherapy*, **8**, 79-96.

Helseth, S. and Slettebo, A. (2004) Research involving children, some ethical issues. *Nursing Ethics*, **11** (3), 298-308.

Hill, J., Fonagy, P., Lancaster, G.A. and Broyden, N. (2007). Aggression and intentionality in narrative responses to conflict and distress story stems, An investigation of boys with disruptive behaviour problems. *Attachment & Human Development,* 9 (3), 223-237.

Hill, J., Pickles, A., Burnside, E., Byatt, M., Rollinson, L., Davis, R. and Harvey, K. (2001) Child sexual abuse, poor parental care and adult depression, evidence for different mechanisms. *British Journal of Psychiatry*, **179**, 104–109.

Hodges, K. (1993) Structured interviews for assessing children. *Journal of Child Psycholology and Psychiatry*, **34**, 49-68.

Hunter, J., Higginson, I. and Garralda, E. (1996) Systematic literature review, outcome measures for child and adolescent mental health services. *Journal of Public Health Medicine*, **18** (2), 197-206.

Janse, A.J., Gemke, R.J., Uiterwaal, C.S., van der Tweel, I., Kimpen, J.L. and Sinnema, G. (2004) Quality of life, patients and doctors don't always agree, a meta-analysis. *Journal of Clinical Epidemiology*, **57**, 653-661.

Jones, R.J., O'Dempsey, T.J. and Greenwood, B.M. (1993) Screening for a raised rectal temperature in Africa. *Archives of Disease in Childhood*, **69**, 437-9.

Juniper, E.F., Guyatt, G.H., Feeny, D.H., Griffith, L.E. and Ferrie, P.J. (1997) Minimum skills required by children to complete health-related quality of life instruments for asthma, comparison of measurement properties. *European Respiratory Journal*, **10**(10), 2285-2294.

Kipnis, V., Subar, A.F., Midthune, D., Freedman, L.S., Ballard-Barbash, R., Troiano, R.P., Bingham, S., Schoeller, D.A., Schatzkin, A. and Carroll, R.J. (2003) Structure of Dietary Measurement Error, Results of the OPEN Biomarker Study. *American Journal of Epidemiology*, **158**, 14-21.

Kleinman, L., Rothman, M., Strauss, R., Orenstein, S.R., Nelson, S., Vandenplas, Y., Cucchiara, S. and Revicki, D.A. (2006) The infant gastroesophageal reflux questionnaire revised, development and validation as an evaluative instrument. *Clinical Gastroenterology and Hepatology*, **4**, 588-596.

Kwiterovich Jr, P.O., Barton, B.A., McMahon, R.P., Obarzanek, E., Hunsberger, S., Simons-Morton, D., Kimm, S.Y.S., Aronson Friedman, L., Lasser, N., Robson, A., Lauer, R., Stevens, V., Van Horn, L., Gidding, S., Snetselaar, L., Hartmuller, V.W., Greenlick, M. and Franklin Jr, F. (1997) Effects of Diet and Sexual Maturation on Low-Density Lipoprotein Cholesterol During Puberty, The Dietary Intervention Study in Children (DISC). *Circulation*, **96**, 2526-2533.

Lancaster, G.A., Dodd, S.R. and Williamson, P.R. (2004) Design and analysis of pilot studies, recommendations for good practice. *Journal of Evaluation in Clinical Practice*, **10** (2), 307-312.

Lancaster, G.A., Rollinson, L. and Hill, J. (2007) The measurement of a major childhood risk for depression, comparison of the Parental Bonding Instrument (PBI) 'Parental Care' and the Childhood Experience of Care and Abuse (CECA) 'Parental Neglect'. *Journal of Affective Disorders*, **101**, 263-267.

Landgraf, J.M. (2005) Practical considerations in the measurement of HRQoL in child/adolescent clinical trials. In *Assessing quality of life in clinical trials* (eds P. Fayers and R. Hays), pp339-367. Oxford: Oxford University Press.

