Developmental Assessments in Preterm Children: A Meta-Analysis

Hilary S Wong[a], PhD, Shalini Santhakumaran[b], MSc, Frances M Cowan[c], PhD, Neena Modi[c], MD, on behalf of the Medicines for Neonates Investigator Group

Affiliations: [a]Department of Paediatrics, University of Cambridge, [b] Imperial College Clinical Trials Unit, Imperial College London and [c]Section of Neonatal Medicine, Imperial College London, United Kingdom

Address correspondence to: Professor Neena Modi, Section of Neonatal Medicine, Imperial College London, Chelsea & Westminster Hospital campus, 369 Fulham Road, London SW9 1NH, United Kingdom, n.modi@imperial.ac.uk

Short title: Developmental Assessments in Preterm Children

Funding Source: This paper represents independent research funded by the National Institute for Health Research (NIHR) under its Programme Grants for Applied Research Programme (Reference RP-PG-0707-10010). The views expressed are those of the author and not necessarily those of the NHS, the NIHR or the Department of Health.

Financial Disclosure: The authors have no financial relationship to disclose.

Conflict of Interest: The authors have no conflicts of interest to disclose.

PROSPERO (International prospective register of systematic reviews) registration: CRD42012002168

This manuscript has not been published previously and is being submitted only to Pediatrics. It will not be submitted elsewhere while under consideration and, should it be published in Pediatrics, it will not be published elsewhere – either in similar form or verbatim – without permission of the editors. All authors are responsible for reported research and have participated in the concept and design, analysis and interpretation of data, drafting or revising, and have approved this manuscript as submitted.

We acknowledge and agree to comply with Pediatrics standard publishing policy.

Abbreviations:
BSID – Bayley Scales of Infant Development
CI – confidence interval
DOR – diagnostic odds ratio
ELBW – Extremely Low Birth Weight
ESS – effective sample size
HSROC – hierarchical summary receiver operator characteristics
MDI – Mental Development Index
NPV – negative predictive value
PPV – positive predictive value
QUADAS-2 – Quality of Diagnostic Accuracy Studies version 2
SD – standard deviation
VLBW – Very Low Birth Weight

Contributors' Statements:

Hilary Wong: Dr Wong designed the study, conducted the systematic review and data collection, carried out the initial analyses, drafted the initial manuscript and approved the final manuscript as submitted.

Shalini Santhakumaran: Ms Santhakumaran conducted the analyses, reviewed and revised the manuscript and approved the final manuscript as submitted.

Frances Cowan: Professor Cowan coordinated the data collection from authors of studies included in the review, critically reviewed the manuscript, and approved the final manuscript as submitted.

Neena Modi: Professor Modi critically reviewed the manuscript and approved the final manuscript as submitted.

ABSTRACT

Context: Developmental outcomes of very preterm (gestational age ≤32 weeks) or very low birthweight (<1500g) children are commonly reported before age 3 years though the predictive validity for later outcomes are uncertain.

Objective: To determine the validity of early developmental assessments in predicting school-age cognitive deficits.

Data sources: PubMed

Study selection: English-language studies reporting at least 2 serial developmental/cognitive assessments on the same population, one between ages 1-3 years and one at ≥5 years.

Data extraction: For each study, we calculated the sensitivity, specificity, positive and negative predictive values of early assessment for cognitive deficit (defined as test scores one SD below the population mean). Pooled meta-analytic sensitivity and specificity were estimated using hierarchical summary receiver operator characteristic curve.

Results: We included 24 studies (n=3133 children). Early assessments were conducted at 18-40 months and generally involved the Bayley Scales of Infant Development or the Griffiths Mental Development Scales; 11 different cognitive tests were used at school-age assessments at 5-18 years. Positive predictive values ranged from 20.0% to 88.9% and negative predictive vales ranged from 47.8% to 95.5%. The pooled sensitivity (95% CI) of early assessment for identifying school-age cognitive deficit was 0.55 (0.46-0.64) and specificity 0.84 (0.78-0.89). Gestational age, birthweight, age at assessment and time between assessments did not explain between-study heterogeneity.

Limitations: The accuracy of aggregated data could not be verified. Many assessment tools have been superceded by newer editions.

Conclusions: Early developmental assessment has poor sensitivity and but good specificity and negative predictive value for school-age cognitive deficit.

(250 words)

INTRODUCTION

The majority of outcome studies of preterm births report neurodevelopmental status at 18 or 24 months post-term age (corrected for prematurity). This practice stemmed from the emphasis of early outcome studies on measuring major disabilities such as severe mental retardation, sensori-neural hearing loss, blindness and cerebral palsy. It has generally been felt that at 18 to 24 months of age, a meaningful assessment of neurodevelopment can be reliably conducted while achieving a good follow-up rate.

However, there are developmental and maturation changes that affect the diagnostic accuracy of findings in the first two years. Follow-up studies into school-age and adolescence have regularly reported high rates of subtle disabilities that impact learning and social integration among seemingly 'non-disabled' survivors of preterm birth.[1-3] These 'high-prevalence/low-severity dysfunctions' include low average IQ scores, learning disabilities, attention and behavioural problems. Although assessment tools such as the Bayley Scales of Infant Development (BSID) provide standardised mental (cognitive) scores from as early as 12 months of age, the correlation of the early mental scores with subsequent IQ at school age is unclear. Two studies reported moderate to substantial agreement between BSID, second edition (BSID-II) Mental Development Index (MDI) at age 2 years and full scale IQ at age 5 years among infants born at less than 30 weeks gestation or with very low birth weight (VLBW; birth weight <1500g).[4,5] Conversely, Hack et al described a considerable reduction in the proportions of extremely low birth weight infants (ELBW; birth weight <1000g) who were diagnosed with cognitive impairment (defined as standardised cognitive scores <70) from 39% at 20 months to 16% at 8 years of age when the children were tested sequentially.[6] Applying the same diagnostic criteria, Roberts et al also found a reduction in the proportions of very preterm (gestational age <27 weeks) and ELBW infants with cognitive impairment, from 27.3% at age 2 years to 19.3% at age 8 years.[7]

We aimed to perform a systematic search of the published literature and review the evidence for the predictive validity of early developmental assessment, conducted between the ages of one to three years, for school-age cognitive deficit in children born very preterm or very low birth weight.

METHODS

This study was registered on international prospective register of systematic reviews PROSPERO (CRD42012002168).

Search strategy

We conducted a systematic electronic search on PubMed to identify relevant English-language literature published since 1st January 1990. Studies published prior to 1990 were not included in order to focus the review on more contemporaneous preterm populations. The following search terms were used both as keywords and subject headings: (combinations of "preterm" or "premature" with "infant" or "neonate" or "children") or ("low birth weight" or "extremely low birth weight") and ("cogniti*" or "neurodevelopment*" or "mental retardation" or "disability" or "intelligence" or "IQ"). The electronic search was supplemented by a manual search of the reference lists of studies that met the inclusion criteria.

Study selection

We sought to include all cohort and matched-control studies on study populations of infants born ≤32 weeks gestation or VLBW, in which at least two serial assessments, consisting of a developmental assessment between 1 - 3 years of age and a cognitive assessment at ≥5 years of age, were conducted and reported using validated standardised psychometric assessments (e.g. BSID, Wechsler Preschool and Primary Scale of Intelligence (WPPSI)). We included a birth weight limit as many neonatal studies used a birth weight-based selection criterion. The titles and abstracts of studies retrieved from the electronic search were screened to identify

relevant studies in the following three categories: (i) studies that reported both early developmental outcomes between ages 1 and 3 as well as school-age cognitive outcomes at ≥5 years, (ii) studies that only reported early developmental outcomes and (iii) studies that only reported school-age cognitive outcomes. In order to seek study populations that had received serial assessments in the relevant time periods but with their results at different ages published in separate articles, we matched the authors and study location of articles in (ii) and (iii) to identify publications on the same population at different time points. We conducted full-text evaluation for studies that satisfied the initial screening process for final inclusion in the review. Studies that only reported outcomes in language or executive function (e.g. memory) were excluded as they would not reflect the overall cognitive function of the study populations.

Data extraction

Data extracted included information on the study (study location, sampling method, eligibility criteria, sample size, attrition rates and developmental and cognitive assessment tools employed) and the study population (years of birth of participants, mean gestational age and birth weight, ages at developmental and cognitive assessments and mean test scores).

For this review, mild-moderate deficit was defined as developmental or cognitive test scores between 1 and 2 standard deviations (SD) below the standardised or control group means. Severe deficit was defined as test scores greater than 2 SD below the standardised or control group means. In studies where a control group of children born at full-term was recruited and assessed simultaneously, the mean and SD of the control group were used as the references for defining the presence of deficit. Data on the number of 'true-positive', 'false-positive', 'false-negative' and 'true-negative' mild-moderate and severe cognitive deficits identified by early assessments were collated from each study. Unpublished data were sought from study authors through email requests.

Quality assessment

The quality of included studies was assessed using a checklist adapted from the Quality of Diagnostic Accuracy Studies version 2 (QUADAS-2) appraisal tool.[8] The QUADAS-2 tool use 'signalling questions' to judge bias in four domains: patient selection, index test, reference standard, and flow of participants through the study and timing of the index test. The applicability of the study to the review question in the first 3 domains was also assessed. In the context of this review, the index tests referred to the early developmental assessments and the reference standards were the school-age cognitive assessments. Table 1 lists the signalling questions and the quality standards set for this review. By appraising against the set standards, each study was given a rating of 'low', 'high' or 'unclear' for risk of bias and concerns regarding applicability in each domain.

Statistical analysis

From each study, the estimated sensitivity, specificity, positive and negative predictive values (PPV, NPV) and the corresponding 95% confidence interval (CI) of identifying any and severe school-age cognitive deficits by early developmental assessments were calculated. Due to the variation in impairment prevalence across studies, meta-analyses on PPV and NPV, which are dependent on prevalence rates, were not performed. Separate pooling of sensitivities and specificities from the studies, which ignore the correlation between the two measures, could lead to an under-estimation of the diagnostic accuracy.[9] Instead, a hierarchical summary receiver operator characteristics (HSROC) curve was used for meta-analysis.[10] The HSROC model accounts for both within-study sampling variation and between-study heterogeneity using random effects. The output includes a summary operating point (pooled values for sensitivity and specificity) with 95% confidence region. Meta-regression was conducted using bivariate models to test for the possible association between sensitivity and specificity and the following study-level variables: mean gestational age, mean birth weight, mean ages at

assessments, time interval between assessments and earliest year of birth of participants. Associations with sensitivity and specificity were tested separately and likelihood ratio test was used to test both associations jointly. As there were too many different types of assessment tools used to be categorised into reasonably homogenous groups, subgroup analysis on the association between the types of assessment tool and diagnostic validity was not robust and therefore, not performed.

To investigate the possibility of publication and other sample size related effects, a funnel plot of the log diagnostic odds ratio (DOR) against 1/(effective sample size (ESS))$^{1/2}$ were tested for plot asymmetry using linear regression of the two variables, weighted by ESS.[11] The DOR is a statistic measure defined as (true-positives x true-negatives)/ (false-positives x false-negatives). The ESS is a function of the number of non-diseased (n_1) and diseased (n_2) participants, where ESS = $(4n_1\ n_2)/\ (n_1 + n_2)$.

All analyses were performed using Stata statistical package version 11.0 (StataCorp, Texas, USA) and SAS 9.3 (SAS Institute, North Carolina, USA).

RESULTS

The electronic literature search yielded 2844 unique citations; two additional studies were identified through manual search. The flow of articles through the search and selection process is depicted in the PRISMA diagram in figure 1. Fifty-four studies met the eligibility criteria. Data required for the review and meta-analysis were extractable directly from 6 articles. The authors of 18 of the remaining 48 studies contributed unpublished data. Therefore, 24 studies (37 articles)[4-7,12-44] were included in this review and their characteristics are detailed in Supplemental table 1. The characteristics of eligible studies that were not included (year of publication, countries where the studies were conducted) were similar to those included in the

review. For simplicity of referencing, studies that are represented by more than one article will be denoted by the first author and year of publication of the earliest article in tables and figures.

The study populations included 3133 children who were born at ≤32 weeks and/or had a birth weight <1500g. The study populations in four studies[12,17,27,29] consist of participants who exceed the gestational age and birth weight limits set in this review. For these studies, only the sub-population of participants who met our criteria was included in the analysis. The mean gestational ages at birth ranged from 25.0 to 33.1 weeks and the mean birth weights were between 675 and 1298g. 37.0% (1159 children) of the included populations was born in the years 1972-1990, 49.6% (1555 children) in 1991-2000 and 13.4% (419 children) in 2000-2005. Of note, there were 20 participants from the study of Cohen that were born in early 1970s. We have not excluded studies on the basis of the time period the participants were born in as it allowed for analysis of the variability of diagnostic validity over time. Children with known genetic syndromes and congenital anomalies were excluded from the studies. Children with severe neurosensory (including blindness and deafness) and motor impairment were likely to be under-represented in the cohort as 13 studies (contributing 55% of the final sample) excluded children who were unable to complete the assessments as a result of their physical disabilities. [5,12-15,17,19,21,29,33-35,37] The actual number of children excluded from the analysis for this reason is unknown as not all studies provided this information. Study participants were assessed between the ages of 18-40 months using the BSID in 13 studies, the Griffiths Mental Development Scales in 6 studies, the Stanford-Binet Intelligence Scale in 1 study and the Brunet-Lezine Scales in 1 study. In 3 studies,[16,23,29] more than one of these assessment tools were used. School-age cognitive assessments were conducted between the ages of 5-18 years and 11 different tests were used.