Lijmer, J.G., Mol, B.W., Heisterkamp, S., Bonsel, G.J., Prins, M.H., van der Meulen, J.H. and Bossuyt, P.M. (1999) Empirical evidence of design-related bias in studies of diagnostic tests. *Journal of the American Medical Association*, **282**(11), 1061-6.

Luiz, D., Faragher, B., Barnard, A., Knoesen, N., Kotras, N., Burns, L. and Challis, D. (2006) *GMDS-ER Analysis Manual.* Oxford: Hogrefe (the Test Agency) and Association for Research in Infant and Child Development.

Marquis, P., Chassany, O. and Abetz, L. (2004) A comprehensive strategy for the interpretation of quality of life data based on existing methods. *Value in Health*, **7**(1), 93-104.

Matthews, J.N., Altman, D.G., Campbell, M.J. and Royston, P. (1990) Analysis of serial measurements in medical research. *British Medical Journal*, **300**, 230-5.

Matza, L.S., Swensen, A.R., Flood, E.M., Secnik, K. and Kline Leidy, N. (2004) Assessment of health-related quality of life in children, a review of conceptual, methodological, and regulatory issues. *Value in Health*, **7**(1), 79-92.

McNunn, A.M., Nazroo, J.Y., Marmot, M.G., Boreham, R. and Goodman, R. (2001) Children's emotional and behavioural well-being and the family environment, findings from the Health Survey for England. *Social Science and Medicine*, **53**, 423-440.

Meltzer, H., Gatward, R. with Goodman, R., Ford, T. (2000) *Mental Health Survey of Children and Adolescents in Great Britain*. London, The Stationery Office.

Must, A. and Anderson, S.E. (2006) Body mass index in children and adolescents, considerations for population-based applications. *International Journal of Obesity*, **30**, 590-594.

Muthén, L.K., and Muthén, B.O. (1998-2007) *Mplus User's Guide. Fifth Edition*. Los Angeles, CA: Muthén & Muthén. (Available from http://www.statmodel.com).

Nugent, J., Ruperto, N., Grainger, J., Machado, C., Sawhney, S., Baildam, E., Davidson, J., Foster, H., Hall, A., Hollingworth, P., Sills, A., Venning, H., Walsh, J.E., Landgraf, J.M., Roland, M., Woo, P. and Murray, K.J. for the Paediatric Rheumatology International Trials Organisation (PRINTO) (2001) The British version of the Childhood Health Assessment Questionnaire (CHAQ) and the Child Health Questionnaire (CHQ). *Clinical and Experimental Rheumatology*, **19,** Supplement 23, S163-S167.

de Onis, M. and Blössner, M. (2003) The World Health Organization Global Database on Child Growth and Malnutrition, methodology and applications. *International Journal of Epidemiology*, **32**, 518-526.

Pal, D.K. (1996) Quality of life assessment in children, a review of conceptual and methodological issues in multidimensional health status measures. *Journal of Epidemiology and Community Health*, **50**, 391-396.

Parsons, S.K., Shih, M.-C., DuHamel, K.N., Ostroff, J., Mayer, D.K., Austin, J., Martini, D.R., Williams, S.E., Mee, L., Sexson, S., Kaplan, S.H., Redd, W.H. and Manne, S. (2006) Maternal perspectives on children's health-related quality of life during the first year after pediatric hematopoietic stem cell transplant. *Journal of Pediatric Psychology*, **31**, 1100-1115.

Patrick, D.L. and Chiang, Y.-P. (2000) Measurement of health outcomes in treatment effectiveness evaluations, conceptual and methodological challenges. *Medical Care*, **38**, Supplement II, II14-II25.

Pepe, M.S. (2003) *The statistical evaluation of medical tests for classification and prediction*. Oxford: Oxford University Press.

Petrou, S. (2003) Methodological issues raised by preference-based approaches to measuring the health status of children. *Health Economics*, **12**, 697-702.

Poustie, V.J., Russell, J.E., Watling, R.M., Ashby, D. and Smyth, R.L., on behalf of the CALICO Trial Collaborative Group (2006) Oral protein energy supplements for children with cystic fibrosis, CALICO multicentre randomised controlled trial. *British Medical Journal*, **332**, 632-5.