The proportion of children diagnosed with developmental impairment (test scores more than 1 SD below standardised or control group mean) varied widely among studies, ranging from

6.0% [14] to 67.0% [6]. The reported prevalence of school-age cognitive deficit was between 5.0% [17] and 67.4% [26] for mild-moderate (1-2 SD below mean) and 0.0% [17,18] and 37.8% [26] for severe impairment (>2 SD below mean). In five studies,[7,21,27,28,34] the categorisation of outcomes was based on the mean and SD of the scores achieved by concurrently recruited term-born controls. Wolke et al used cohort-specific cut-off points derived from a normative sample representative of the total population of infants in the Bavarian region to categorise impairments.[41] It should be noted that the study population in Smith et al[34] was from low- to middle-socioeconomic groups and the mean test scores achieved by the control group was about 0.5 SD below the normative mean. Using the results from the control group in this case could lead to an under-estimation of the prevalence of impairment in this study. If the test standardised norm values were used, the prevalence of cognitive impairment diagnosed at 8 years of age would increase from 24.0 to 36.0% for mild-moderate and from 6.0 to 6.6% for severe impairment.

Bias and applicability of included studies

The proportions of studies considered to be at 'low', 'high' and 'unclear' risk for bias and applicability concerns according to the QUADAS-2 appraisal are displayed in figure 2. The quality of individual study and the reasons for being considered at high risk for bias or concern for applicability are detailed in Supplemental table 2. The loss in follow-up of more than 30% of the eligible birth cohort was a main source of selection bias in the included studies. Whilst the overall risk of bias was low, the applicability of the results to our current population of preterm infants is concerning. This is because many of the included studies were conducted more than 20 years ago; the characteristics of the study populations would be different now and some of the assessment tools used have been superseded by newer versions.

Predictive validity of early developmental assessment

The sensitivities, specificities, PPV and NPV of early assessment for identifying any and severe cognitive deficit estimated from each study are presented in the forest plots in figure 3. In studies where participants were examined at different time points within the two age ranges we studied, only the results from the assessment performed at the oldest age are presented here. This gives a final sample size of 3060 children for the meta-analysis.

There was significant heterogeneity in the reported sensitivities and specificities among studies ($p<0.001$ for both). The estimated sensitivities of diagnosing any impairment ranged from 17.0% to 90.5%. There appears to be a wider range and poorer precision (wider confidence intervals) in the estimated sensitivity than specificity across studies. This may reflect the presence of heterogeneity or due to estimates of sensitivity being based on smaller samples than estimates of specificity. The HSROC curves providing the pooled measures are presented in figure 4. The summary points corresponded to a pooled sensitivity of 55.0% (95% CI 45.7 - 63.9%) and pooled specificity of 84.1% (77.5 - 89.1%) for the identification of any impairment. For the diagnosis of severe impairment, the pooled sensitivity was 39.2% (26.8 - 53.3%) and pooled specificity was 95.1% (92.3 - 97.0%). As the BSID-II was the most commonly used developmental test, a post-hoc meta-analysis of the subgroup of 11 studies that only used this tool for early developmental assessment showed a pooled sensitivity of 54.9% (39.5 – 69.3%) for any impairment and 43.6 (23.5 – 66.0%) for severe impairment; the corresponding specificities were 84.3% (70.1 – 92.5%) and 96.4% (90.0 – 98.8%). These values are similar to the results for the whole group.

None of the study-level variables examined (gestational age, birth weight, ages at assessments, time interval between assessments and year of birth) were associated with sensitivity or specificity and therefore did not explain the heterogeneity present between studies (table 2).

PPV estimates were most precise (narrower confidence intervals) in studies in which the prevalence for any impairment was greater than 40%, and ranged from 63.0% to 80.6%. For impairment prevalence less than 40%, PPV estimates for the prediction of any cognitive impairment were between 20.0% and 88.9%. In general, the NPV of early developmental assessments were high (range for 'any impairment' 47.8% to 95.5%, particularly in predicting the absence of severe impairment (NPV range for 'severe impairment' 68.9% to 100%.

Significance testing confirmed that asymmetry was not present in the funnel plot of the log DOR against the inverse of the square root of the ESS (figure 5, p=0.22), indicating the absence of sample size-related effects in the meta-analysis.

DISCUSSION

Through a systematic review of the literature, we found a substantial number of studies published in the past 20 years that have reported the early neurodevelopmental outcomes and later school-age cognitive abilities of children born very preterm or VLBW. Whilst early assessments were generally accurate in predicting the *absence* of school-age cognitive deficits (high NPV), the identification and prediction of children who would have cognitive difficulties were weak. Meta-analysis of the data suggested that almost half of children who might experience cognitive difficulties at school-age were classified as having normal neurodevelopmental function at ages 1-3 years. Even for cases of severe cognitive deficit, the accuracy in early detection was low (meta-analytic sensitivity of 39.2%).

This review sought to answer a clinically relevant question that for individual cohort studies, would involve lengthy follow-up and significant resources. One of the key strengths of the review is the systematic and comprehensive literature search that is highly sensitive in capturing all available data relevant to the research question in different settings. As the sensitivity estimates from individual studies were based on small number of participants with

cognitive impairment, the corresponding 95% CI were very wide. The use of a meta-analytic approach increases the sample size and improves the precision of the pooled estimate.

However, we recognise weaknesses in our study. It is possible that the included studies represent a biased sample as a large number of eligible studies were not included because of non-response, refusal or data no longer accessible. However, the non-included studies share similar study characteristics (year of publication, countries the studies were conducted, inclusion criteria, assessment tools) to those in the review and the funnel plot symmetry confirmed no publication bias by sample size. Hence there is no reason to presume that different conclusions would be drawn. We used available aggregated data and were unable to verify data accuracy. Although we had attempted to focus the review on studies published since 1990, only 14 of the 24 included studies recruited participants born after 1990 and none was born in the last 10 years. The past couple of decades have seen an overall reduction in the proportions of survivors of very preterm birth with adverse neurodevelopmental outcomes at age 2 years [45-47] so we can expect that the characteristics of the current preterm population to be different those from past eras. Furthermore, the assessment tools used in the included studies, although validated and contemporary at the time of each study, have mostly been superseded by newer editions. For example, the BSID is now in its third edition.[48] Recent studies have suggested that children achieve higher scores on the third edition of the Bayley Scales compared with the second edition when concurrently tested with both versions.[49,50] Therefore, caution should be exercised when extrapolating from results based on earlier versions of the assessment tools. Although psychometric properties differences exist, all the assessment tools provide comparative information of an individual's development in reference to age-appropriate normative data on the same scale.

We investigated the source of heterogeneity between studies using meta-regression. This method has a few drawbacks. The statistical power to detect associations between the study

estimates and the explanatory variables is related to the magnitude of the relationship between them, and is typically considered low in meta-regression.[51] This was compounded by the narrow range of values available for each of the explanatory variables under evaluation. For example, the mean gestational age of the included studies ranged from 25.9 to 33.1 weeks. Hence, we cannot exclude the possibility of a type II error. More importantly, meta-regression is subject to ecological fallacy (or aggregation bias). Therefore, in order to identify factors reliably that influence the validity of early developmental assessments, it would be necessary to utilise individual patient-level data.

Early intervention programs, initiated within the first 12 months of post-term life, are known to promote neurodevelopment among preterm infants.[52] We do not have the necessary information on whether study participants were offered or received early intervention to evaluate its effect. A Cochrane review of 25 randomized controlled trials of early intervention programs reported that the cognitive benefit observed in infancy and pre-school age did not persist into school-age. However, most of these programs terminate within the first year after birth and little work has been done on cognitive rehabilitation or training programs that are sustained beyond toddlerhood. Neurodevelopmental assessment at two or three years of age is often used as the endpoint for post-discharge follow-up of very preterm or VLBW infants. Depending on the diagnosis at this stage, children are either referred for further intervention and support or discharged from follow-up. Reassuringly, we found the false-positive rate for early diagnosis of impairment to be low. It is likely that children with more severe impairments would be correctly identified at this stage. However, children with milder impairments, who are harder to diagnose, may miss out on the potential advantages of cognitive intervention programs.

Cognitive function in infancy is a poor predictor of later IQ in the general population.[53] This may reflect real changes in cognitive function during childhood, unveiling of deficits in

complex task performance that were non-essential in early childhood, or the increasing effect of social and environmental influences on cognitive outcomes over time. Other explanations may be the impact of behaviour and attention during testing at different ages as well as the differences in the content and psychometric properties of early neurodevelopmental and later cognitive assessment tools. It has been reported that IQ scores from childhood to adulthood were more stable for very preterm/VLBW than term-born individuals, particular among those with severe cognitive impairment.[54]

In 2013, a meta-analysis similar to our study, on the predictive value of the BSID on later very preterm and/or VLBW outcomes was published by Luttikhuizen dos Santos et al.[55] They reported a strong positive correlation between BSID MDI in the first three years after birth and later cognitive scores (pooled correlation coefficient: 0.61, 95% CI 0.57 - 0.64) that accounted for 37% of the variance in cognitive functioning. There are several important methodological differences between this meta-analysis and our study. Only studies using the BSID were included in the Dutch meta-analysis and studies published before 1990 were not excluded. The meta-analysis incorporated early neurodevelopmental data obtained before the age of one year and nearly half of the follow-up data were based on testing before school-age. The convergent validity of MDI scores and cognitive scores may reflect the short interval between testing in this case. More crucially, the statistical measures used in our study (sensitivity and specificity) and the published meta-analysis (correlation coefficient) evaluate different test properties. Whilst sensitivity and specificity assessed the stability of diagnosis defined as a dichotomous variable, correlation coefficient measures the strength and direction of a linear relationship between two continuous variables. In a hypothetical scenario where the one-year BSID MDI always fall 20 points below the IQ measured at 10 years, the measured correlation would be perfect but the sensitivity would still be poor.

CONCLUSIONS

Early neurodevelopmental assessment has high specificity and NPV but low sensitivity in identifying later school-age cognitive deficit. A significant number of older children and adolescents born very preterm or VLBW experience difficulties in school and are a group that might have benefitted from earlier support and intervention had their cognitive deficits been recognised. We would encourage future studies of the factors affecting the diagnostic and predictive accuracy of early neurodevelopmental assessments, in order to identify follow-up schedules that have maximal likelihood of detecting impairment.

ACKNOWLEGEMENTS

Members of the Medicines for Neonates Investigator Group are Prof Deborah Ashby (Imperial College London, UK), Prof Peter Brocklehurst (University College London, UK), Prof Kate Costeloe (Queen Mary University of London, UK), Prof Elizabeth Draper (University of Leicester, UK), Prof Azeem Majeed (Imperial College London, UK), Prof Neena Modi (Imperial College London, UK) Prof Stavros Petrou (University of Warwick, UK), Prof Alys Young (University of Manchester, UK), Mrs Jane Abbott and Ms Zoe Chivers (Bliss Charity, London, UK) and Mrs Jacquie Kemp (London, UK). We thank the following investigators who contributed unpublished and supplemental data for the review: Dr Haim Bassan (Tel Aviv University, Israel), Prof Arend Bos (Beatrix Children's Hospital, Groningen, the Netherlands), Dr Marie-Laure Charkaluk (Lille Catholic University, Paris Descartes University, France), Dr Sarale Cohen (retired), Prof Olaf Dammann (Tufts University, Boston, USA), Prof Linda de Vries (University Medical Centre, Utrecht, the Netherlands), Prof Ermellina Fedrizzi (University of Padua, Italy), Prof Vineta Fellman (Lund University, Sweden), Prof Peter Gray (Mater Mothers' Hospital, Brisbane, Australia), Prof Howard Kilbride (University of Missouri-Kansas City, USA), Prof Neil Marlow (University College London, UK), Prof Jennifer Pinto-Martin (University of Pennsylvania, Philadelphia, USA), Prof Jon Skranes (Norwegian University of Science & Technology, Trondheim, Norway), Prof Karen Smith (The University of Texas Medical Branch, Galveston, USA), Prof Mary Sullivan (University of Rhode Island, Kingston, USA), Prof Paul Swank (The University of Texas Health Science Center at Houston, USA), Prof H. Gerry Taylor (Case Western Reserve University, Cleveland, USA), Dr Viena Tommiska (Helsinki University Central Hospital, Finland), Dr Norbert Veelken (Asklepios Klinik Nord, Hamburg, Germany), Prof Dieter Wolke (University of Warwick, UK) and Prof Lianne Woodward (Harvard Medical School, Boston, USA).