Powell, C.V.E., McNamara, P., Solis, A. and Shaw, N.J. (2002) A parent completed questionnaire to describe the patterns of wheezing and other respiratory symptoms in infants and preschool children. *Archives of Disease in Childhood*, **87**, 376-379.

Ransley, J., Donnelly, J.K., Khara, T.N., Botham, H., Arnot, H., Greenwood, D.C. and Cade, J.E. (2001) The use of supermarket till receipts to determine the fat and energy intake in a UK population. *Public Health Nutrition*, **4**(6), 1279-1286.

Rigby, R.A. and Stasinopoulos, D.M. (2005) Generalized additive models for location, scale and shape. *Applied Statistics*, **54**, 507-554.

Royston, P. and Wright, E.M. (1998) A method for estimating age-specific reference intervals ('normal ranges') based on fractional polynomials and exponential transformation. *Journal of the Royal Statistical Society Series A*, **161**, 79-101.

Rutter, M., Tizard, J. and Whitmore, K. (1970) *Education, health and behaviour*. London: Longman.

Schafer, J.L. and Graham, J.W. (2002) Missing data: our view of the state of the art. *Psychological Methods*, **7** (2), 147-177.

Singer, J.D. and Willett, J.B. (2003). Applied Longitudinal Data Analysis. New York: Oxford University Press.

Skrondal, A. and Rabe-Hesketh, S. (2004) *Generalized latent variable modelling*. Boca Raton, Florida: Chapman and Hall/CRC.

Snaith, R.P. (2003) The hospital anxiety and depression scale. *Health and Quality of Life Outcomes*, **1**, 29 (Available from http,//www.hqlo.com/content/1/1/29).

Snijders, T.A.B. and Bosker, R.J. (1999) *Multilevel analysis*. London: Sage Publications.

Spiegelhalter, D.J., Gore, S.M., Fitzpatrick, R., Fletcher, A.E., Jones, D.R. and Cox, D.R. (1992) Quality of life measures in health care. III, resource allocation. *British Medical Journal*, **305**, 1205-1209.

Stewart, B., Lancaster, G.A., Lawson, J., Williams, K. and Daly, J. (2004) Validation of the Alder Hey triage pain score. *Archives of Disease in Childhood*, **89**, 625-30.

Streiner, D.L. and Norman, G.R. (2003) *Health measurement scales, 3rd edition*. Oxford: Oxford University Press.

Thompson, M.L., Edland, S.D., Gibbons, L.E. and McCurry, S.M. (2001) Estimating reference ranges from stratified two-stage samples. *Journal of the Royal Statistical Society Series A*, **164** (3), 505-516.

United Nations, Centre for Human Rights (1989) *Convention on the rights of the child*. Geneva: United Nations.

Varni, J.W., Burwinkle, T.M., Seid, M. and Skarr, D. (2003) The PedsQL™ 4.0 as a pediatric population health measure, feasibility, reliability, and validity. *Ambulatory Pediatrics*, **3**, 329-341.

Varni, J.W., Limbers, C.A. and Burwinkle, T.M. (2007) How young can children reliably and validly self-report their health-related quality of life?: An analysis of 8,591 children across age subgroups with the PedsQL 4.0 generic core scales. *Health and Quality of Life Outcomes*, 5 (1). (Available from http://www.hqlo.com/content/pdf/1477-7525-5-1.pdf)

Vermunt, J.K. and Hagenaars, J.A. (2004) Ordinal longitudinal data analysis. In *Methods for human growth research* (eds R.C. Hauspie, N. Cameron and L. Molinari). Cambridge: Cambridge University Press.

Vickers, A.J. (2004) Statistical considerations for use of composite health-related quality of life scores in randomised trials. *Quality of Life Research*, **13**, 717-723.

Vickers, A.J. and Altman, D.G. (2001) Statistics notes, analysing controlled trials with baseline and follow up measurements. *British Medical Journal*, **323**, 1123-1124.

Wittes, J. (2002) The use of soft endpoints in clinical trials, the search for clinical significance. In *Statistical methods for quality of life studies* (eds M. Mesbah, B.F. Cole and M.-L. Ting-Lee), pp129-140. Netherlands: Kluwer Academic Publishers.