REFERENCES

1. Marlow N. Neurocognitive outcome after very preterm birth. *Arch Dis Child Fetal Neonatal Ed.* May 2004;89(3):F224-228.
2. Doyle LW, Anderson PJ. Adult outcome of extremely preterm infants. *Pediatrics.* Aug 2010;126(2):342-351.
3. Aylward GP. Cognitive and neuropsychological outcomes: more than IQ scores. *Mental retardation and developmental disabilities research reviews.* 2002;8(4):234-240.
4. Munck P, Niemi P, Lapinleimu H, Lehtonen L, Haataja L, PIPARI Study Group. Stability of cognitive outcome from 2 to 5 years of age in very low birth weight children. *Pediatrics.* Mar 2012;129(3):503-508.
5. Potharst ES, Houtzager BA, van Sonderen L, et al. Prediction of cognitive abilities at the age of 5 years using developmental follow-up assessments at the age of 2 and 3 years in very preterm children. *Developmental medicine and child neurology.* Mar 2012;54(3):240-246.
6. Hack M, Taylor HG, Drotar D, et al. Poor predictive validity of the Bayley Scales of Infant Development for cognitive function of extremely low birth weight children at school age. *Pediatrics.* Aug 2005;116(2):333-341.
7. Roberts G, Anderson PJ, Doyle LW, the Victorian Infant Collaborative Study Group. The stability of the diagnosis of developmental disability between ages 2 and 8 in a geographic cohort of very preterm children born in 1997. *Arch Dis Child.* Oct 2010;95(10):786-790.

8. Whiting PF, Rutjes AW, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Annals of internal medicine.* Oct 18 2011;155(8):529-536.
9. Deeks JJ. Systematic reviews of evaluations of diagnostic and screening tests. In: Egger M, Davey Smith G, Altman DG, eds. *Systematic reviews in health care: meta-analysis in context*. 2nd ed. London: BMJ; 2001:248-282.
10. Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Statistics in medicine.* Oct 15 2001;20(19):2865-2884.
11. Deeks JJ, Macaskill P, Irwig L. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. *Journal of clinical epidemiology.* Sep 2005;58(9):882-893.
12. Bassan H, Stolar O, Geva R, et al. Intrauterine growth-restricted neonates born at term or preterm: how different? *Pediatric neurology.* Feb 2011;44(2):122-130.
13. Bowen JR, Gibson FL, Leslie GI, Arnold JD, Ma PJ, Starte DR. Predictive value of the Griffiths assessment in extremely low birthweight infants. *J Paediatr Child Health.* Feb 1996;32(1):25-30.
14. Bruggink JL, Van Braeckel KN, Bos AF. The early motor repertoire of children born preterm is associated with intelligence at school age. *Pediatrics.* Jun 2010;125(6):e1356-1363.
15. Charkaluk ML, Truffert P, Marchand-Martin L, et al. Very preterm children free of disability or delay at age 2: predictors of schooling at age 8: a population-based longitudinal study. *Early Hum Dev.* Apr 2011;87(4):297-302.
16. Claas MJ, de Vries LS, Bruinse HW, et al. Neurodevelopmental outcome over time of preterm born children </=750 g at birth. *Early human development.* Mar 2011;87(3):183-191.
17. Cohen SE. Biosocial factors in early infancy as predictors of competence in adolescents who were born prematurely. *J Dev Behav Pediatr.* Feb 1995;16(1):36-41.
18. Fedrizzi E, Inverno M, Botteon G, Anderloni A, Filippini G, Farinotti M. The cognitive development of children born preterm and affected by spastic diplegia. *Brain Dev.* 1993 Nov-Dec 1993;15(6):428-432.
19. Gray PH, Burns YR, Mohay HA, O'Callaghan MJ, Tudehope DI. Neurodevelopmental outcome of preterm infants with bronchopulmonary dysplasia. *Arch Dis Child Fetal Neonatal Ed.* Nov 1995;73(3):F128-134.
20. Gray PH, O'Callaghan MJ, Rogers YM. Psychoeducational outcome at school age of preterm infants with bronchopulmonary dysplasia. *Journal of paediatrics and child health.* Mar 2004;40(3):114-120.
21. Gray D, Woodward LJ, Spencer C, Inder TE, Austin NC. Health service utilisation of a regional cohort of very preterm infants over the first 2 years of life. *Journal of paediatrics and child health.* Jun 2006;42(6):377-383.
22. Pritchard VE, Clark CA, Liberty K, Champion PR, Wilson K, Woodward LJ. Early school-based learning difficulties in children born very preterm. *Early human development.* Apr 2009;85(4):215-224.
23. Kilbride HW, Daily DK, Claflin K, Hall RT, Maulik D, Grundy HO. Improved survival and neurodevelopmental outcome for infants less than 801 grams birthweight. *American journal of perinatology.* Apr 1990;7(2):160-165.
24. Kilbride HW, Daily DK. Survival and subsequent outcome to five years of age for infants with birth weights less than 801 grams born from 1983 to 1989. *Journal of perinatology : official journal of the California Perinatal Association.* Mar-Apr 1998;18(2):102-106.

25. Marlow N, Wolke D, Bracewell MA, Samara M. Neurologic and developmental disability at six years of age after extremely preterm birth. *The New England journal of medicine.* Jan 6 2005;352(1):9-19.
26. Johnson S, Fawke J, Hennessy E, et al. Neurodevelopmental disability through 11 years of age in children born before 26 weeks of gestation. *Pediatrics.* Aug 2009;124(2):e249-257.
27. McGrath MM, Sullivan MC, Lester BM, Oh W. Longitudinal neurologic follow-up in neonatal intensive care unit survivors with various neonatal morbidities. *Pediatrics.* Dec 2000;106(6):1397-1405.
28. Orchinik LJ, Taylor HG, Espy KA, et al. Cognitive outcomes for extremely preterm/extremely low birth weight children in kindergarten. *Journal of the International Neuropsychological Society : JINS.* Nov 2011;17(6):1067-1079.
29. Reuss ML, Paneth N, Pinto-Martin JA, Lorenz JM, Susser M. The relation of transient hypothyroxinemia in preterm infants to neurologic development at two years of age. *N Engl J Med.* Mar 1996;334(13):821-827.
30. Pinto-Martin JA, Whitaker AH, Feldman JF, Van Rossem R, Paneth N. Relation of cranial ultrasound abnormalities in low-birthweight infants to motor or cognitive performance at ages 2, 6, and 9 years. *Dev Med Child Neurol.* Dec 1999;41(12):826-833.
31. Pinto-Martin J, Whitaker A, Feldman J, et al. Special education services and school performance in a regional cohort of low-birthweight infants at age nine. *Paediatric and perinatal epidemiology.* Mar 2004;18(2):120-129.
32. Lorenz JM, Whitaker AH, Feldman JF, et al. Indices of body and brain size at birth and at the age of 2 years: relations to cognitive outcome at the age of 16 years in low birth weight infants. *J Dev Behav Pediatr.* Dec 2009;30(6):535-543.
33. Skranes J, Vik T, Nilsen G, Smevik O, Andersson HW, Brubakk AM. Can cerebral MRI at age 1 year predict motor and intellectual outcomes in very-low-birthweight children? *Dev Med Child Neurol.* Apr 1998;40(4):256-262.
34. Smith KE, Landry SH, Swank PR. The role of early maternal responsiveness in supporting school-aged cognitive development for children who vary in birth status. *Pediatrics.* May 2006;117(5):1608-1617.
35. Tommiska V, Heinonen K, Kero P, et al. A national two year follow up study of extremely low birthweight infants born in 1996-1997. *Archives of disease in childhood. Fetal and neonatal edition.* Jan 2003;88(1):F29-35.
36. Mikkola K, Ritari N, Tommiska V, et al. Neurodevelopmental outcome at 5 years of age of a national cohort of extremely low birth weight infants who were born in 1996-1997. *Pediatrics.* Dec 2005;116(6):1391-1400.
37. Veelken N, Stollhoff K, Claussen M. Development of very low birth weight infants: a regional study of 371 survivors. *Eur J Pediatr.* Sep 1991;150(11):815-820.
38. Dammann O, Drescher J, Veelken N. Maternal fever at birth and non-verbal intelligence at age 9 years in preterm infants. *Developmental medicine and child neurology.* Mar 2003;45(3):148-151.
39. Vermeulen GM, Bruinse HW, de Vries LS. Perinatal risk factors for adverse neurodevelopmental outcome after spontaneous preterm birth. *European journal of obstetrics, gynecology, and reproductive biology.* Dec 1 2001;99(2):207-212.
40. Rademaker KJ, Uiterwaal CS, Groenendaal F, et al. Neonatal hydrocortisone treatment: neurodevelopmental outcome and MRI at school age in preterm-born children. *The Journal of pediatrics.* Apr 2007;150(4):351-357.

41. Wolke D, Meyer R. Cognitive status, language attainment, and prereading skills of 6-year-old very preterm children and their peers: the Bavarian Longitudinal Study. *Dev Med Child Neurol.* Feb 1999;41(2):94-109.
42. Gutbrod T, Wolke D, Soehne B, Ohrt B, Riegel K. Effects of gestation and birth weight on the growth and development of very low birthweight small for gestational age infants: a matched group comparison. *Arch Dis Child Fetal Neonatal Ed.* May 2000;82(3):F208-214.
43. Wolke D, Schulz J, Meyer R. Longterm developmental outcome of very prematurely born infants. Bavarian longitudinal study. *Monatsschrift Kinderheilkunde* 2001;149:S53-S61.
44. Saigal S, den Ouden L, Wolke D, et al. School-age outcomes in children who were extremely low birth weight from four international population-based cohorts. *Pediatrics.* Oct 2003;112(4):943-950.
45. Salt A, D'Amore A, Ahluwalia J, et al. Outcome at 2 years for very low birthweight infants in a geographical population: risk factors, cost, and impact of congenital anomalies. *Early human development.* Feb 2006;82(2):125-133.
46. Wilson-Costello D, Friedman H, Minich N, et al. Improved neurodevelopmental outcomes for extremely low birth weight infants in 2000-2002. *Pediatrics.* Jan 2007;119(1):37-45.
47. Doyle LW, Roberts G, Anderson PJ. Changing long-term outcomes for infants 500-999 g birth weight in Victoria, 1979-2005. *Archives of disease in childhood. Fetal and neonatal edition.* Nov 2011;96(6):F443-447.
48. Bayley N. *Technical manual for the Bayley Scales of Infant and Toddler Development (third edition)*. 3rd ed. San Antonio, TX: Psychological Corporation; 2006.
49. Moore T, Johnson S, Haider S, Hennessy E, Marlow N. Relationship between test scores using the second and third editions of the Bayley Scales in extremely preterm children. *The Journal of pediatrics.* Apr 2012;160(4):553-558.
50. Lowe JR, Erickson SJ, Schrader R, Duncan AF. Comparison of the Bayley II Mental Developmental Index and the Bayley III Cognitive Scale: are we measuring the same thing? *Acta paediatrica (Oslo, Norway : 1992).* Feb 2012;101(2):e55-58.
51. Bossuyt P, Davenport C, Deeks J, Hyde C, Leeflang M, Scholten R. Chapter 11: Interpreting results and drawing conclusions. In: Deeks JJ, Bossuyt PM, Gatsonis C, eds. *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 0.9.*: The Cochrane Collaboration; 2013.
52. Spittle A, Orton J, Anderson PJ, Boyd R, Doyle LW. Early developmental intervention programmes provided post hospital discharge to prevent motor and cognitive impairment in preterm infants. *Cochrane database of systematic reviews (Online).* 2015;11:CD005495.
53. Aylward GP. Prediction of function from infancy to early childhood: implications for pediatric psychology. *Journal of pediatric psychology.* Oct 2004;29(7):555-564.
54. Breeman LD, Jaekel J, Baumann N, Bartmann P, Wolke D. Preterm Cognitive Function Into Adulthood. *Pediatrics.* Sep 2015;136(3):415-423.
55. Luttikhuizen dos Santos ES, de Kieviet JF, Konigs M, van Elburg RM, Oosterlaan J. Predictive value of the Bayley scales of infant development on development of very preterm/very low birth weight children: a meta-analysis. *Early Hum Dev.* Jul 2013;89(7):487-496.