Wong, C.S., Gipson, D.S., Gillen, D.L., Emerson, S., Koepsell, T., Sherrard, D.J., Watkins, S.L. and Stehman-Breen, C. (2000) Anthropometric measures and risk of death in children with end-stage renal disease. *American Journal of Kidney Diseases*, **36**(4), 811-819.

Wong, D. and Baker, C. (1995) *Reference manual for the Wong-Baker faces pain rating scale*. Duarte, California: Mayday Pain Resource Center.

World Health Organisation Quality of Life Assessment Group (WHOQOL) (1996) What is quality of life? *World Health Forum*, **14**, 354-6.

Wyrwich, K.W., Nienaber, N.A., Tierney, W.M. and Wolinsky, F.D. (1999) Linking clinical relevance and statistical significance in evaluating intra-individual changes in health-related quality of life. *Medical Care*, **37**(5), 469-478.

Young, J.G., O'Brien, J.D., Gutterman, E.M. and Cohen, P. (1987) Research on the clinical interview. *Journal of the American Academy of Child and Adolescent Psychiatry*, **26**, 613-620.

Zaslow, M., Halle, T., Martin, L., Cabrera, N., Calkins, J., Pitzer, L. and Geyelin Margie, N. (2006) Child outcome measures in the study of child care quality. *Evaluation Review*, **30**(50), 577-610.

Zhou X.-H., Obuchowski N.A. and McClish D.K. (2002) *Statistical methods in diagnostic medicine*. New York: John Wiley and Sons.

Table 1: Main stages in the development of a health outcome measure for use with children

Stage	Issues	Study design	Methods of analysis
1. Developing the instrument			
(a) Physiological Device	Does this device agree sufficiently well with one in current use?	Method comparison study	[1]Mean difference, 95% limits of agreement [2]Kappa with 95% CI
	How do I handle repeated measures and duplicates?		[1]ANOVA or confirmatory factor analysis [1,2]Multi-level modelling
	How do I compare this device to a known gold standard?	Diagnostic test study	[1]ROC curve, Area under Curve with 95% CI [2]Sensitivity, specificity, positive predicted value, 95% CI
	Problems: poor study design eg. no device calibration, time lapse between sequential measurements; insufficient control of sources of variability in design and/or analysis; repeated measures treated as independent observations to boost small samples		
(b) Questionnaire	Creation of items	Focus groups, interviews, expert review	Qualitative analysis to identify themes and assess face, content validity
	Do the items (eg. measured on Likert scale) work reasonably well?	Pilot study to evaluate item performance on a small sample	Mean, minimum, maximum responses; % missing; floor and ceiling effects; item-total correlations
	Problems: lack of use of age-appropriate language and instrument formatting; range of vocabulary available in different languages may affect translation and scoring; poor item performance because of cultural misunderstandings		
2. Establishing reliability and validity			
Reliability	Does the instrument give stability and consistency of measurement?	Intra-rater/test-retest (using the same assessor on two different occasions) Inter-rater (using 2 different assessors on the same occasion)	[1]ICC (also mean difference, 95% limits of agreement may be useful) [2]Kappa, weighted kappa with 95% CI, latent class analysis
	Does the instrument have good internal properties (reliability)?	Item performance evaluation on larger sample	As in Stage 1 above for pilot study.
	(for questionnaire type instruments typically using dichotomous items or Likert scales)	Examine internal structure and number of domains	Cronbach's alpha, split-half coefficient, Kuder-Richardson 20 method [1,2]Factor analysis, [2]item-response analysis
	Problems: Incomplete data for occasions/assessors eg. because of upset child or parent did not return for 2nd assessment; sample size too small for robust factor analysis/item-response analysis		
Validity	Does the instrument measure what it is supposed to be measuring?	Convergent/divergent validity	[1]Correlation (Pearson or Spearman rank)
		Extreme groups; known-groups; concurrent validity using external criterion; predictive validity	[1]T-test, Mann-Whitney U, Cuzick test for trend, ANOVA, Kruskal-Wallis etc. [2]Chi-squared test
		Criterion validity (with gold standard)	As in Stage 1 above for diagnostic test study
	Problems: may not be feasible to assess all listed types of validity as may require too many assessments for a child to tolerate; no known gold standard criterion may be available for children		