Supplemental table 1 Characteristics of studies included in review

Author year	Country	Population	Years of birth	Sampling method	Sample size	Mean (SD) or median (IQR) GA (weeks)	Mean (SD) or median (IQR) BW (grams)		Ages at assessments (month for early, years for school-age)	Assessment tools used	Mean (SD) assessment scores‡
Bassan 2011	Israel	BW <10th percentile for GA*	1992-1997	SC	32	33.1 (2.2)	1182 (229)	Early:	24m	BSID-II	95.8 (19.1)
								School-age:	6	WPPSI-R	103.4 (17.7)
Bowen 1996	Australia	BW <1000g	1985-1988	SC	45	27.6 (2.3)	864 (90)	Early:	12m, 36m	GMDS	-
								School-age:	5	S-B-IV	94.4 (11.2)
Bruggink 2010	the Netherlands	"Preterm"	1992-1997	SC	50	30.0 (1.9)	1184 (292)	Early:	19m	BSID-II	100.5 (11.2)
								School-age:	8	WISC-III	92.2 (10.6)
Charkaluk 2011	France	GA <33 weeks	1997	PB	313	29.8 (2.1)	1355 (406)	Early:	24m	Brunet-Lezine Revised	96.7 (12.7)
								School-age:	5	KABC	94.7 (18.7)
Claas 2011	the Netherlands	BW ≤750g and GA ≥24 weeks	1996-2005	SC	101	28.0 (24.8 - 34.4)†	675 (480 - 750)†	Early:	24m	BSID-II/ GMDS	-
								School-age:	5.5	WPPSI/ RAKIT/ SON-R	-
Cohen 1995	USA	"Preterm"*	1972-1974	SC	20	28.1 (2.1)	1111 (187)	Early:	24m	BSID	103.9 (21.1)
								School-age:	5, 8,12,18	S-B-III/WISC-R/ WAIS-R	101.8 (19.0)
Fedrizzi 1993	Italy	Spastic diglegia	1984-1991	SC	11	29.6 (1.6)	1474 (321)	Early:	36m	GMDS	72.6 (14.5)
								School-age:	6	WPPSI	76.4 (18.9)
Gray 1995	Australia	GA 23-33 weeks with diagnosis of BPD	1989-1990	SC	126	28.2	1065	Early:	24m	GMDS	108.5
								School-age:	8	WISC-III	90.5
Gray 2006	New Zealand	GA <32 weeks or BW <1500g	1998-2000	SC	99	27.8 (2.4)	1065 (321)	Early:	24m	BSID-II	86.1 (17.3)
								School-age:	6	WPPSI-R	95.4 (15.2)
Hack 2005	USA	BW <1000g	1992-1995	SC	200	26.4 (2)	811 (125)	Early:	20m	BSID-II	75.6 (16.0)
								School-age:	8	KABC	87.8 (19.0)
Kilbride 1990	USA	BW <801g	1983-1990	MC	129	25.9 (1.6)	698 (82)	Early:	12-24m, 36m	BSID/ S-B-III	84.4 (10.0)
								School-age:	5	S-B-III	85.7 (11.6)
	UK	GA <26 weeks	1995	PB	212	25.0 (0.7)	748 (116)	Early:	30m	BSID-II	81.7 (14.5)

Marlow 2005								School-age	6, 11	KABC	83.8 (18.0)
McGrath 2000	USA	BW <1850g with neonatal diagnoses*	1985-1989	SC	88	29.6 (2.2)	1200 (285)	Early:	18m	BSID-II	105.2 (19.0)
								School-age:	8	WISC-III	96.3 (18.4)
Munck 2012	Finland	BW <1500g	2001-2004	SC	124	28.7 (2.8)	1061 (260)	Early:	24m	BSID-II	101.2 (16.3)
								School-age:	5	WPPSI-R	99.3 (17.7)
Orchinik 2011	USA	GA <28 weeks or BW <1000g	2001-2003	SC	139	25.9 (1.6)	818 (174)	Early:	20m	BSID-II	77.2 (17.3)
								School-age:	6	BIA	86.3 (21.1)
Potharst 2011	the Netherlands	GA <30 weeks or BW <1000g	2003-2004	SC	100	28.7 (1.6)	1040 (253)	Early:	24m,36m	BSID-II	102.0 (14.0)
								School-age:	5	WPPSI-III	93.0 (17.0)
Reuss 1996	USA	BW 501-2000g*	1984-1987	MC	231	29.2 (2.9)	1142 (223)	Early:	24m	BSID/ S-B-III	-
								School-age:	6, 9, 16	S-B-IV/ WISC-III/ WASI	-
Roberts 2010	Australia	GA 22-27 weeks or BW 500-999g	1997	PB	186	26.5 (2.0)	832 (164)	Early:	24m	BSID-II	-
								School-age:	8	WISC-R	94.4 (14.2)
Skranes 1998	Norway	BW <1500g	1988	PB	21	29.0 (2.0)	1218 (193)	Early:	12m	BSID	99.0 (18.3)
								School-age:	6	WPPSI	96.0 (16.4)
Smith 2006	USA	BW <1500g from lower socio-economic groups	1990-1992	MC	161	29.7 (2.5)	1114 (267)	Early:	40m	S-B-IV	86.2 (10.6)
								School-age:	6, 8 ,10	S-B-IV	85.1 (12.4)
Tommiska 2003	Finland	BW <1000g	1996-1997	SC	72	27.1	778	Early:	24m	BSID-II	95.5
								School-age:	5	WPPSI-R	101.0
Veelken 1991	Germany	BW <1500g	1983-1986	PB	234	29.9 (2.8)	1196 (211)	Early:	18-20m	GMDS	97.3 (15.9)
								School-age:	9	KABC	88.3 (17.6)
Vermeulen 2001	the Netherlands	GA ≤32 weeks or BW <1500g	1991-1993	SC	185	29.2 (2.1)	1183 (313)	Early:	18	GMDS	99.0 (13.9)
								School-age	7-10	WISC-R	100.6 (14.0)
Wolke 1999	Germany	GA <32 weeks	1985-1986	PB	254	29.6 (1.5)	1298 (340)	Early:	20m	GMDS	90.8 (22.8)
								School-age:	6, 8	KABC	88.2 (18.6)

*For these studies, only the sub-population of participants who were born ≤32 weeks gestation and/or with birthweight <1500g were included in the review.
†Data presented are the medians and inter-quartile ranges.
‡Where participants received multiple assessments, the mean (SD) score for the assessment performed at the oldest age was presented.
Abbreviations: GA = gestational age. BW = birthweight. SC = single-centre. MC = multi-centre. PB = population-based. BIA = Brief Intellectual Ability. BSID/BSID-II = Bayley Scale of Infant Development 1st or 2nd edition. GMDS = Griffiths Mental Development Scales. KABC = Kaufman Assessment Battery for Children. RAKIT = Revision Amsterdam Children's Intelligence Test. S-B-III/IV = Stanford-Binet Intelligence Scale 3rd or 4th edition. SON-R = Snijders-Oomen Nonverbal Revised. WAIS-R = Wechsler Intelligence Scale for Adults-Revised. WASI = Wechsler

Abbreviated Scale of Intelligence. WISC-III/R = Wechsler Intelligence Scale for Children 3rd or revised edition. WPPSI/-R = Wechsler Pre-school and Primary Scale of Intelligence 1st or revised edition.

Supplemental table 2 Quality assessment of included studies using the QUADAS-2 appraisal tool

Study	Risk of bias				Applicability concerns			Reasons for being considered high risk for bias or applicability concerns, as judged against the standards set, with statements being numbered according to the domain it is applied to.
	Patient selection [1]	Index test [2]	Reference standard [3]	Flow & timing [4]	Patient selection [5]	Index test [6]	Reference standard [7]	
Bassan 2011	↑	↔	↔	↑	↑	↔	↑	[1] Inclusion criteria: birth weight below 10th percentile for gestational age. [4] Final cohort represents <70% of eligible population. [5] Study population restricted to children with birth weight below 10th percentile for gestational age. [7] Assessment tool developed before 1990 (WPPSI-R, 1989).
Bowen 1996	↔	↔	↑	↔	↑	↑	↑	[3] Assessors not blinded to results of developmental assessment. [5] Study population was born before 1990. [6] Assessment tool developed before 1990 (GMDS, 1970). [7] Assessment tool developed before 1990 (S-B-IV, 1986).
Bruggink 2010	?	↔	↔	↑	↔	↔	↔	[1] Recruitment/ sampling method not stated. [4] Final cohort represents <70% of eligible population.
Charkaluk 2011	↔	↔	↔	↑	↔	↑	↑	[4] Final cohort represents <70% of eligible population. [6] Non-universal assessment tool (Brunet-Lezine Revised, a French psychometric test). [7] Assessment tool developed before 1990 (KABC, 1983).
Claas 2011	↔	↔	↑	↔	↔	↑	↑	[3] Assessors not blinded to results of developmental assessment. [6] Assessment tool developed before 1990 (GMDS, 1984). [7] Non-universal assessment tools (RAKIT and SON-R, Dutch psychometric tests) used.
Cohen 1995	↔	↔	↔	↔	↑	↑	↑	[5] Study population was born before 1990. [6] Assessment tool developed before 1990 (BSID, 1969). [7] Assessment tool developed before 1990 (S-B-III, 1973; WISC-R, 1974 and WAIS-R, 1981).

Fedrizzi 1993	↑	↔	↑	↑	↑	↑	↑	[1] Inclusion criteria: diagnosis of spastic diplegia. [3] Assessors not blinded to results of developmental assessment. [4] Final cohort represents <70% of eligible population. [5] Study population restricted to children with spastic diplegia. [6] Assessment tool developed before 1990 (GMDS, 1970). [7] Assessment tool developed before 1990 (WPPSI, 1967).
Gray 1995	↑	↔	↔	↔	↑	↑	↑	[1] Inclusion criteria: diagnosis of broncho-pulmonary dysplasia. [5] Study population restricted to children with broncho-pulmonary dysplasia. [6] Assessment tool developed before 1990 (GMDS, 1970). [7] Assessment tool developed before 1990 (WPPSI-R, 1989).
Gray 2006	↔	↔	↔	↔	↔	↔	↔	
Hack 2005	↔	↔	↔	↔	↔	↔	↑	[7] Assessment tool developed before 1990 (KABC, 1983) used.
Kilbride 1990	↔	↔	↔	↔	↑	↑	↑	[6] Assessment tool developed before 1990 (BSID, 1969). [7] Assessment tool developed before 1990 (S-B-III, 1973).
Marlow 2005	↔	↔	↔	↔	↔	↔	↑	[7] Assessment tool developed before 1990 (KABC, 1983).
McGrath 2000	↑	↔	↔	↑	↑	↔	↔	[1] Inclusion criteria: meets *a priori* medical criterion (not specified). [4] Final cohort represents <70% of eligible population. [5] Study population was born before 1990.
Munck 2012	↔	↔	↔	↔	↔	↔	↑	[7] Assessment tool developed before 1990 (WPPSI-R, 1989).
Orchinik 2011	↔	↔	↔	↔	↔	↔	↔	
Potharst 2012	↔	↔	↔	↑	↔	↔	↔	[4] Final cohort represents <70% of eligible population.
Reuss 1996	↔	↔	?	↑	↑	↑	↔	[3] Blinding of assessors not stated. [4] Final cohort represents <70% of eligible population. [5] Study population was born before 1990.

								[6] Assessment tool developed before 1990 (BSID, 1969 and S-B-III, 1973).
Roberts 2010	↔	↔	?	↔	↔	↔	↔	[3] Blinding of assessors not stated.
Skranes 1998	↔	↔	↔	↑	↑	↑	↑	[4] Final cohort represents <70% of eligible population. [5] Study population was born before 1990. [6] Assessment tool developed before 1990 (BSID, 1969 and S-B-III, 1973). [7] Assessment tool developed before 1990 (WPPSI, 1967).
Smith 2006	↑	↔	↔	↑	↑	↑	↑	[1] Inclusion criteria: from middle to lower socioeconomic groups. [4] Final cohort represents <70% of eligible population. [5] Study population restricted to children from middle to lower socioeconomic groups. [6] Assessment tool developed before 1990 (S-B-IV, 1986). [7] Assessment tool developed before 1990 (S-B-IV, 1986).
Tommiska 2003	↔	↔	?	↔	↔	↔	↑	[3] Blinding of assessors not stated. [7] Assessment tool developed before 1990 (WPPSI-R, 1989).
Veelken 1991	↔	↔	↔	↑	↑	↑	↑	[4] Final cohort represents <70% of eligible population. [5] Study population was born before 1990. [6] Assessment tool developed before 1990 (GMDS, 1970). [7] Assessment tool developed before 1990 (KABC, 1983).
Vermeulen 2001	↔	↔	↔	↔	↔	↑	↑	[6] Assessment tool developed before 1990 (GMDS, 1970). [7] Assessment tool developed before 1990 (WISC-R, 1974).
Wolke 1999	↔	↔	↔	↔	↑	↑	↑	[5] Study population was born before 1990. [6] Outdated assessment tool (GMDS, 1970) used. [7] Outdated assessment tool (KABC, 1983) used.

↔ = low risk; ↑ = high risk; ? = unclear risk
Abbreviation for assessment tools are followed by year of publication:
BSID = Bayley Scale of Infant Development, 1969
GMDS = Griffiths Mental Development Scales,1970 and 1984
KABC = Kaufman Assessment Battery for Children, 1983

RAKIT = Revision Amsterdam Children's Intelligence Test, 1987
S-B-III/IV = Stanford-Binet Intelligence Scale 3rd edition, 1973 and 4th edition, 1986
SON-R = Snijders-Oomen Nonverbal Revised, 1998 and 2003
WAIS-R = Wechsler Intelligence Scale for Adults-Revised, 1981
WISC-R = Wechsler Intelligence Scale for Children-Revised, 1974
WPPSI = Wechsler Pre-school and Primary Scale of Intelligence-Revised, 1989