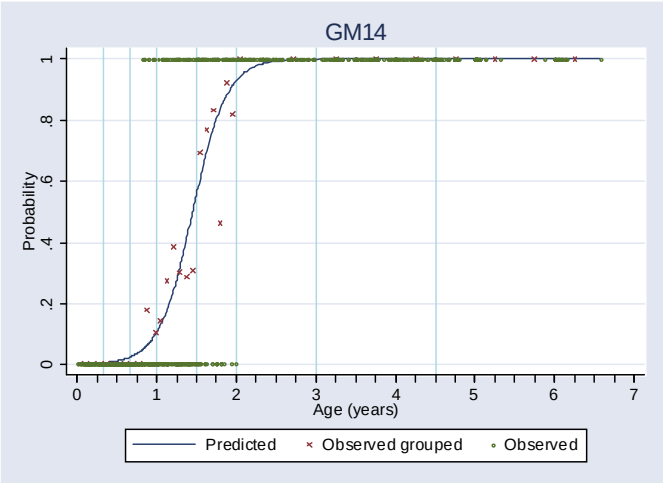
3. Measuring change over time and discriminating between subjects			
Magnitude of change	How do I assess the magnitude of change from baseline to endpoint?	Pilot study and/or main intervention study	[1]Cohen's effect size, Guyatt's responsiveness statistic, standard error of measurement
	Is there a better way of taking into account pre-test/baseline measures (eg. to avoid regression to mean)?	Randomised or non-randomised two group comparison on large sample	[1]ANCOVA to compare groups [2]Logistic/ordinal regression
Measurement error	How do I handle measurement error (particularly in dietary data)?	Take additional measurements using a reference instrument	Use reference instrument (eg. diet diary, biomarker, till receipts) to calibrate/adjust results
Multiple measurements	How do I account for multiple measurements taken over time?		Select the baseline and most important follow up time point only and analyse as above; Use summary measures (eg. mean, peak, area under curve, gradient) to obtain one measure for each child; [1,2]Longitudinal data analysis, growth curve modelling
	Problems: Establishing meaningful cut-offs and magnitudes of change that can be interpreted in different groups of children; missing data especially when follow up visits are necessary for sick children; use of proxy respondents and age-specific questionnaires; scaling of items		
4. Reference values for a healthy population			
Reference range	How do I create reference values for a normal, healthy population?	Take measurements from a large consecutive or random sample of children from school or community	[1]Mean, 95% reference range or [1]Median, 2.5th and 97.5th centiles
Gender-specific	What if the measurements vary by gender (usually identified by a bimodal distribution)?	Separate the sample measurements into those for boys and girls	Calculate reference range for each gender group separately
Age-specific	What if the measurements vary for children of different ages?	Separate sample measurements into age groups (could take quota sample); Use whole sample	Calculate age-specific reference ranges for each age group separately [1]Linear regression, fractional polynomials [2]Logistic regression, regression splines
Z-scores	How do I measure and compare deviations from the average across different groups of children?		Express measurements in terms of z-score units from the mean (ie. reference mean for that age and gender) and compare summary statistics for z-scores across groups
Anthropometric measurements	Is there a way of dealing with variability in growth and non-linear change eg. in weight for age, weight for height, BMI?		Z-scores, LMS (Lambda, Mu, Sigma) method (skewed data), GAMLSS (Generalised Additive Models for Location, Shape and Scale)
	Problems: Need large samples of children from reference population; need established reference means or medians to calculate z-scores; more sophisticated methods (eg. LMS, GAMLSS) require specialist software		

[1]measurements taken on a continuous scale (can include ordinal data with multiple categories) satisfying the assumptions of the method, [2]binary and/or ordered categorical measurements
CI=confidence interval

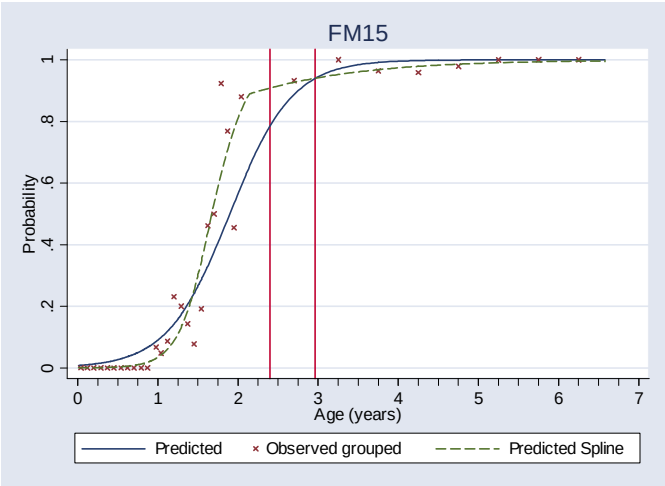
Table 2: Biases found in the design of method comparison studies and diagnostic test studies

Type of bias	**Reason**
Reference standard bias	Use of invalid reference standard
Spectrum bias and selection bias	Inappropriate patient sample or sampling technique
Review bias	Unblinded comparison
Verification bias	Gold standard test not applied to all patients eg. if invasive, costly, difficult to perform
Treatment paradox bias and disease progression bias	Measured value alters between tests either because patients are treated or there is a time delay before 2nd test
Measurement error	Variability in conduct/interpretation of test eg.instruments not calibrated, insufficient training of raters

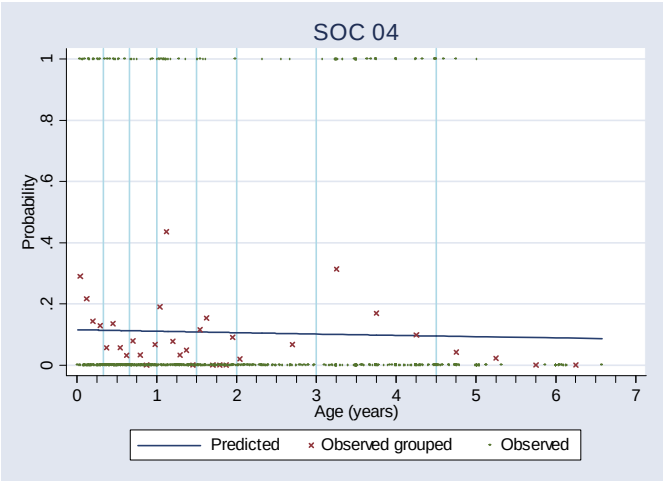
Figure 1: Examples of questions that performed well (a) fitted by logistic regression (b) fitted by triple joint spline regression and questions that performed poorly because (c) family dependent (d) badly worded, misunderstood



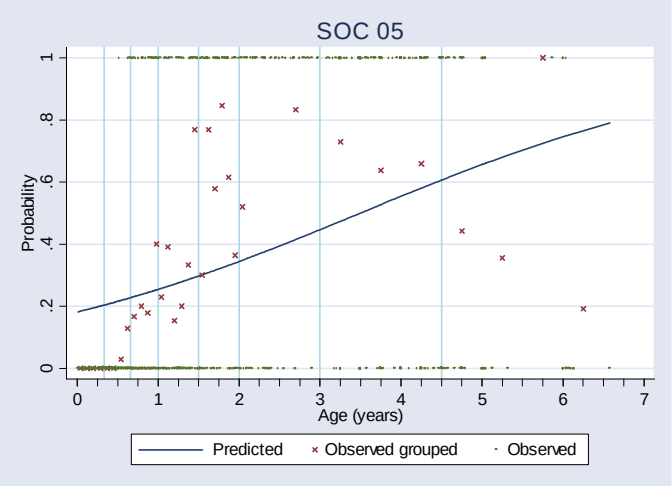
(a) Gross Motor item 'Walks well by self'



(b) Fine Motor item 'Builds tower of two cubes'



(c) Social item 'Spends most of time on mum's back'



(d) Social item 'Shy with strangers'